A Quantitative Approach to Concept Analysis

Rogelio Nazar

TESI DOCTORAL UPF / ANY 2010

DIRECTORES DE LA TESI

Dr. Jorge Vivaldi Palatresi (Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra)

Prof. Dr. Leo Wanner (Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra)

UNIVERSITAT POMPEU FABRA

Acknowledgments

First of all, I thank my parents, for all the love, protection and education. Also my brothers, sisters, nieces and friends I left in Argentina, for their support and for forgiving me for my absence all these years. I express also my gratitude and appreciation to my advisors, Jorge Vivaldi and Leo Wanner, for the knowledge they shared with me, and for their enthusiasm and patience. This thesis would have been impossible without the support of Institut Universitari de Lingüística Aplicada. I am in debt with the three chairs of the Institut during my time there, Teresa Cabré, Teresa Turell and Mercè Lorente, for their intellectual, financial and affective support. This research was supported first through an ADQUA scholarship from Generalitat de Catalunya and later by a contract with the Ministry of Education of Spain. Hinrich Schütze provided valuable feedback for the firsts steps of this thesis during my research stay in University of Stuttgart, and also accepted to write an external evaluation. Juan Manuel Torres-Moreno also accepted to write an evaluation and to be a member of the jury. I thank as well those who accepted to be members or alternate members of the jury: Teresa Cabré, Guadalupe Aguado, Jean Véronis, Horacio Saggion, Maarten Janssen, Horacio Rodríguez, Mick O'Donnell, Kim Gerdes and Irene Castellón. Jean Véronis also shared with me the dataset to replicate his experiments and provided feedback. I am very much in debt to my friend Chris Norrdin, who corrected many of my grammar errors. He is not to blame for the remaining errors because the text suffered intense post-editing. The same goes for errors of facts or reasoning, which are my own and not of my supervisors. I benefited from discussions with my colleagues at the Institut, my thanks to Teresa Cabré, Jaume Llopis, Gabriela Ferraro, Sabela Fernández, Jenny Azarian, Evanthia Triantafyllidou, Marta Sanchez, Maarten Janssen, Teun van Dijk, Irene Renau, Janet deCesaris, Paz Battaner, Vanesa Vidal, Araceli Alonso, Raquel Casesnoves, Rosa Estopa, Lluís De Yzaguirre, Manuel Souto, Albert Morales, Juan Manuel Pérez and many others. Horacio Panella and Francisco Vallejo programmed a nice flash-interface for my system. Ricardo Baeza-Yates provided corpus and access to the Yahoo BOSS platform. Vanessa Alonso, Sylvie Hochart and Jesus Carrasco provided indispensable tactical support. A group of specialists in the field of medicine kindly accepted to evaluate the results of knowledge extraction algorithms presented in this thesis: Antoni Valero, Jaume Franci, Hugo Vitale, Jorge Nazar and Graciela Nazar. Also a group terminologists contributed with the evaluation: Amor Montane, Carles Tebé, Carme Bach, Natalia Seghezzi, Irina Kostina, Alba Coll, Gabriel Reus and Iria da Cunha. Finally, thanks to Adriana Gorri for being always there and, after all, because it was her idea to come to this beautiful country.

Abstract

The present research focuses on the study of the distribution of lexis in corpus and its aim is to inquire into the relations that exist between concepts through the occurrences of the terms that designate them. The initial hypothesis is that it is possible to analyze concepts by studying the contexts of occurrence of the terms. More precisely, taking into account the statistics of term co-occurrence in context windows of *n* words. The thesis presents a computational model in the form of graphs of term co-occurrence in which each node represents single or multiword terms. Given a query term, a graph for that term is derived from a given corpus. As texts are analyzed, every time that two terms appear together in the same context window, the nodes that represent each of these terms are connected by an arc or, in case they already had one, their connection is strengthened. This graph is presented as a model of learning, and as such it is evaluated with experiments in which a computer program solves tasks that involve some degree of concept analysis. Within the scope of concept analysis, one of those tasks is to tell whether a word or a sequence of words in a given text is referring to a specific concept and to derive some of the basic properties of that concept, such as its taxonomic relations. Some other tasks can be to determine when the same word is referring to more than one concept (cases of homonymy or polysemy) as well as to determine when different words are referring to the same concept (cases of synonymy or equivalence between languages or dialectical variations). As a linguistic interpretation of these phenomena, this thesis derives a generalization in the realm of discourse analysis: the properties of the co-occurrence graphs are possible because authors of argumentative texts have a tendency to name some of the basic properties of the concepts that they introduce in discourse. This happens mainly at the beginning of texts, in order to ensure that principles among reader and writer are shared. Each author will predicate different information about a given concept, but the authors that treat the same topic will tend to depart from a common base and this coincidence will be expressed in the selection of the vocabulary. This coincidence in the selection of the vocabulary, because of its cumulative effect, can be studied with statistical means.

Resumen

El presente trabajo se centra en el estudio de la distribución del léxico en corpus y su cometido es el análisis de las relaciones existentes entre los conceptos a través de los términos que estos designan. La hipótesis de partida es que podemos analizar conceptos estudiando los contextos de aparición de los términos que los designan, utilizando para ello las estadísticas de coocurrencia de los términos en ventanas de contexto de *n* palabras. La tesis presenta un modelo computacional en forma de grafos de coocurrencia de términos donde los nodos representan términos simples o sintagmáticos. Dado un término analizado, se deriva un grafo para ese término a partir de un corpus. A medida que los textos se analizan, cada vez que dos términos aparecen juntos en una misma ventana de contexto, los nodos que los representan se conectan entre sí mediante un arco o bien fortalecen su conexión si ya la tenían. Este grafo es presentado como un modelo de aprendizaje, y como tal es evaluado mediante experimentos en que un ordenador resuelve tareas propias del análisis conceptual. Estas tareas incluyen determinar cuándo una palabra o secuencia de palabras dentro de un texto hace referencia a un concepto definido, así como determinar algunas de las propiedades más importantes de este concepto, tal como sus relaciones taxonómicas. Otras tareas son las de determinar cuándo una misma palabra puede hacer referencia a más de un concepto (casos de homonimia o polisemia) o determinar cuándo distintas palabras hacen referencia a un mismo concepto (casos de sinonimia o equivalencia entre lenguas o variedades dialectales). Como una interpretación lingüística de estos fenómenos, esta tesis extrae una generalización en el plano del análisis del discurso: las propiedades de los grafos de coocurrencia léxica surgen gracias a la tendencia que tienen los autores de textos argumentativos de mencionar algunas de las propiedades más importantes de los conceptos que introducen en el discurso. Esto ocurre sobre todo al inicio del discurso, con el objeto de asegurar que los principios entre lector y autor son compartidos. Cada autor predicará distintas informaciones acerca de un determinado concepto, pero los autores que traten sobre un mismo tema tendrán tendencia a partir de una misma base y esta coincidencia se manifestará en la selección del léxico que, por su efecto acumulativo, puede ser estudiada de manera estadística.

Table of Contents

Index of Figures

Index of Tables

Chapter 1: Introduction

The aim of this research is to inquire into the relations that exists between concepts and the distributional behavior of the terms that designate them. Within the framework of distributional semantics, the thesis proposes a representation of a concept as a function that takes as arguments a term that designates such concept and a set of contexts of occurrence of such term. This aim entails a statistical interpretation of what a concept is: a configuration of terms (and, by implication, of other concepts) that typically surround the instantiations of a term or, more generally, a sign. This research raises questions of theoretical interest and also opens diverse practical applications for natural language processing. To name only some of them, one can think of named entity recognition, taxonomy extraction, word sense disambiguation, semantic-neology extraction and term translation. This research focuses, hence, on the study of the behavior of terms in text using statistics of co-occurrence, in order to find relations between the concepts that are designated by such terms.

In principle, the application of a mathematical apparatus such as statistics to the analysis of concepts may seem disconcerting at first glance. Nevertheless, the quantitative approach reveals itself as the natural option, as will be shown later. This study will show how terms have a tendency to cluster, forming configurations of terms of related meaning in a form that recalls a molecular organization. In this context, the term *configuration* is used in the sense of a meaningful totality, as the German term *Gestalt:* a unity of a system that is more than the elements by which it is conformed. These configurations have a measurable overlap with the meaning of a given term *t*, which suggests that concepts can be automatically represented on the basis of associational learning. A simple model of associational learning can consist of the analysis of a sufficiently large number of contexts of occurrence of *t.* With such a model, one can identify a pattern of recurrence of a set of lexical units that configure the typical context of *t.*

The thesis is thus committed to the study of the structuring of vocabulary into meaningful configurations that can be obtained by statistical association. The study is presented as a mathematical model of a learning process based on association by repetition of a combination of events. It takes as input a term *t* and a collection of documents where *t* occurs, and outputs a co-occurrence graph that contains a structure of terms related to the meaning of *t* in such collection. Terms become associated when they appear at short distances from each other more frequently than expected

by chance. This numerical relation is a correlate of the importance that repetition of the combination of events has in the mechanisms of human cognition, where one sees repetition of verbal stimuli and reinforcement of associational strength. Statistics is, therefore, the most suitable tool for the study of this phenomenon. An examination of text using corpus statistics reveals the knowledge needed for the construction of a reduced-scale simulation of the associational network of concepts. For a given term, this network is the result of a convergence of clusters of terms that represent the canonical features of the concept designated by such term.

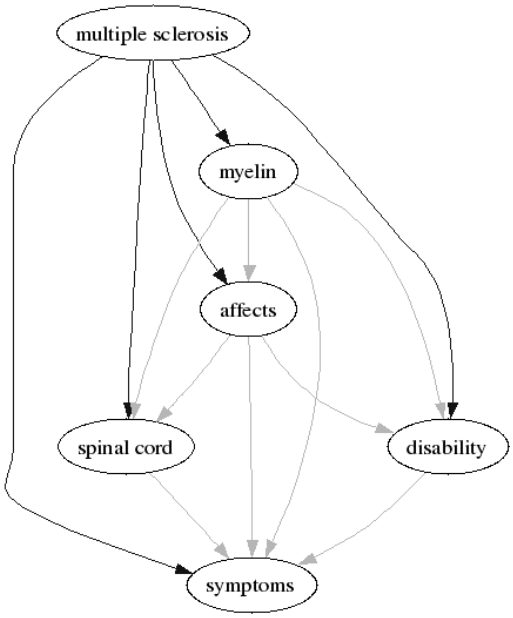

Figure 1: A simple co-occurrence graph of "Multiple Sclerosis".

Figure 1 shows an example of a simple graph of lexical co-occurrence with the term *Multiple Sclerosis.* The nodes represent terms from documents downloaded from the web using *Multiple Sclerosis* as a query term. Connections between nodes link terms which have significant frequency of co-occurrence in a window of fifteen words. Statement (1) is the definition for the same term provided by WordNet[1].

> 1) *Multiple Sclerosis: (n) A chronic progressive nervous disorder involving loss of myelin sheath around certain nerve fibers.*

One can see that there is a certain degree of overlap between the graph and the definition. The words are not exactly the same, however the

[1] http://wordnet.princeton.edu/perl/webwn?s=multiple+sclerosis&sub=Search+WordNet&o0=1&o1=1 [accessed June 2010].

conceptual proximity is evident. Certain degree of inference can be performed to fill the eventual absence of a relevant term, such as the hypernym *nervous disorder*. It is, thus, possible to say that if *Multiple Sclerosis* affects the *spinal cord* (and more specifically, its *myelin*) then *Multiple Sclerosis* is a *nervous disorder*.

The question that arises is: why are these terms together in the graph? The answer is that in the documents devoted to this topic, as well as in science in general, authors tend to define the terms they use and to present the state of the art of the particular topic they are writing about. If a text refers to an object, it is to be expected that the text will also predicate some of the canonical attributes of this object.

1.1 Basic definitions

This section is aimed to specify some of the most important notions used in the present work. These definitions rely on many assumptions, and more attention will be devoted to clarify them in Chapter 3. Here, a preliminary delimitation should be enough. The notions defined in this section are: *concept, term, referential expression*, *discourse, analytical statement* and *synthetical statement.*

Concept: According to Saussure (1916), concepts, characterized as *facts of consciousness*, are opposed to terms as two faces of the same coin. In contrast, in this thesis, a concept will be defined as a knowledge representation unit that is interpreted or described through the relation that it has with other knowledge representation units. Still a Saussurean interpretation, it stresses the fact that language is a system *où tout se tient*. Consider example (2) from the WordNet[2] lexical databse.

> 2) *Minimal brain dysfunction: A condition (mostly in boys) characterized by behavioral and learning disorders.*

The concept *minimal brain dysfunction* is related to the more general concept of *condition* and other more specific concepts such as *behavioral disorder* and *learning disorder.* A concept is thus defined in relation to other concepts in order to make sense and to be integrated with the rest of cognitive constructions.

Another fundamental characteristic of concepts, from a logical perspective, is their possibility of being defined in *extension* and *intension* (Bunge, 1960; 1974). The extension is the set of objects that can be referred to by terms or that are instances of a given concept, while the intension is the set of attributes that an object must posses to be considered an instantiation of a concept.

[2]http://wordnet.princeton.edu/ [accessed June 2010].

Term: As a generalized definition, one can adopt that provided by Cabré & Estopà (2005), who state that a lexical unit in a given text is a term if it designates a specialized concept inside a discipline. In the present thesis, however, the expression *term* will be used in a more classical sense, as a single or multiword unit that designates an object, concept or category in a proposition, leaving aside its normative status. A proposition, again in a classical sense, is the content of a sentence with independence of its particular wording (Bunge, 1974). Thus, two sentences can be formulated different and yet convey the same information. When terms refer to specific entities, terms are a subset of the more general category of referential expressions. Terms can be polysemic -or homonym- when they refer to more than one concept. For instance, the same term *Sudoku* can refer to a rat-bite induced fever or to a game, depending on the context. The specific difference between terms and the rest of all lexical units is that their function is designative, while lexical units that are not terms are often found to serve also as predicates. This would not be fully true from the terminologist's point of view, because specialized terms can also function as predicates, that is, they can take arguments (l'Homme, 2005). Having said this, it is still clear that a predicate cannot designate a concept. It would probably be more convenient to treat predicative terms as a special class, such as "specialized predicates".

Referential Expression: As terms, referential expressions have a designative function. However, they are more general since there can be units that are referential expressions but not terms, like proper nouns or demonstrative pronouns. This thesis uses Frege's (1892) definition of reference, as the property of discourse by which it designates concrete or defined entities, whether they be a singular object or a concept, while the rest of the vocabulary is used to establish predicative relations between these entities. Russell (1905) and Strawson (1955) use the term in the same way, despite other differences in their perspective. Bunge (1974) is, however, more restrictive since, for him, reference is possible when the sign designates particular instances, thus, strictly speaking there could be no reference to categories, concepts or ideas. This is because Bunge separates concepts in individual (e.g. *Mars*) from collective (e.g. *planet*). As a consequence, class concepts should be regarded as predicates. This view, however, is not generally shared by most computational linguists: "...by a referring expression we mean a linguistic form that is used to pick up an entity in some real or imaginary world, by complex entity we mean not only singular individuals […] but also what philosophers of language would call non-singular individuals" (Dale, 1992, p.1). In principle, proper nouns are those that are used to designate instances, while common nouns designate classes (Bosque, 1999), however this is less clear in some cases, for instance when the proper noun is used in plural to refer to classes of objects, as in "we have two Picassos in this collection" (Coseriu, 1967). In the case of the specialized terms, they belong to a

special class because they are like proper nouns, since they posses a referential value (Wüster, 1979; Cabré, 1999) and still they designate categories rather than instances. In the context of this research the difference between concrete entities or abstractions from them is not considered relevant, as both, material and conceptual are treated as constructs and handled "as if they were autonomous -Platonic Ideas- without however assuming that they are such" (Bunge, 1974, p. 13). Consequently, in the context of this research it will be assumed that there is no contradiction in the fact that expressions refer to concepts.

Discourse: Discourse is, in the context of this thesis, a general notion of corpus as a sample of the documents that circulate in a community. In most of the experiments within this work, the Internet is used as the corpus because it represents a sample of the scattered knowledge, without taking into account the question of the truth or falsity of the propositions contained in that corpus. What is important is that the conclusions of the thesis can also be applied to other more reliable text sources.

Analytical and Synthetical Statements: Analytical propositions are propositions whose predicate is implicit in the subject by definition while synthetical propositions are those whose predicate adds information that is new or not implicit in the subject. The prototypical example of an analytical statement is the definition. As an example, consider the definitional statement (3).

> 3) *A circle is a figure consisting of a line whose points have the same distance to the center.*

If one splits the definition into its grammatical subject and predicate, the subject being "A circle" and the predicate being the rest of the sentence, one sees that the predicate does not add new information since it is contained in the subject precisely by definition. The definition is, hence, a meta-linguistic device. It is addressed to those who do not know the code, that is, the meaning of the subject.

These two notions of analytical and synthetical statements appear sometimes with different names (truths of reason / truths of fact) already in the work of Leibniz (1765), Hume (1777) and Kant (1787). They are still in use, with different terminology (semiotic statements / factual statements), in fields such as semantics, semiotics and philosophy (Quine, 1951; Eco, 1968). It is also important to bear in mind that this distinction between statements is historical and not logical, because it depends always on the historical contexts of the utterance.

These terms are important for this thesis because, if terms that are syntagmatically related to a term *t* are the basic conceptual features of the meaning of *t*, then these terms are included in the analytical statements of

t. Starting with the assumption that the set of all possible statements can be divided into these two categories of analytical and synthetical statements, this thesis proposes a formal distinction between them. This distinction is based on a mathematical model that takes lexical distribution into account and identifies the set of the analytical statements as a prominent group. More space will be devoted later to the clarification of this terminology. For now, the most important idea is that there are patterns across large corpora that reflect the way concepts are structured in language, and that these patterns of recurrence is a linguistic phenomenon, present in every language.

1.2 General Outline of the Approach

The approach developed in this thesis is to build, for a given term t, a graph of lexical co-occurrence such that each term t_i that occurs in the context of t is assigned to a node. The connection between nodes t and t_i is strengthened every time the terms that these nodes represent are found in the same context window. With each activation, the rest of the connections is weakened. As the learning progresses, nodes that were assigned a particular term but received no further stimulation, that is, that found no significant connections with neighbors, tend to be de-emphasized and finally eliminated. At the end of the process, the remaining nodes are the most interconnected ones. The main claim in this thesis is that the resulting nodes represent the canonical features of the concept designated by a term at the specific moment in which the documents of the sample were produced.

The analysis of the resulting graph opens up new lines of research in distributional semantics. The observation of the geometrical properties of the graph that a given term generates reveals a difference between the terms that function as referential expressions from those general words of the lexicon that have a predicative function. Referential expressions show a defined set of neighbors (a lexical cohort) and their graphs are more densely interconnected. The rest of the words, which lack this referential function, produce poorly interconnected graphs since the documents where they occur are not usually related with respect to a given concept. The only condition for the distinction to take place is that the referents that appear in the analyzed text are shared knowledge. By shared knowledge it is meant that the referent is socially perceived and is therefore present in a representative corpus of some community of discourse. In terms of Lara (1997, p. 92, [my translation]): “It is stated that any utterance, to be meaningful, needs to be adjusted to a social consensus on the meaning of signs in language, which is what defines its meaning and relevance. This consensus is made in the 'information space'

created between the individual, the society and the world around them. It is this consensus that establishes the significance in their social and communication context and any message or information, to be successful, needs to account for this pool of common knowledge and shared experiences, which is nothing but the horizon of interpretation that every member of society has to understand and to be understood".

The web is, currently, a repository of the world's shared knowledge. Thus, if a unit is socially perceived, it is to be expected that it will be represented in the web, having a minimum number of documents on the subject. Names of famous people, places, specific concepts and terms are in this situation, and they conform the set of entities of shared extra-linguistic knowledge. For this reason, most of the experiments shown in this research use the web as a corpus.

Furthermore, with the inspection of the degree of the overlap of different graphs, an implicit hierarchical structure begins to emerge. For instance, if A is a graph for the term *multiple sclerosis,* B the graph for *bronchial asthma,* and C the graph for *disease,* A shows a strong association with C (C is a central node of graph A) while B, which is also associated with C, has no relation with A, nor does A with B. This type of hierarchical relations between term parallels the taxonomical organization of the concepts of a discipline.

Another interesting property of the graphs is shown in the case of polysemous terms, since graphs show different clusters of nodes for each of the meanings that the expression has. When polysemous or homonym words are queried by this statistical method, contexts of occurrence of these words are clustered according to each different use.

Still from a semantic point of view, similarities between the co-occurrence graphs of different terms can indicate if such terms are equivalent, i.e, they refer to the same concept. The equivalence can hold inside the same language, in the case of synonyms, or across different languages or dialectal variations within the same language. In the case of equivalent terms, their respective graphs show a high degree of overlap, meaning that they share part of their profile of co-occurrence. Interestingly, synonyms are less likely co-occur in the same context. They may co-occur more frequently in large context windows, however, as a consequence of a stylistic reason: authors tend not to repeat themselves too much, therefore they try to use synonyms to refer to the ideas that were already introduced in their texts.

1.3 The Contribution of this Thesis to Theoretical Linguistics

A brief sketch of the place of this work in theoretical linguistics seems appropriate. As stressed at the beginning of the introduction, the thesis focuses on the laws that rule the distribution of lexis in discourse. Traditionally, discourse analysis has been clearly differentiated from the rest of linguistics, having its own principles and methodology. Corpus linguistics, which is the most suitable field for this thesis, lies somewhere between theoretical linguistics and discourse analysis because it studies *parole* rather than *langue*. In comparison to discourse analysis, which works on a document-by-document basis, corpus linguistics analyses texts on a large scale, inquiring about characteristics of language that are beyond the scope of direct human observation.

This thesis is a contribution to theoretical linguistics in the specific field of quantitative (or distributional) semantics, because it proposes an explanation of how subtle structures found in discourse, namely, patterns of repetition of lexical combination, reflect the way in which speakers conceptualize entities of the extra-linguistic world. It is a contribution, in particular, to the tradition of distributional properties of lexical units in discourse, which were initially studied by Harris (1954) and were the basis for later psycholinguistic research on the phenomenon of lexical priming (Aitchison, 1987; Belinchon et al., 1992; Bybee & Hopper, 1997; Loritz, 1999; Vega & Cuetos, 1999; Hoey, 1991; Hoey, 2004; Hoey, 2005). The priming effect is defined as the capability of a linguistic unit to relate to other units or to predict other units in discourse. According to this principle, terms are related in memory by semantic association. Thus, when one thinks about a term or referent, a variety of other terms are primed or activated. For instance, *table* activates *chair*; *bread* activates *butter*, etc[3].

This thesis hypothesizes that something similar occurs in the case of terms that have a precise denotation, such as scientific terminology or proper nouns, only that in this case they tend to activate their salient conceptual

[3] Skipper & Zevin (2010) have empirically measured this brain activation with functional magnetic resonance imaging (fMRI) in subjects that are exposed to auditory stimuli of both predictable and unpredictable sentences, where the element that is to be predicted by the listener is followed by a pause (a filled gap, such as "um...") . They found that in the case of predictable sentences, those of the type of "The pond was full of croaking, um... frogs", there is a significant amount of brain activation during the pause, before the predicted element is uttered, and reduced when the listener finally receives the missing element. In contrast, in the case of unpredictable sentences, of the type "she was talking about, um..., frogs", there is a low brain activity during the pause, until brain activity is triggered by the missing element, *frogs*.

features. Thus, *penicillin* is associated with *antibiotics* and *infections*, while *Napoleon* is associated with *Emperor of France*; *King of Italy* and *Waterloo*, among others. These units are associated in memory but also are syntagmatically related in text (Nazar, 2005; Wettler et al., 2005) and this is why it is possible to develop a model or a mechanism that can predict the conceptual relation that exists between these terms by using term co-occurrence in large corpora[4].

Units can also be related from a paradigmatic perspective. That is, they may not occur in presence of each other and still have a tendency to share part of their typical profile of co-occurrence. This is the case, for instance, of synonyms or different expressions that refer to the same concept and do not necessarily show a priming effect (Plaut, 2005).

One of the main characteristics of this approach is its high degree of generality, namely in the following two aspects. In the first place, the conceptual relations that are to be expected among terms are not previously defined or categorized. They may range from very general, such as hypernymy, which plays a fundamental role in human cognition, to very domain specific ones, such as the kind of action of an enzyme on certain molecules. The second aspect is that no explicit knowledge of a particular language is introduced. The decision not to use any of the available language resources is scientifically motivated since it was considered necessary to separate language independent phenomena at the theoretical level. This thesis presents a study of terminology *in itself*, and not the terminology of a specific thematic area and language. The intrinsic laws that rule terminological organization in argumentative discourse are the structural regularities of discourse that can be conceptualized from a statistical point of view. In any case, in an eventual concrete application, it may be reasonable to incorporate different sources of information, meaning knowledge about the language or the specific domain in question. Notice, however, that it is not obvious that the introduction of explicit linguistic knowledge will increase the performance of a distributional method in practical applications, and in fact there are reports (Sahlgren, 2008) that point out that the addition of explicit knowledge can be even detrimental in some cases. That is a claim that still needs to be backed with empirical evidence, however, it is not the purpose of this thesis to sustain or reject it. The introduction of explicit knowledge and the comparison of performance of an algorithm with and without this knowledge is left for future work.

[4] Thanks to recent advances in neuroimaging and machine learning (Mitchell et al., 2008) it is now even possible to try to predict patterns of brain activation using only statistics of term co-occurrence on large corpora.

1.4 The Contribution of This Thesis to Applied Linguistics

Beyond its pure theoretical contribution, this research also offers practical applications in a wide range of fields. Chapter 9 is devoted to the enumeration and explanation of various possible fields of application of the approach developed in the scope of this thesis. Among them is automatic terminology extraction. Also, with further syntactic and semantic analysis of the relations between terms, it will be possible to automatically generate a conceptual representation of a term. Applications for information retrieval (IR) could be based on such conceptual representation as a means of document clustering. IR systems based on this approach could also offer services of term retrieval for those who do not know or cannot remember the exact denomination of a given concept. Bilingual terminology acquisition is another field of application that has been evaluated and shows promising results.

At first, these areas of application may seem completely different, and in fact they present different types of problems. However, the solution is in all cases based on the same principle: the geometrical properties of the lexical co-occurrence graphs.

1.5 Structure of this Thesis

The structure of the thesis is the following: After this introduction, Chapter 2 offers a brief presentation of the basic hypotheses and the limits of their scope. Chapter 3 offers a background on the basic notions used in this thesis. Chapter 4 is devoted to the discussion of the existent formalisms for conceptual representation. Chapter 5 summarizes the corpus of related work divided into two main sections: the symbolic and the statistical approaches to the automatic extraction of knowledge. Chapter 6 contains a detailed description of the basic algorithms that are used. Chapter 7 shows the experiments undertaken to test the hypotheses. Chapter 8 is devoted to the conclusions from this work, and Chapter 9 discusses the wide range of possible directions of future work including practical applications. Finally, after the list of the cited bibliographic references, appendices with extra-information about the outcome of the experiments follow.

Chapter 2: Working Hypotheses

This chapter presents the working hypotheses and an in-depth examination of its details, as well as an outline of some of the limits of the scope of these hypotheses.

2.1 Main Hypotheses

The general hypothesis of this thesis assumes the dichotomy between analytical and synthetic statements presented in section 1.1. Recall these are two types of propositions in argumentative discourse: the propositions that present and define the terms to be used (analytical statements) and propositions that contribute some knowledge that cannot be directly inferred from the definition of terms (factual or synthetical statements). In general terms, the hypothesis of the thesis is that there exists a correlation between term co-occurrence and conceptual structures as a consequence of the fact that analytical statements about a given concept are salient than the synthetic propositions about such concept. This general hypothesis can be divided in two more specific hypotheses:

1. There exists a statistical association between the terms that refer to conceptual features (or essential properties) of a given concept and the term that refers to this concept.
2. In discourse, there exists a statistically observable difference between the terms that have a referential function and the rest of the words.

Let us explain these hypotheses in more detail by means of some examples. Example (1), below, is analytical because it predicates established knowledge about its subject, while (2) is factual because of its relation to a current state of affairs, that is, because it provides new -or contingent- information.

1) The Brandenburg Gate is situated in Berlin.
2) There is a demonstration at the Brandenburg Gate.

Depending on what is really happening in that particular place of Berlin at the moment of utterance, statement (2) is or is not true. In contrast, in normal conditions the truth of statement (1) needs no verification.

One should not, however, be deceived by the simplicity of the example. Example (2) implies a reference to the exact moment of the utterance, which makes it a historical statement. That is: *There was a demonstration* (of some importance) *at the Brandenburg Gate on the day* such and such. This may become an analytical statement if the event becomes part of the socially shared knowledge in relation to the Brandenburg Gate. It is a

matter of degree when a historical statement is incorporated into the general knowledge of an average educated person.

Consider Examples (3) and (4). Example (3) is analytical. Example (4), in contrast, appears in a certain moment of history, as a synthetic statement. It may happen, if the claim is successful and historians agree, that with time this becomes an established knowledge about Napoleon too, although at the present moment, the fact that he has died of arsenic poisoning is not directly inferred from the referent *Napoleon*.

3) Napoleon was the Emperor of France and King of Italy.

4) Napoleon was poisoned with arsenic[5].

Of course, propositions such as (1) and (3), even with different formulations, are today far more frequent than those as (2) and (4). This is a property of discourse that provides the necessary conditions to capture the meaning of a term as the configuration of frequent neighbors.

The second hypothesis is a consequence of the first. As mentioned above, reference is intuitively easy to recognize. Without this ability, the text would be incomprehensible. From pure introspection, the functions of reference and predication of the three elements involved in (4) can be easily recognized. *Napoleon* as a reference to a historical character, *arsenic* as reference to a chemical substance and *poisoned* as the predicate that links the two referents. In Aristotle's terminology (The Categories), *Napoleon* and *arsenic* would be categories or topics while *poisoned* is the element that unifies the categories into a proposition, being an indispensable element for the sentence to make sense. This distinction between sense and reference continues until present day's linguistic terminology: only the categories have reference, while sense is a property of the proposition (Coseriu, 1967).

Some referential expressions may also contain implicit forms of propositions. For instance, one can refer to *Sir Walter Scott* by the expression *the author of Waverlay*, if one assumes that the message will reach an audience that is acquainted with that information. The expression has a reference function but contains an implicit proposition: *Scott is the author of Waverlay*.

The term *reference function* should not be confused with *referential function* used by Jakobson's (1960). In the context of this thesis, the term is used more restrictively. The referential function of an expression is understood as the property of an expression (in a given utterance and

[5]Apparently, the most accepted version is that he died of stomach cancer. However, after an analysis of the rests of his hair at the end of the XX Century, the version of the poisoning began to circulate (Weider, 1998). This version, however, is discredited by Lugli et al. (2007), who point out that he actually died of gastric cancer.

context) that designates a specific entity. For Jakobson, in contrast, it is the content or information, in general, that is carried by a message, opposed to other functions such as emotive, conative or phatic. In the context of this thesis the notion of referential function is used in opposition to predicative function. In every text, entities are referred to and something is predicated about them. There is a basic difference between the actants of some discourse and what is predicated about them. There is the correlate in the syntactic level: the sentence has subject and predicate. Of course, not always the grammatical subject coincides with the semantic actant. Such belief is the misleading cue of grammaticalism (Bunge, 1974). A better terminology for denoting this distinction may be theme and rheme (Van Dijk, 1993), theme for that of what the statement is about and rheme for what is said about it.

Some objection may arise at this point, based on the argument that, ultimately, all words can have reference, even the word *poisoned*, since it refers to the concept of poison. With independence of whether or not it is true that all words can denote concepts, it is important to examine this objection: it is perfectly admissible that the same word can have a referential function in one context and not in another. Reference is a property of discourse (a text in a context) and not of the lexicon. It is a form of deixis and a special kind of speech act (Kronfeld, 1990), which is, at the same time, an act of *parole* (the realization of language in a specific utterance, in the specific context where the meaning of the message is anchored) and an act of *langue* (the language as a system, as the knowledge of the language that the native speaker has at some moment).

This thesis will show a series of structural properties that can be useful for finding a behavioral difference between the types of units that have referential function in discourse and those which have not. A first condition for an expression to have a referential value is to be socially perceived, meaning that it is a "cultural unit" (Eco, 1968). The claim here is therefore that referential units have a defined lexical cohort while the rest of the vocabulary, which does not have a denominative function, do not show this behavior. The units that have a predicative function instead of a referential value tend to engage in relations with a more diverse vocabulary, a phenomenon that can be traced by co-occurrence statistics and dispersion measures.

It would be, however, naive to regard the problem of the distinction between sense and reference without taking into account the diachronic perspective (and historical, in the sense of extra-linguistic), given that reference is a relation between discourse and reality. Therefore, the principle of immanence -which many linguists would still like to adhere- does not apply in this research.

A linguistic sign may be both referential or non-referential depending on the context, and this must be understood without conflict nor contradiction. Consider, for instance, the expression "September 11". That was not referential before September 11, 2001. However, from that day on, it began to be a referring expression. There is no need for empirical evidence, however there are traces of this shift of status of the expression that can be observed with simple methods taking the diachronic axis into account. Figure 2, for instance, shows the frequency distribution words *septiembre*, *octubre* and *noviembre,* the Spanish words for *September*, *October* and *November* in the archives of the EL PAÍS newspaper in the period 1976-2007. The vertical axis shows the relative frequency and the horizontal axis the editions of each year. All these words should have more or less the same frequency and thus a relatively horizontal line. The word *septiembre* (*September*) is the only one to show a peak on the years 2001-2002, and having peaks in small intervals of time is, as it will be shown in Section 7.2.3, the typical behavior of a referential expression.

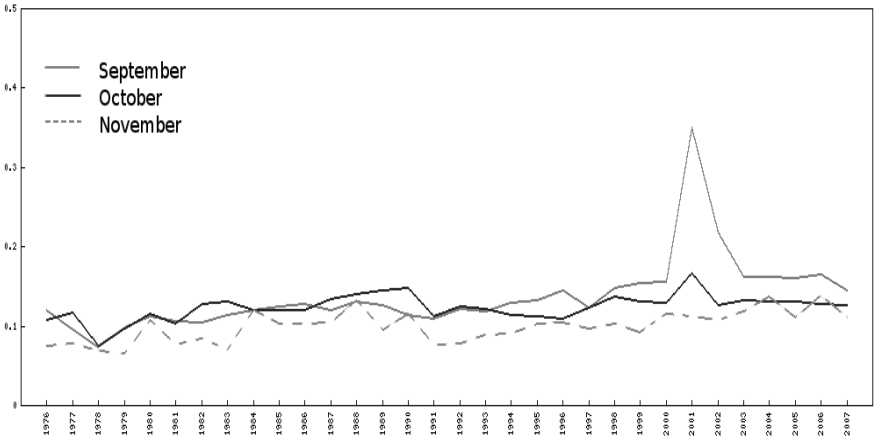

Figure 2: Frequency distribution of names of months in dyachonical axis with a peak in the year 2001 for "September"

From the synchronic perspective, it should also be obvious that an important number of lexical units show co-occurrence relations with others too, even if they are not referential expressions. Relational names such as *mother* and *son* are examples of this kind. There exist also relations of co-occurrence of the elements that form collocations (Smadja & Mckeown, 1990; Wanner et al., 2006), such as *to contract* and *obligation*, or *to repay* and *loan*. But the case of the referential units is different because the phenomenon is more pronounced (and this difference can be quantitatively determined, as it will be shown in Chapter 7). The positions of the collocates are, as well, independent of their order in discourse and the collocational distance is more flexible than with typical verb-noun collocations.

The study of the status of a referential unit is thus based on the statistics of a sufficient number of contexts of occurrence. *Napoleon* and *arsenic* both have, as a sort of "orbit", a defined set of related terms, while *poisoned*, instead, shows no such behavior, or if it does, it is always to a lesser extent. The predicate *poisoned* appears always with a greater variety of vocabulary.

This phenomenon can be explained by the concept of textual strategy (Eco, 1979): writers have a model of the reader they are addressing. This model is what a writer thinks about the knowledge of a possible reader and how he or she is supposed to deal with the interpretation of the text. When a text mentions a referent for a first time, the writer has to bear with the dilemma of not wanting to lose the readers who do not know that specific referent and at the same time not underestimating the readers who do know it. This is where rhetorical strategies come to the aid of the writer, such as appositions, periphrasis or expressions that function as co-reference, helping to avoid the repetition of a referent's name and, at the same time, giving the opportunities to mention its attributes. These phenomena can be considered implicit forms of analytical propositions. There are plenty of examples in every text and in every language. Take for example the following paragraph:

> *"La próxima semana, Diego Maradona recibiría el alta médica y podría abandonar el sanatorio Güemes", admitió el director del centro de salud, doctor Héctor Pezzella. El astro fue internado el 28 de marzo, contra su voluntad.* (Clarín Newspaper, 04/08/2007).
>
> (*Next week, Diego Maradona will be discharged and could leave the Güemes clinic, as admitted by the director of the hospital, doctor Héctor Pezzella. The star was taken to the hospital on March 28th, against his will*).

This fragment, taken from an Argentinian newspaper, shows an example of co-reference. In the first sentence, the journalist uses the name of Diego Maradona; in the second, he uses the expression "el astro" (the star). This is an implicit analytical proposition. Explicitly, the proposition would be *Maradona is a star.*

This prevailing presence of analytical propositions is what makes possible the extraction of relevant information with the aid of statistical techniques, even when the corpus is unstructured and noisy. What is relevant is what a discourse community states as analytical propositions about a referent. *Former football player* and *star* are not synonymous with Maradona, but they are relevant attributes and, in context, they function as co-referents.

Using a corpus, and with the aid of measures of statistical association examined in detail later in Chapter 6, it is possible to obtain some other relevant attributes of the cultural unit *Maradona*. In a sample of around 660 fragments of texts retrieved from the web where this unit occurs, one finds out that among the most frequent *n*-grams there are words of related meaning, precisely the topics related to this person. Some of these expressions, with brief explanations, are listed in Table 1.

Unit	**Explanation**
Diego Armando Maradona	His full name
hand of god	A famous goal that left England out of the 1986 World Cup
1986 world cup	1986 Football World Cup
Buenos Aires	Argentina's Capital
Villa Fiorito	Where he was born
Boca Juniors	The team where he played
cocaine addiction	His problem
Coppola	His famous best friend.
hepatitis	A new problem for him.
hospitalization	Consequence of the previous.
football; *player; legend ...*	Need no explanation.

Table 1: Collocates of "Maradona"

Of course there are also frequent but non-significant expressions, such as *rights reserved;* "empty" expressions if one considers meaning as "content". By treating them as statistic noise, there are many ways to filter these meaningless units. What reveals their condition of noise is the fact that expressions such as these appear in a great variety of documents.

With respect to the question of the validity of the propositions that relate the referent with its attributes, the only possible answer is that whether a given proposition is true or false is not a linguistic problem. It is different in the case of an ontology, which must only reflect true knowledge. There is a certain parallelism between false or inaccurate statements and ungrammatical statements: ungrammatical statements are very frequent on the Internet, but the grammatical version of them is usually much more frequent. Normally, in language there is one use that ends up imposing itself over the others, and eventually the correct form is assimilated to the one that is most used. This is referred to as the "Principle of repeated Bodges":

> *...if it's simpler and more convenient to do something in the 'wrong' way than in the 'right' way, then people will tend to do it; and if enough people do it in the 'wrong' way over a long period of time, then one of these 'wrong' ways often ends up becoming the right way. And this is how systems evolve towards the way normal people really are, end the needs and uses normal people really have; and away from the grand theories and structures created for them by their 'betters'.* (Cairns, 2007, p. 19).

Not everyone will welcome the hypothesis that the *truth* is the most extended or repeated version. Different versions, even contradictory versions, are competing against each other with different frequencies of occurrence.

2.2 Empirical Evidence in Support of the Hypotheses

Propositions in discourse are characterized by their diversity, although there also exists a space of convergence. This space includes a variety of analytical propositions that are different in surface structure but have a similar deep structure. The terms *surface structure* and *deep structure*, borrowed from early transformational grammar (Chomsky, 1957), here mean that two sentences may share the same deep structure if they convey roughly the same meaning or information, even if they are not identically formulated. Two sentences may present variations in syntax or in the selection of the vocabulary and still convey the same content. Consider statements (5) and (6), which are contexts of occurrence of the term *enzyme* with one of its most significant collocates, *phosphofructokinase* in a corpus downloaded from the web. A common conceptual structure underlies these sentences: from them, one may infer that *phosphofructokinase* is an *enzyme* involved in the process of *glycolysis*.

5) **Phosphofructokinase** (PFK) is the most important regulatory **enzyme** of **glycolysis**.

6) Different modes of activating **phosphofructokinase**, a key regulatory **enzyme** of **glycolysis**, in working vertebrate muscle.

In a sufficiently large corpus of fragments of texts where the terms *enzyme* and *glycolysis* co-occur, the frequency distribution of words also fulfills the Zipf-Mandelbrot law (Mandelbrot, 1957). Among the most frequent words (excluding grammar words) are *glycolytic enzymes*, *Phosphofructokinase*, *Hexokinase* and *Pyruvate Kinase.* There are a number of association measures that can help to refine this calculation

and reject possible frequent terms in case they are not significant, and they will be studied more carefully in Chapter 6.

The properties described so far have no relation to any particular search engine. On the contrary, they are properties of any large corpus. Figure 3 shows the case of a frequency distribution of a sample taken from IULA's LSP corpus[6] (Cabré et al., 2006; Vivaldi, 2009). These numbers come from Spanish and English medical texts, but the pattern is the same in other domains and other languages. Given the Spanish medical term *actina*, there is a set of lexical units that most frequently occur with that term inside the same sentence. This set confirms that the most frequent neighbors of a term are those that are usually mentioned in the definitions of that term in a terminological dictionary[7]:

> ***Actina* :"***proteína muscular* en *filamentos* que unida a las partículas de *miosina* constituye la actomiosina, causa de la *contracción y relajación musculares*."
>
> (**Actin**: a muscular protein in filaments, which united with particles of myosin, constitute actomyosin, the cause of the contraction and relaxation of muscles).

The correspondence between the two sets of collocates of equivalent terms, shown in Figure 3, is remarkable, even despite important difference of scale (65 vs. 418 sentences). Both terms are surrounded by the same a very similar terminology. Correspondingly, the distribution of verbs in sentences with *actina* and *miosina*, reveals that among the most frequent and uninformative verbs such as *ser*, *estar*, *poder, haber*, also the verb *unir* (to bind) is found. This verb is essential in the conceptual relation between *actin* and *myosin*, as seen in the dictionary definition.

[6] This LSP (Language for Special Purposes) corpus is a collection of scientific documents in different fields and different languages, as well as a collection of text from the press. It has more than 53 million words and is available for consultation on-line on http://bwananet.iula.upf.edu [accessed June 2010].

[7] Diccionario terminológico de ciencias médicas, 13ª edición. Barcelona, Masson, 1992.

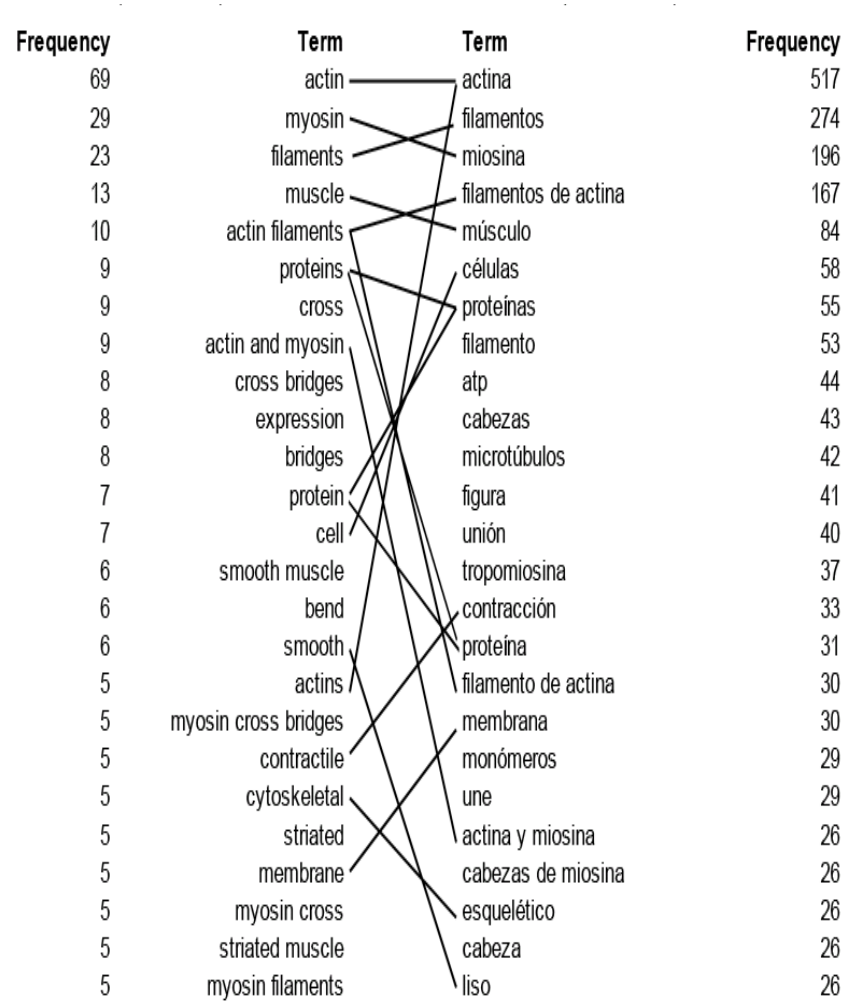

Figure 3: Most frequent *n*-grams in 65 contexts of the English term "actin" (left) and 418 contexts of Spanish equivalent "actina" (right)

It is not possible to assume that word associations are symmetric as in the above example of *actin* and *myosin.* Relations between terms can be asymmetric. This may happen in the case of syntagmatic relations -for instance, *omen* strongly predicts *good* but not the other way around- but also in a purely conceptual relation -for instance, *Renascence* predicts *Florence* but not the other way around. This is the case, too, of many hypernymy relationships and some adjective-noun pairs. Empirical evidence for this asymmetry in psycholinguistic research is provided by experiments on priming: subjects recognize faster and more accurately the word *butter* if it is preceded by *bread.* In turn, *lemon* primes to *yellow*, but not conversely. Plaut (1995) states that syntagmatic relations that

show a priming effect -considered "word associations" in psycholinguistics- can be asymmetrical while other semantically related words which do not show the priming effect are always in a symmetric relation, as it is the case with *bread* and *cake*.

The asymmetry of some semantic relations is important from both a cognitive and practical point of view. *Florence* is more relevant to *Renascence* than the other way around. This has consequences in a practical application such as a search engine. If the query is *Florence*, it is more likely -and, arguably, preferable- to obtain geographical and touristic information rather than information about a period in history. Of course, this depends on the analyzed corpus. The situation would be different if working with a corpus of historiography.

To finish this section about empirical evidence, it should be borne in mind that there exists great dependency on the corpus used to obtain the evidence and that there are risks of a possible bias because of the corpus. Problems of sampling methods that can arise with respect to the influence of the corpus. Some, for instance, may argue against the assumption of the unstructured status of the corpus under analysis. That is, if the corpus is constituted by the results of a web search engine, there is an important bias as a consequence of the particular ranking algorithm used by this engine. This argument, however, is ruled out by control study experiments carried out using the IULA's Technical Corpus (CT-IULA), which does not have any kind of algorithm for the ranking of the results. The CT-IULA uses the Corpus Query Processor (CQP), the program for the extraction of concordances (Christ et al., 1999), which can sort results only according to user parameters (e.g. internal order, alphabetically or randomly). In the control experiments that were undertaken with random sorting, no difference was noticed using the CT-IULA and using different search engines (as already shown in Figure 3). The difference in using an Internet search engine is, obviously, that there are more possibilities for querying because of the amount and variety of the available data.

2.3 Limits of the Hypotheses

There are a number of limitations to the proposed hypotheses. In first place, problems of document sampling may arise in the case of some concepts from knowledge which is still not consolidated and therefore still not well documented on the web. Another limitation is the fact that relevance is determined on the basis of subjective conditions. Finally, another limitation of the hypotheses is that they do not cover the important notion of feedback, which would be essential for a method of concept analysis based on associational learning.

With respect to the first issue, that is, the problems of sampling, it is still open to discussion whether a sample set of a hundred or a few hundreds of documents within the population of several thousands that the search engine may have in correspondence to a specific query, is or is not a fair sampling procedure. This is, though, the general problem of the representativeness of the sample, which is a recurrent issue in computational linguistics and statistics in general.

Problems of sampling may arise in a different sense: if there is not enough documentation about a subject or if the corpus gathered is poor or noisy, then it may happen that a term considered relevant or essential in relation to a given concept does not co-occur frequently enough to obtain a place in the graph of the concept. This depends on the availability of information. However, with the arrival of the web, this problem is becoming less and less important.

With respect to the second limitation, it is obviously true that the concept of relevance is grounded on subjective conditions. The same information may be relevant to some person and irrelevant to another. Even for the same person, some information may be irrelevant at one moment and relevant at a later stage. As an objection to an absolute value of information, von Foerster (1984), reporting an experiment with electroencephalography on cats, explains that a specific beep sound is meaningless until a cat learns that the sound announces feeding. In this case, meaning is objectively measured by evoked potentials. Not much can be said about this objection, and there is no doubt that the value of information is always relative. The only issue is perhaps the importance of culture. The cat can only rely on its instinctive provision and experience. Humans, however, count on a transmission of information that is detached from experience. Therefore, what is relevant in some culture may be relevant for the individual too, at some moment of life, which is to assume a correlation between relevancy and what is widespread knowledge.

Finally, with respect to not using any feedback from the environment, the decision was made to limit the scope of this thesis. Taking feedback into account in a model of associational learning would be another interesting line of research that could lead up to the development of systems that automatically learn from the failure or success of their strategies. Of course, a number of problems arise here because one should determine the ways for the algorithm to distinguish failure from success, although this would be solved most probably by human evaluation. Feedback was the basic principle in the era of cybernetics (Wiener, 1948) and represents a promising line to continue the research presented in this thesis. In any case, and as has already been stated in Chapter 1, this thesis does not pretend to be a model of human cognition because of course this would include explaining creativity or problem solving among other abilities

such as linguistic processing. On the contrary, this thesis is aimed at describing how language structures concepts in discourse, a property that can be explained in geometrical terms, which would be still consistent with some probabilistic interpretations of memory and human cognition, such as Anderson's:

> *Information is encoded probabilistically into a network of cognitive units. The nodes of this network vary in strength with practice and with the passage of time. The strength of a node controls the amount of activation it can emit [and] receive.* (Anderson, 1983:208).

Chapter 3: Basic Notions of Concept Analysis

This chapter offers a review of the work from many authors who dealt with different forms of concept analysis. More precisely, the following sections are devoted to the historical evolution of the term *concept*, an account that of course cannot pretend to be exhaustive nor detailed, since a complete treatment would be beyond the scope of this thesis. The work of authors from a wide range of disciplines would have to be consulted, amongst them, philosophers, epistemologists, psychologists, neuro-psychologists, linguists, psycholinguists, computational linguists and possibly others.

3.1 Further Delimitation of the Meaning of the Term *Concept*

The definition of concept presented in Chapter 1, although useful for purposes of this thesis, is not safe from controversy. As mentioned in the introduction, the definitions provided in this thesis rely on many assumptions. Therefore, the purpose of the following sections is to state as explicitly as possible what is understood by the term *concept*.

Without a transcendent knowledge, the question of how to define the meaning of the term "concept" can only have a tautological answer: a concept is to be defined by the relations it has with other concepts. Quine (1951) also points out that there is a tautology in presenting a definition for the term *concept*, because to be definable is an intrinsic property of concepts, and then, how is it possible to define a definition? One might object, then, that not all concepts can be analyzed, though this may lead to a new problem: if there are primitives or basic concepts that are not decomposable, from which other more complex concepts are elaborated (Wierzbicka, 1996), then one is faced with the problem of defining what a basic concept is.

One could say that such basic concept can be data from sensible experience, under the form of perception, whose atom would be the *sensation* of, for example, yellowness. Or yet some abstract, inner and undefinable notion, such as what is *good* (Katz, 1985). The notion of sensation, however, as non-analyzable piece of basic data, is inconsistent with already consolidated traditions in the psychology of perception. According to Merleau Ponty (1945), the basic unit in the cognitive plane cannot be an atom such as the sensation of yellow. Rather, it must be a contrast between that color with another one, meaning a relation such as figure and background. In other words, from a cognitive perspective the

basic unit is a relation rather than an atom. Instead of concepts in isolation, each concept is a configuration or *Gestalt* (Arnheim, 1954).

This paradox was already known by Saussure (1916): it is impossible to define an element of a system in isolation from the rest of the elements. According to Saussure, language is a system "où tout se tient", a type of solidarity of the elements of a system in which a given element is defined as not being the rest of the elements of the system. The same answer is given to the question of "what is a phoneme". The same objective sound could map to different phonemes for individuals belonging to different phonological systems. Tantamount to this is the relation of the signifier with the signified. It is not a relation of a sign with an object of a real or imaginary world. A better explanation is that the sign points to a position in the cognitive structure that allocates an abstraction of a plurality of possible instances. Yet not all concepts refer to perception or instances. Language creates the *fruit* where there are only individual apples, pears and oranges (Trujillo, 2006). Language constructs concepts to fragment the world, thus there are no concepts in reality (Wüster, 1979).

Up to this point, the comments already introduced are valid with respect to the relational aspect of the representation of concepts. Another problem is the language of the representation. The attempt to define the concept of concept by means of the same language entails an unavoidable recursion, and this is the reason to appeal to mathematics as a metalanguage. This thesis, thus, studies a quantitative aspect of the concepts which is the repetition of the combination of terms related to that concept. The thesis studies concepts as a correlate of a successive number of experiences where a pattern is repeated. In discourse, this pattern is a configuration of terms which are in a recurrent combination in the context (the immediate surrounding) of the occurrences of a term in text, a term that is used to designate a specific concept. This redundancy is an aspect that reminds us of Greimas' (1966) notion of isotopy, a beam of lexical recurrence that is characteristic of coherent discourse.

To say that a concept is the consequence of a repetition of stimuli, consisting of combinations of words in determined contexts, would be highly controversial. In fact, it may not be within the range of competence of a work in linguistics to make such a claim and is probably not in the scope of science in general (Katz, 1985). A follower of the Platonist tradition in philosophy, Katz believes that concepts have existence in themselves, with independence of the cognitive subject. Things are different from the point of view of an anthropologist such as Malinowski (1923), whose influence can be perceived in what was the Firthian tradition in linguistics, later developed as the systemic functional school (Halliday, 1994). Malinowski does not deal specifically with concepts. However, his analysis of communicative situations assumes a definition

of concepts that is akin to the one presented in this thesis. Both in primitive as well as in modern cultures, concepts are indiscernible from action and daily work, those practices that are repeated following a social convention. This is far from the Platonic idealism. Concepts depend on habits, on patterns of combination of elements that let individuals infer that their actions will have a communicative function in a certain situation.

In the Saussurean signifier-signified dichotomy, there is a fundamental and arbitrary relation between the concept and the term by which it is designated. This should not lead us to an idea of univocity, which is a one-to-one correspondence between concepts and terms, meaning that every concept is designated by a single term. This conception has been superseded by Hjelmslev (1943), who demonstrated that besides denotative semiotic systems, which relate a plane of content to a plane of expression, there are also connotative semiotic systems, where a whole denotative semiotic, including a plane of content and a plane of expression, forms the plane of expression of another content. Wüster explicitly avoids the subject of the connotative semiotic systems:

> *Terminology specialists use the expression* concept*, and not* signified*, for a basic reason: in their opinion, the signified of a term (the concept) is used up with the* denotative *signified, also known as* conceptual signified*, and disregards, in general, the connotations.* (Wüster, 1979, p. 22 [my translation]).

More recent terminology is proposed by Cabré & Estopá (2005). According to these authors, in specialized literature one can fin, on the one hand, units of specialized knowledge, analogous to Wüster's definition of concept, and, on the other hand, *terminological units* which designate them:

> *We denominate generically Units of Specialized Knowledge (USK) the units of different descriptive levels that constitute the nodes of knowledge of a text or that form part of them. The essential condition to consider that a unit is a USK is the type of knowledge that it transmits (the cognitive and semantic conditions) and its use in discourse (pragmatic and discursive conditions).* (Cabré & Estopá, 2005:77 [my translation]).

The metaphor of the "node of knowledge" stresses the idea of knowledge as a net. The function and purpose of a node in a net is the relation that it holds with the rest of the nodes. A term or terminological unit, is, in turn, "a lexical unit whose structure corresponds to an original lexical unit or is the result of the lexicalization of a phrase, that possess a specific meaning

in the field in which it is used and is necessary in the conceptual domain where it belongs" (Cabré & Estopá, 2005:77).

3.2 Historical Roots

Among the earliest references on the subject of concept analysis there is the work of Plato, who suggests that concepts (or ideas) have an objective and eternal existence (and are therefore more perfect) in comparison to the myriads of material manifestations of the universe, which are in constant flow. The material world is therefore a mere reminiscence of a perfect ideal world. This is in sharp contrast with pre-Socratic philosophers such as Heraclitus (ca. 500BC), who believed that transitoriness is precisely what is essential to the universe. For Heraclitus, one cannot step twice into the same river, objects are never identical to themselves and Plato's idealism would have probably been regarded as mere nostalgia for the absolute.

In Plato's Meno (380 B.C), for example, Socrates is impatient with a friend of his because the latter, after being questioned about the definition of virtue in general, begins to explain the different types of virtues. Socrates wants a definition of the concept of virtue, not the enumeration of its particular manifestations. He wants the *simile in multis*. What is, similarly, a figure?, he asks. There are different kinds of figures, but they have a common name and people say they are figures. Thus, what is that common nature by which one designates as figure those that are straight as well as round?

With Aristotle arrives the principle of identity and non-self-contradiction, according to which substance cannot be something different from itself. In his Categories, Aristotle offers a systematic semantic analysis where it is already possible to see the inquisitive attitude and categorization schema that is proper of the scientific spirit.

> *Expressions which are in no way composite signify substance, quantity, quality, relation, place, time, position, state, action, or affection. To sketch my meaning roughly, examples of substance are 'man' or 'the horse', of quantity, such terms as 'two cubits long' or 'three cubits long', of quality, such attributes as 'white', 'grammatical'. 'Double', 'half', 'greater', fall under the category of relation; 'in a the market place', 'in the Lyceum', under that of place; 'yesterday','last year', under that of time. 'Lying', 'sitting', are terms indicating position, 'shod', 'armed', state; 'to lance', 'to cauterize', action; 'to be lanced', 'to be cauterized', affection.* (Aristotle, Categories, Part 4).

Categories are basic constituents of propositions and are not compositional by themselves. This means that none of these categories or terms implies a true or false statement. Aristotle is among the first philosophers that categorized entities in ontologies, applying definitions by genus and species. Entities are therefore distinguished in primary substance, the instance or individual, and secondary substance, the class to which the individual pertains. Being one and the same, substance admits no contraries and no variation of degree, while it can suffer modifications in quality in the course of time.

Aristotle also distinguished those units of the lexicon that, without making reference to substance, serve only as predicates and make sense in relation to substance, for example adjectives ("superior", "double", "similar") but not only these, because many nouns ("habit", "knowledge", "attitude") as well as verbs ("to lie", "to stand", "to be seated") are relative too. They cannot stand by themselves if they have no relation to substance.

Benveniste (1966) offers a review of this work of Aristotle from a linguistic perspective and points out that these are not distinctions that are proper of the things in themselves, but a classification that stems from language (Table 2).

substance =>	nouns
quantity, quality, relation=>	adjectives
place, time=>	adverbs
state, action, affection=>	verbs (perfect, active and passive)

Table 2: Correspondence with grammatical categories.

"He thought that he was defining the attributes of objects, while he poses nothing more than linguistic beings: it is language which, with its own categories, allows us to recognize and specify them" (Benveniste, 1966, p. 70 [my translation]). This criticism is actually unjustified, because if one reads carefully the "Categories", it is evident that Aristotle is not unaware of the grammatical categories. Aristotle is going further than a grammatical distinction, one could say his point of view is trans-grammatical. In any case, it is dubious that such an enterprise is possible, thus Benveniste's comment on the limits that the grammar imposes over mind is correct in the sense that what can be thought is what can be said. From this perspective, rather than concepts, Aristotle is defining the conceptual projection of a given linguistic state. In fact, other languages have different categories. As an example, Benveniste mentions the importance given in Western philosophy to the verb "to be", and the treatment it receives as a concept like any other, when it is a particularity of the Greek language.

According to Benveniste, tables of categories are outdated because modern epistemology considers spirit as a virtuality rather than as frame, and as dynamic rather than structured. This does not imply, however, that tables of categories and their evolution in history cannot be instructive for an inquiry of concepts.

The Aristotelean categories served as the basis for the much celebrated Kantian categories:

> *It was an enterprise worthy of an acute thinker like Aristotle to make a search for these fundamental concepts. But as he did so on no principle, he merely picked them up as they came his way, and at first procured ten of them, which he called categories* [...]. *Afterwards he believed that he had discovered five others, which he added under the name of post-predicaments. But his table still remained defective. Besides, there are to be found in it some modes of pure sensibility* [...]*, and an empirical concept* [...]*, none of which have any place in a table of the concepts that trace their origin to the understanding. Aristotle's list also enumerates among the original concepts some derivative concepts* [...]*; and of the original concepts some are entirely lacking.* (Kant, 1787, p. 114).

I *Of Quantity* Unity Plurality Totality	II *Of Quality* Reality Negation Limitation
III *Of Relation* Of Inherence and Subsistence (*substantia et accidens*) Of Causality and Dependence (*cause and effect*) Of Community (reciprocity between agent and patient)	IV*Of Modality* Possibility -- Impossibility Existence -- Non-existence Necessity -- Contingency

Table 3: Kant's table of categories.

Table 3 shows Kant's list of pure concepts, those that are a priori categories of mind and not learned from experience. These are innate constrains of the mind, as time and space, necessary to apprehend reality, the faculty of judgment or thought. All possible statements derive from these four types of categories, in abstraction from their content and taking into account only their form.

I *Quantity of Judgments* Universal Particular Singular	II *Quality* Affirmative Negative Infinite
III *Relation* Categorical Hypothetical Disjunctive	IV *Modality* Problematic Assertoric Apodictic

Table 4: Kant's table of propositions.

Kant's primitives have an indisputable relevance in semantics because they are the most important philosophical precedent of the research on semantic and lexical universals, a topic which has been a recent object of study by authors such as Wierzbicka's (1996). Consider, for instance, the Kantian spirit of her collection of semantic primitives, shown in Table 5.

substantives	I, YOU, SOMEONE, SOMETHING, PEOPLE
determiners	THIS, THE SAME, THE OTHER
quantifiers	ONE, TWO, MANY (MUCH), ALL
mental predicates	THINK, KNOW, WANT, FEEL
speech	SAY
actions and events	DO, HAPPEN
evaluators	GOOD, BAD
descriptors	BIG, SMALL
time	WHEN, BEFORE, AFTER
space	WHERE, UNDER, ABOVE
partonomy and taxonomy	PART (OF), KIND (OF)
metapredicates	NOT, CAN, VERY
Interclausal linkers	IF, BECAUSE, LIKE

Table 5: Semantic and Lexical Universals according to Wierzbicka (1996).

A key moment of the philosophical reflection on concept analysis is the way in which concepts (or categories) are related to each others in propositions which can be analytical and synthetical. Analytical and synthetical propositions are always in a dialectical relation. Both analytical and synthetical propositions can be, with respect to quantity, universal, particular or singular, and affirmative or categorical with respect to quality; they can involve any type of relation, but, considering these two types of propositions with respect to Modality, the first would be apodeictic and the second assertoric. Let us consider the distinction between analytical and synthetical propositions according to Kant:

> *Either the predicate B belongs to the subject A, as something which is contained (though covertly) in the conception A; or the predicate B lies completely out of the conception A, although it stands in connection with it. In the first instance, I term the judgment analytical, in the second, synthetical. Analytical judgments (affirmative) are therefore those in which the connection of the predicate with the subject is cogitated through identity; those in which this connection is cogitated without identity, are called synthetical judgments. The former may be called explicative, the latter augmentative judgments; because the former add in the predicate nothing to the conception of the subject, but only analyze it into its constituent conceptions, which were thought already in the subject, although in a confused manner; the latter add to our conceptions of the subject a predicate which was not contained in it, and which no analysis could ever have discovered therein.* (Kant, 1787, IV,p.).

One way of distinguishing both types of statements is using the proof of negation. The negation of an analytical statement, such as (1), is self contradictory.

1) *#Bodies are not extended.*

If the terminology seems outdated, reconsider the problem in Cruse's (1986) terms, and let us talk about canonical traits of the meaning of an expression, or a degree of expectancy of such traits with respect to a concept, not compromising then on what is essential to an entity, which is not a linguistic problem. That is, if one says that a dog has four legs, one is decomposing the subject into its canonical features. The canonical status of a feature can be proved with the "but" test: it would sound odd to say "it is a dog, but it has four legs".

Other changes in the terminology of these categories of analytical and synthetical statements were introduced by Eco (1968), under the inspiration of Frege (1884) and Quine (1951). Eco uses the analytical and synthetical distinction in the field of linguistics and semiotics, calling them *semiotic* and *factual* statements, respectively. The difference is that he believes that propositions are analytical or synthetical due to existent codes and not to natural qualities of objects. Then, a sentence like */Napoleon died on 5th May 1821 on St. Helena/* was once a synthetic proposition but since then, the same utterance has become an analytical proposition, because culture has assigned to *Napoleon* the attribute *dead in St. Helena*. This introduces an important property of these categories, which is their evolution over time. As the definition of concepts evolve, some propositions that once where synthetic, like (3), now come to be analytical.

2) *#The whale is a fish.*
3) *The whale is a mammal.*

The difference depends thus on the circumstances of the utterance and the extra-linguistic conditions of the world. In principle all propositions are synthetic, since they are rooted in the experience of the speaker. However, many other statements are inferred from previously established knowledge, thus they are propositions that come from deduction and are, as a consequence, analytical.

The conceptual system goes through different crises, both at the individual level in the cases of revealing experiences and at the social level, in the case of scientific and/or cultural revolutions. The whole semiotic system is in permanent flux, and the status of a statement depends on historical accidents and a relatively stable consensus that elaborates these accidents, producing and reproducing some principles and knowledge that are supposed to be shared among the participants in communication. Strictly speaking, logic will not offer a clear distinction between these two categories.

Kantian categories had a major impact when they were published, but they begun to be forgotten a few decades later, when they were absorbed by the Hegelian system (Redding, 1997). When Hegel (1830), instead, inquires on how knowledge takes place, he rejects the separation between the individual which is the subject of cognition and the Thing-in-itself, which Kant thought was unknowable. For Kant, cognition is constrained by reduction of instances of perception to general discursive concepts. Hegel, however, rejects this separation because he considers both as moments of a dialectical unity of Thought and Being, thus reason is capable of realizing itself in the world.

Hegel examines objects of consciousness and concludes that they necessarily change into something different than themselves. In this sense they are "contradictory". Perception is first a shape of consciousness, the most immediate manifestation of being without any internal semantic structure that more elaborated concepts contain. But the mere "being" cannot be meaningful if it is not in opposition to another concept, namely, "nothing". "Being" and "nothing" are opposed and entail a third concept that is "becoming", which is now a complex concept because it can be decomposed into two moments. Perception becomes a concept through a distinction between essential and accidental properties, that is, the identity of individual substance. Concepts become judgments within larger patterns of inference, and judgments become syllogisms in an ever growing – ever developing complexity[8].

[8]It is interesting to notice that much of the psychoanalytic tradition can be read in Hegelian terms. Freud's "Totem & Tabu" (1913) is an account of neurosis presented in the same three dialectic moments: desire, repression and

Hegel argued against Kant because his conclusion would be skepticism, which would be immoral, or contradictory, if they hold to a certainty of the existence of a (yet unknowable) Thing-in-itself. He also argued against those who advocate irrationalism, asserting that they incur in self-contradiction, since it is absurd to use reason to argue against the existence of reason.

Hegel's influence begun to diminish after the nineteenth century, due mainly by critics from the analytical movement in logic (with members like G. Frege, B. Russell and K. Popper), because they claimed that Hegel's logic was not scientifically rigorous (Redding, 1997). However, the most devastating critic came perhaps from German philosophy itself: Heiddeger (1927) considered that Hegel had trivialized the work of Plato and Aristotle into a dogmatic metaphysic, disabling in this way the crucial question of Being. 'Being' had been considered the most universal and the emptiest of the concepts, and therefore a definition of being is not required since everyone uses it constantly: its meaning is taken for granted as clear and self-evident.

> *...when Hegel at last defines 'Being' as the 'indeterminate immediate' and makes this definition basic for all the further categorial explications of his 'logic', he keeps looking in the same direction as ancient ontology, except that he no longer pays heed to Aristotle's problem of the unity of Being as over against the multiplicity of 'categories' applicable to things. So if he said that 'Being' is the most universal concept, this cannot mean that it is the one which is clearest or that it needs no further discussion. It is rather the darkest of all.* (Heiddeger, 1927, p. 23).

Being is a universal concept but it is not an entity, and therefore it cannot be defined as pertaining to a class or genus. Heidegger's is an intelligent move, and it crumbles Hegel's dialectical system, in which the concept of Being ceases to be the most simple one and, therefore, it is impossible to sustain a system by using this concept as the first and most basic step to move forward and define the rest of the concepts.

---

sublimation, that will have their counterpart in the division of the subject in Id, Super Ego and Ego. Klein's (1932) most primitive concept is the nipple and its absence or negation. Because of this negation, the first idea is possible. Lacan (1956), in turn, introduces the negation (as “the name of the father”) as the entry point of the individual into the symbolic order.

3.3 Modern Semantics

When structuralist linguistics consolidated, it tried to separate itself from philosophical speculation. As a consequence, the ontological question has been forgotten. Lyons (1963) offers an overview of pre-structuralist and structuralist semantics of 18th and 19th centuries, in which the meaning of every lexical unit is determined by the paradigmatic and syntagmatic relations that are established between that unit and the rest of the units within the linguistic system. The framework is still compatible with modern semanticists such as Cruse (1986) or Miller (1995), but the structuralists stressed the fact that it is not possible to identify first the units and then, in a second phase, to inquire which combinatorial relations exist between them because both, units and relations, are identified simultaneously.

Syntagmatically related elements are elements *in praesentia,* i.e., the choices made by the speaker, while paradigmatically related elements are the elements *in absentia,* i.e., the options that the speaker has. Examples of syntagmatic semantic relations are pairs such as *bite* and *teeth*; *lick* and *tongue*; *bark* and *dog*; *bite* and *dog*; etc. In contrast, examples of units that are in paradigmatic relation can be *animate* and *inanimate*; *man* and *woman*; *up* and *down*; *go* and *come*; *mother* and *son*; *best – good – regular – bad – worst*; etc.

The kind of semantic relation that is of most interest to this thesis is the one that holds between terms and concepts. According to Frege (1892), the units that hold this relation are referential expressions (section 1.1.3). With the terms *Sinn* and *Bedeutung*, usually translated as sense and reference, he attempts to divide between expressions that point to entities of the world, i.e., the ones that have reference, and units used to predicate about them. Typically, proper nouns are used as referential units, but also indexical expressions such as pronouns or even ostensible gestures of the participants of the act of communication can have a referential function. In the current historical context, statement (4), below, has a sense that can be understood, but it has no reference since there is no king of France (Russell, 1905).

(4) *The king of France is bald.*

Reference is, as a consequence, only a property of discourse (a statement in a context), and never of lexemes out of context, because before 1789 the same statement (4) had both sense and reference.

Lyons identifies referents as entities. He even provides a classification for different types of entities. First order entities are people, animals and things, since *they exist* and are publicly visible. Of second order are events, processes, circumstances, because *they happen*. Finally, third

order entities are those which are abstracted from time and space, ideas or the entities of thought. He based these arguments on the semiotic triangle he ascribed to Ogden & Richards (1923). This triangle (Figure 4), in contrast to Saussure's signifier-signified dichotomy, puts the terms of the vocabulary in relation to notions of thought and then to objects of the world.

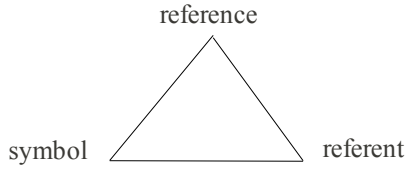

Figure 4: The semiotic triangle of Ogden & Richards (1923)

Eco (1968) criticizes this triangle as a misinterpretation that Ogden and Richards made of Peirce's work, as a consequence of the concept of reference that they had learned from Frege's (1892). According to Eco, the matter of the referent must be excluded from the studies of meaning. *Evening star* and *morning star* are, thought Frege, two expressions that have the same referent (*Bedeutung*) but different modes of meaning (*Sinn*). Eco argues, on the contrary, that for people who used those expressions, morning star and evening star were two different things, two different cultural units. In fact, there are no such things as objects until they occupy a place or a function in relation to other objects in the cognitive structure, in what Eco defines as *culture*.

> *At the moment when the Australopithecus uses a stone to destroy the skull of a monkey, there is still no culture, even when an element of nature has been transformed into a tool. Let us say that culture begins when (and we don't know if the Australopithecus meets these conditions): a) a thinking agent ascribes a new function to the stone (it is not necessary to polish it to convert it into a burin); b) the agent names it 'stone that is useful for something' (it is not even necessary to say it in loud voice or to tell someone); c) the agent can later recognize the stone as 'the object that corresponds to the function X and has the name Y'.*(Eco, 1968, p. 25, [my translation]).

This is why there is no such a thing as the referent in the original Peircean triadic theory of sign (Figure 5). The *interpretant* occupies that vertex, and is defined as the effect of the sign in the receiver, that is, in turn, another sign.

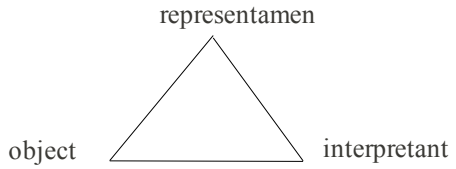

Figure 5: Peirce's (1867) original version of the semiotic triangle.

In this triadic semiosis, a symbol, or *representamen*, puts in relation the object that it represents with the *interpretant*. The object is not the referent nor a singular object. It is a set, a logical entity. The dog is my idea of the dog, an intensional definition. Peirce's sign is therefore of an unlimited circularity. The definition of referent of a sign is in terms of an abstract entity.

> *So, what is the meaning of a term? From the semiotic point of view it can't be anything else than a cultural unit. In every culture this unit is simply something that is defined and distinguished as an entity. It can be a person, a place, a thing, a feeling, a situation, a fantasy, an hallucination, a hope or an idea".* (Eco, 1968, p. 62 [my trans.]).

Also from a semiotic point of view, Peirce (1867) offers another answer. According to him, at the beginning and end of all conception, are substance and being. However, as many other philosophers have already pointed out, the conception of being has no content, since substance is inapplicable to a predicate and being is equally inapplicable to a subject. The conception of being is in every proposition. A proposition always has a term that expresses the substance and another that predicates a quality of that substance. Being is what unites the quality to the substance and quality is the first conception passing from being to substance.

> *If we say "The stove is black", the stove is the substance, from which its blackness has not been differentiated, and the "is", while it leaves the substance just as it was seen, explains its confusedness, by the application to it of blackness as a predicate.* (Peirce, 1867. CP1.549).

In the proposition *"*This stove is black*",* the conception of stove is more immediate than its blackness. The blackness is pure abstraction, as in *"*there is blackness in the stove*"*. Every proposition involves a related thing, its ground, its correlate and a mediating representation which Peirce calls the interpretant, a representation that relates a thing to its conception in the mind. Therefore, the relation of substance to being depends on the quality that is reference to a ground, a relation that is reference to a correlate, and a representation, that is reference to an interpretant. The

representamen is, as a consequence, what refers to ground, correlate, and interpretant.

The reference of signs to objects, termed "process of semiosis", can be of one of the following categories:

1st. The relation to the object as a community in some quality may be termed *Likeness* (or iconicity).

2nd. The relation to the object in a correspondence in fact may be termed *Indexical*.

3rd. The relation to the object in a conventional character may be termed *Symbolyc*.

The work of Peirce is usually not mentioned in linguistic studies, being practically ignored in the work of prominent semanticists such as Lyons (1963), Heger (1974) or Leech (1981). In subsequent debates in mainstream semantics, the concept of reference is also ignored. The concept of reference is implicit in Jackendoff (1990), when he speaks about a fundamental tension in the meaning of words between the objects of the real world and those entities that can only exist and be communicated through language. But he uses the term "construal", which suggests a constructivist idea of cognition:

> *A concept [is] a mental representation that can serve as the meaning of a linguistic expression, [...] a level of mental representation called conceptual structure is seen as the form in which speakers encode their construal of the world.* (Jackendoff, 1990: 11).

The concept is what associates a particular entity with the class to which it belongs. Jackendoff has his own list of primitives and principles of combination. A parallel with syntactic structures is evident. Syntactic structures can be of infinite variety, and also an infinite number of concepts can be invoked in the production and interpretation of sentences. This is the list of primitives, which Jackendoff calls the "conceptual part of speech": Thing; Event; State; Action; Place; Path; Property and Amount. Each of these elements forms a conceptual constituent, as Jackendoff calls them, that has a correspondence with the syntactic constituents.

8) *John ran toward the house.*

Thus, in sentence (8), *John* and *house* correspond to the constituent Thing; *toward the house* to the constituent Path, and the complete sentence to the constituent Event. This does not imply that there is necessarily an univocal correspondence between syntactic constituent, morphological category and conceptual constituent, because, for instance, a nominal phrase can be an event, a thing or a property.

The problem of reference is apparently not included in the later studies of Jackendoff (e.g., 2002). The term is only mentioned when he argues against Frege's notion of *Bedeutung* in that the entity "language" is not the same when talking to a linguist as when talking to a dentist. But reference, as said before, is a property of expressions in contexts and not of lexemes, thus it is understood that the term *expression* includes the context in which it occurs.

The point made by Jackendoff with respect to the tension between objects of a real world and entities of thought is in the center of the discussion about sense and reference. It is clear that referential units can be the names of particular objects, such as Mona Lisa, places such as the Louvre, events such as the Hurricane Katrina or persons such as Thomas Jefferson. The rest of the units of the lexicon, even when they have meaning, do not designate a particular entity. Common nouns, as said before, may have a designative function, but they refer to classes, which can in turn refer to objects; places; events; persons, etc. This division has some problems, though. A class is an entity of thought, as are memories of the Louvre or the thousands of reproductions of the Mona Lisa. Katrina is the Hurricane that took place in New Orleans, which is still a category or entity of thought. This single event in New Orleans is also a very complex and chaotic interaction of sub-events. Certainly there are well defined events, like social rituals, but there are many others, such as natural events, where it is difficult to determine the beginning and the end. Can it be said that an event is an entity? This problem is posed by Borges (1944) in his fiction "Funes, the Memorious", where he describes a man who had been injured in an accident and had lost the ability to forget. This circumstance made him incapable to assimilate concepts normally, as conceptualization is the abstraction from details and instantiations. It was not possible for him to assimilate abstractions or platonic ideas, as he was prone to consider each instance of reality as a different entity, giving thus different names to the same person perceived in different moments.

3.4 Cognitive Perspectives

Different metaphors are used in the literature to refer to the structure of mental representations. Among them, the ideas of maps and networks are among the most frequent. The metaphor of the map for human cognition has many advocates in psychological traditions such as the constructivist (Bateson, 1972) and systemic (Watzlawick et al., 1985) schools. They often use the term *map* to refer to the internal system of mental representations. These authors state that knowledge is represented in human cognition in the same manner that a map represents a territory,

which is in the line of Korzybski's (1951) general semantics and his idea that "the map is not the territory". By this slogan he means that, contrary to Aristotelean philosophy, the representation of things are not the things *themselves*. A map of internal representations is all humans have to make sense out of the experience of each individual in the world, since experiences are always fragmented and limited by the sensory, cognitive, linguistic and cultural systems.

The metaphor of the network is more recent and, inspired on neurological systems, takes the frequency effect into account. However, there is no clear understanding of the relation that exists between the frequency effect (the co-occurrence of events) and the structuring of the representation of those events in the mind. There is evidence about the correlation between lexical co-occurrence and conceptual structuring (Wettler et al., 2005). However, the correlation does not prove that the semantic memory is learned purely by repetition, and naturally, an associational network cannot pretend to be a model of learning in human cognition. It would be reductionist to regard cognition in that way. The repetition of the verbal stimuli was important in the behaviorist approach to language (Skinner, 1957), a paradigm that was shown to be suspect due to the criticism of Chomsky (1959). To admit the importance of the verbal stimuli would have contradicted Chomsky's argument of the poverty of the stimulus, one of the cornerstones of the Universal Grammar (UG) theory. This debate is the continuation of the epistemological dichotomy between rationalism and empiricism that is already in Kant's (1787) rejection of Hume's (1777) concept of causality, which was grounded merely on an association produced by the repetition of the experience of different events one after the other. For Kant, instead, causality was one of the innate categories of mind, like time and space, or, as generative linguists later added, the Universal Grammar. Nowadays, it seems clear that a research in linguistics is concerned with properties of language rather than with theories of human cognition. However, there is an increasing interest in the effect of frequency on language and cognition, particularly in the field of cognitive linguistics (Gries & Divjak, to appear) which until recently was not akin to statistical methods (Lewandowska-Tomaszczyk, 2007). In experimental psychology, particularly in the research on semantic association, statistical methods have played a fundamental role. These lines of research have existed long before their discovery by corpus linguistics and have undergone a parallel development since then (Lenci, 2008; Shulte im Walde & Melinger, 2008).

From a cognitive perspective, but from a completely different framework and terminology, the work of Langacker (1998) has posed relevant questions on how concepts are structured and networked. Langacker divides the whole cognitive structure in what he calls *things* and *relations*:

> *The terms* thing *and* relation *are used in a technical sense and defined quite abstractly. By* thing *I do not mean just a physical object, but rather anything that can be characterized as a region in some domain. When used as a noun, for instance,* yellow *profiles (that is, it designates) a region in color space.* [...] *The term* relation *is also used in a very general sense. We can think of a conceived relationship as residing in cognitive operations assessing the location, relative position, or interaction of entities within a domain. Like things, relations can stand in profile, i.e. They can be designated by linguistic expressions. When used as an adjective, for example,* yellow *profiles* [a] *relationship.* (Langacker, 1998, p. 181).

It seems that the Aristotelian terms *category* and *quality* could be used instead of nominal and relational predications. Yet, cognitive linguists (Rosch, 1975; Cuenca & Hilferty, 1999; Cuenca, 2000) are, in general, reluctant to accept Aristotelean categories to address human cognition. They reject, for instance, the idea that the cognitive system stores concepts using intensional definitions. That is to say, they do not consider the necessary and sufficient conditions that an object must meet to be allocated in memory under a specific category. Instead, cognitive linguists claim that objects (or concepts) are classified following the degree of resemblance in relation to an instance that is considered prototypical of a class. This notion involves the concept of fuzzy borders between categories in cases of non-typical units, like, say, the *duckbill platypus*. The notion of prototype is a relevant contribution here because it provides a theoretical ground to decide the relevance of an attribute of a referent even when it is not part of its intensional definition. For example, having the capacity to fly is not entailed in the concept of bird, however this capacity is an attribute of the prototypical bird.

3.5 Neurolinguistic Accounts

Loritz (1999) offers a neurolinguistic interpretation of the problem of semantic primitives that is relevant for the representation of concepts that has been sketched in this chapter. A follower of Grossberg (1986) and Piaget (1963), his argumentation is nevertheless akin to Hegel's dialectical thought: First is perception, that later becomes comprehension and then body movement, and therefore planned action with a specific objective. In this respect, thought and language are “embodied”. This idea of embodiment also has multiple points of contact with Merleau-Ponty's (1945) “flesh ontology” although it is not explicitly mentioned. The thesis

of embodiment of Lakoff & Johnson's (1980) cognitive linguistics is another ignored reference, despite the coincidence in the body-centrism.

Loritz explains how a child sees his arm moving and only then becomes conscious of the fact that he can move it. The first words are *mamamamama* and *dadadaadada*, but as soon as the child learns to walk, the coordination required for body movement is reused to organize grammar and thought. The words *mamamama* and *dadadadada* become a rhythmic dipole (off-beat / up-beat, as when walking, giving one step and then the other) and the syllables of the word *mama* and *dada* become well defined. When the child learns to walk, there is an exponential development of vocabulary and syntax. Furthermore, the discourse being linear as is the walking path, entails goals and planning.

If the first word is *mama*, the second is not *dada* but *no!*[9]. Reminiscent of Hegel's contradictions, from Loritz' perspective the negation is a fundamental moment because it causes a sudden neurochemical alteration, a crisis and a *rebound*, that will be followed by new cognitive disposition and subsequently new plans to interact with reality. Learning does not depend, though, only on rebounds. Children learn by generalization of patterns. It is not necessary to correct them when they produce sentences such as (1), because they will be more frequently exposed to different instances of the correct pattern (2).

1) #Irv poured the glass with water.
2) Irv poured water into the glass.

Water is what one usually pours, however, one can also pour glasses with other things such as ice water or *Chateau Petrus 1961,* and then, because there is new information in the sentence, examples (3) and (4) are grammatical.

3) Irv poured the glass with ice water.
4) The waiter poured the glasses with Chateau Petrus 1961.

> *Children can make this and many fine distinctions because the massively parallel architecture of their cerebrum computes these patterns with a granularity approaching 1 part in* $10^{7,111,111}$*. Discrete rules like locative movement fail to account* [examples 3 and 4] *because they deny that language can be complex to this degree.* (Loritz, 1999, p. 189).

In this passage, Loritz supports the idea that, as grammatical representations, conceptual structures emerge as a result of multiple interactions of the users of language with instantiations of terms in

[9]This is Lacan' (1956) "name of the father" (see note 38). It is the same idea exposed in Freud (1913) and Levi-Strauss (1947): the prohibition of the incest, that is, the repression of the basic instinct and the acceptance of symbolic rules, is what marks the passage of the individual from nature to culture (and to neurosis).

contexts. With independence of ontological status of a concept, meaning whether they have their own existence or they are merely a pattern of repetitions, what is undeniable is the importance that patterns of repetition have for human memory, an effective way of learning that can work even at an unconscious level (Bekinschtein et al., 2009).

Chapter 4: Varieties of Conceptual Representation

This chapter presents a group of formalisms used for the representation of concepts. There exists a plethora of knowledge representation proposals. Only those formalisms considered to be among the most successful, are reviewed in this chapter. These include Quillians' Semantic Networks, Sowa's (1991; 1992; 2000a; 2000b) Concept Graphs, Formal Concept Analysis (Ganter & Wille, 1999) the Conceptual Spaces Framework (Gärdenfors & Williams, 2001), Novak's (Novak & Cañas, 2006) Concept Maps and the Topic Maps (Rath, 1999; Park & Hunting, 2003). None of these formalisms is intrinsically related to this thesis, yet they are an antecedent of the thesis because they are ways to encode relations between nodes or objects in a non-verbal manner. More precisely, what these representation formalisms have in common with the present thesis is the idea of displaying concepts, objects and/or attributes in a two dimensional space using geometric distance as a representation of semantic relatedness. The last part of the chapter also comments on some graphic proposals for the visualization of conceptual representations which, while not formalisms on their own right, can still serve as inspiration for a graphic representation of concepts.

4.1 Semantic Networks

The first antecedent of a semantic network is likely to be the work of Quillian (1968) in Artificial Intelligence (AI). Quillian's Semantic Networks comprise information in nets of interdependent units of knowledge. The relations or connections between the nodes are labeled. Thus, for instance, one may have a node “pen” related to a node “tool” by a connection labeled “IS A”. Another node, related to “pen”, is “writing”, the connection between them is labeled “FOR”. In this network, concepts are hierarchically structured, and because of the concept of inheritance that is implicit in such structure, one can see that the semantic network proves to be a very economic system to store information. That is, one does not need to specify that a “dog” is an “animal” if it has been already said that a “dog” ISA “mammal”.

Quillian's work was well received in psychology in the sixties and seventies (Loritz, 1999), where compatible models of semantic memory were about to be developed, mainly the *logogen model* (Morton, 1969) and the *spreading activation model* (Collins & Loftus, 1975), both of them antecedents of the later connectionist model of language (McClelland & Rumelhart, 1985).

Sowa (1992) offers a complete review of all the types of existent semantic networks, where he distinguishes six different types: 1) *Definitional networks*, based on *is-a* relations, support the rule of inheritance for propagating properties from more general to more specific concepts. "Since definitions are true by definition, the information in these networks is often assumed to be necessarily true" (Sowa, 1992, p.1). In the terms of this thesis, this would be a network of analytical statements. 2) *Assertional networks*, which are based on propositions. Therefore, unlike definitional networks, the information they carry is assumed to be contingently true. This would be a network of synthetical statements. 3) *Implicational networks*, which state relations of implication by connecting nodes. Their design makes them especially appropriate to be used as a basis for inference engines. 4) *Executable networks*, including an automated procedure that uses the information of the network for a given purpose. 5) *Learning networks*, which expand their representations by acquiring knowledge from examples. As new knowledge is acquired, nodes can be added or deleted according to the weights of the connections between nodes. 6) *Hybrid networks*, which implement the different possibilities of combination of the networks mentioned above.

The formalisms that are reviewed in this chapter can be considered (or can be used to represent) both definitional and assertional networks. Apart from knowledge representation in Artificial Intelligence, semantic networks were also applied in fields such as Machine Translation, for instance in Fujitsu's ATLAS English-Japanese Translation System (Uchida, 1987).

4.2 Ontologies

Ontologies, which are intrinsically related to semantic networks, are the specification of the knowledge of a certain domain (Gruber, 2008). An ontology is an arrangement of concepts, usually hierarchically organized, and represents entities, ideas, and events, along with their properties and relations according to a system of categories. Elements of an ontology are, thus, concepts, attributes and relations, although it is common to see ontologies which also include instances. Ontologies are usually hand-crafted by human coders and are restricted to specific domains, like it is the case of the UMLS[10] and the SNOMED[11] ontologies in medicine. SNOMED, for instance, is an ontology of medical terms organized in concepts, where each concept has a unique ID number, associated with its denominations and definition. Concepts are divided in categories such as Procedures, Substances, Disorders, Organisms, etc. In each class, concepts are taxonomically structured. Several browsers for this ontology

[10]The UMLS Ontology http://www.nlm.nih.gov/research/umls/about_umls.html [accessed June 2010].
[11]http://www.snomed.org/ [accessed June 2010].

are currently available on the web[12]. There also exist ontologies of general knowledge, such as the CYC[13] ontology. In the case of lexical ontologies, the best known examples are WordNet[14] and EuroWordNet[15]. These databases organize lexical units and lexical relations in synonymy sets (*synsets*), by grammatical category, and by relations of hypernymy, hyponymy, antonymy and meronymy, among others. Wordnet is an example of ontology that includes instances. For instance, both *epistemologist* and *Socrates* appear as hyponyms of philosopher. Figure 6 shows an example of hypernomy chain provided by WordNet 3.0 for the term *keratitis*.

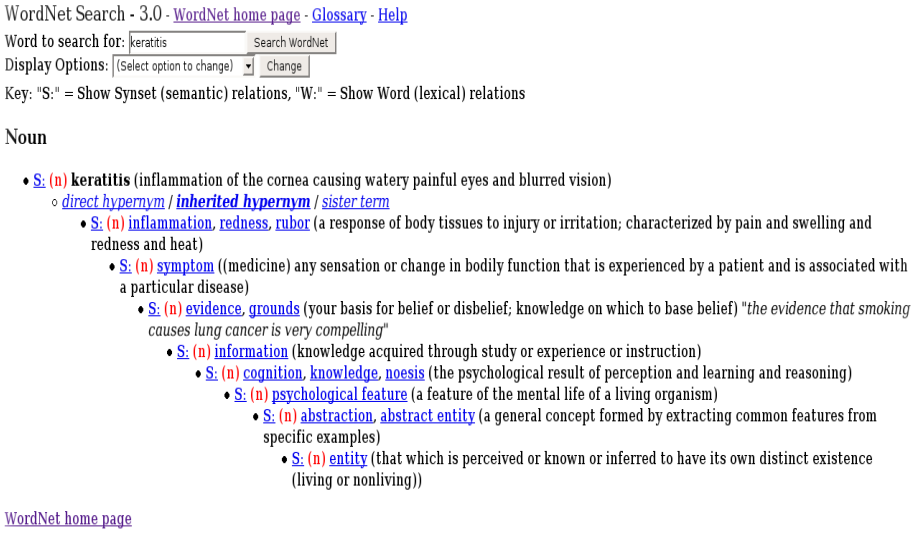

Figure 6: Screenshot of the WordNet Web Interface.

There are, currently, several attempts to extract ontologies by automatic means. Statistical methods are also used in the form of machine learning algorithms for acquiring new knowledge to the CYC ontology (Taylor et al., 2007). The idea is to obtain new knowledge automatically and to classify that knowledge into the ontology. Buitelaar et al. (2005) also report ongoing efforts towards the automatic generation of ontologies. Aguado et al. (2002) have pointed out the differences in approach that have existed in the field of ontology generation, with two parallel efforts

[12]See, for instance, the SNOMED CT Core Browser of Virginia-Maryland Regional College of Veterinary Medicine http://terminology.vetmed.vt.edu/SCT/menu.cfm [accessed June 2010].
[13]The CYC Ontology Project: http://www.cyc.com/ [accessed June 2010].
[14]http://wordnet.princeton.edu/ [accessed June 2010].
[15]http://www.illc.uva.nl/EuroWordNet/ [accessed June 2010].

in the field of Corpus Linguistics and Artificial Intelligence. Both fields have been evolving independently for decades, and therefore these authors advocate for an integration of linguistic and ontological annotation of corpus in a single hybrid formalism. Sowa (1992), who also acknowledges this split, shows however that both fields have a common linguistic precedent in Tesnière's (1959) dependency analysis, which was adopted by Schank's (1975) conceptual dependency analysis, shifting thus the emphasis from words to concepts. From this point, linguistics and Artificial Intelligence developed independently but now, with the advent of corpus linguistics, as pointed out by Aguado et al. (ibid), they are finally integrating again.

With independence of their automatic or hand-crafted origin, the role of ontologies is to be used in automatic procedures that require some sort of common sense knowledge and reasoning (Lenat, 2006). Ontologies are often used by inference engines in tasks such as query expansion in information retrieval systems (Lorente, 2005), or in information extraction, especially in a specialized domain. Some degree of common knowledge, in contrast, is needed in the case of information extraction from general domains, such as newspaper articles. The main difficulty is that to construct an ontology of common knowledge is a very difficult task:

> [Douglas Lenat, from the CYC ontology] *stated that he and his co-workers found that the encyclopedias from which they had hoped to draw "common sense knowledge" actually contain what might be considered the complement of common sense. As an example, he cited the Encyclopedia Britannica article on alchemy. Common knowledge, he said, leads us to think about alchemists changing lead into gold, yet in the alchemy article all sorts of famous alchemists and their works are described but not once is lead or gold mentioned.* (Grefenstette, 1994, p. 15).

4.3 Concept Graphs

Concept Graphs (Sowa 1991; 1992; 2000a; 2000b) are another kind of formalism for knowledge representation that has existed since the early eighties in the framework of Artificial Intelligence. Based on Peirce's (1878) Existential Graphs, Sowa's formalism has enough expressive power to convey the information of a text, or at least its logical interpretation, as expressed by predicate logic. A conceptual graph consists of nodes which refer to concepts or objects; connections between nodes indicate the specific type of semantic relation in which these nodes are engaged. Such graphs of nodes and connections are built from five

semantic primitives: existence, correference, relation, conjunction and negation (Table 6).

Primitive	**Informal Meaning**	**English Example**
Existence	Something exists.	There is a dog.
Correference	Something is the same as something.	The dog is my pet.
Relation	Something is related to something.	The dog has fleas.
Conjunction	A and B.	The dog is running, and the dog is barking.
Negation	Not A.	The dog is not sleeping.

Table 6: Semantic primitives, according to Sowa (2000a: 5).

Other logic operators can be defined based on these five primitives, such as universal quantifiers, implication and disjunction (Table 7).

Operator	**English Example**	**Translation to Primitives**
Universal	Every dog is barking.	not(there is a dog and not(it is barking))
Implication	If there is a dog, then it is barking.	not(there is a dog and not(it is barking))
Disjunction	A dog is barking, or a cat is eating.	not(not(a dog is barking) and not(a cat is eating))

Table 7: Logical operators and their primitives (Sowa, 2000a).

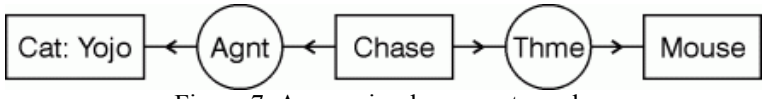

Figure 7: A very simple concept graph.

Figure 7 offers an example of the conceptual graph that corresponds to the statement *Yojo is chasing a mouse*. All information contained in such a predication is encoded in the graph. Yojo and Mouse are the arguments of Chase, the first being the *agent* and the second the *theme*. Graphs can supposedly be automatically generated from running text, especially from the scientific literature, which is homogeneously structured and stylistically constrained. To do this, one would have to go through a

process that involves different steps of syntactic and semantic-dependency analysis, identifying functional groups governed by a head unit. The semantic relations between nodes may be extracted by looking for a series of explicit markers in the text which are instantiations of the primitives shown in Table 6.

4.4 Formal Concept Analysis

Formal Concept Analysis (FCA) is a fertile field of research (Wolff, 1994; Ganter & Wille, 1999; Priss, 2006). It is rooted in the seminal work of 19th-Century mathematician Evariste Galois, who developed concept lattices as a sophisticated way to group entities or concepts in relation to their properties, considered a predecessor of the vector space model and of the variety of modern document clustering techniques.

In FCA, concepts have two fundamental elements that are objects and attributes. Objects are defined in terms of attributes such that different objects have different attributes. A concept lattice is a graph with a set of objects distributed across the space of the graph and connected to each other according to the attributes they have in common in such a way that allows to quickly identify which objects are similar, because they are clustered according to their shared attributes. Figure 8 shows an example of a concept lattice where the objects are animals that are grouped according to their respective common attributes. Attributes are typed in lower case and the objects in upper case. The object FINCH, for instance, is in that position because of the attributes 'flying' and 'bird'. EAGLE inherits those attributes but it has one more, which is being an animal which preys on others.

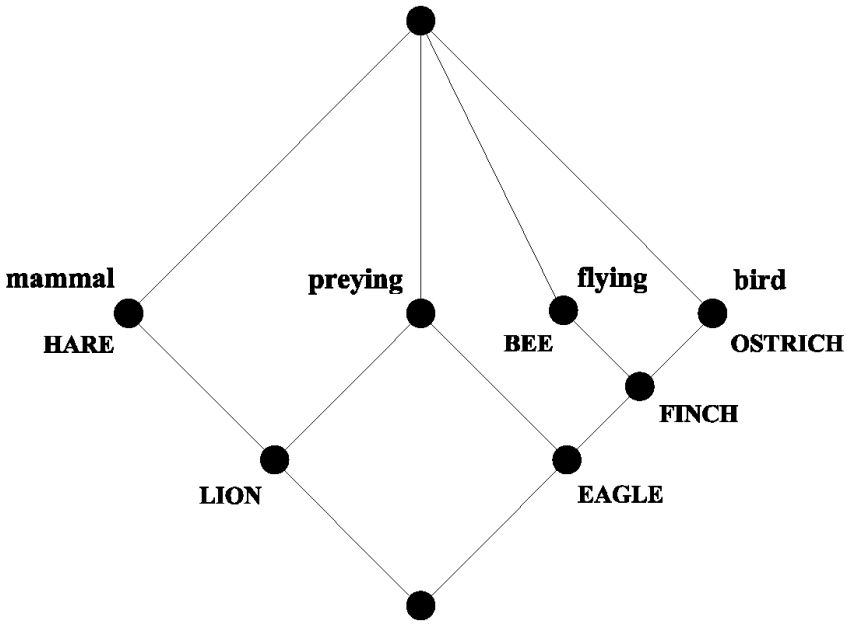

Figure 8: Example of a concept lattice (taken from Wolff, 1994).

Each node is a potential intersection of a number of objects and attributes. The height at which a node is placed in a concept lattice is also significant. Nodes that are placed at the top are the most specific concepts because they have more attributes and fewer objects associated with them. Nodes at the bottom of the lattice are more general because they have fewer attributes and therefore more objects associated with them. When there is an underspecification of the conditions that an object must meet to be categorized as a concept, more objects will be able to be associated with that concept. A normal matrix (Table 8), with the animal sets as rows and the attributes as columns, conveys the same information. In a lattice, however, information is more intuitively expressed.

	mammal	preying	flying	bird
HARE	X			
LION	X	X		
EAGLE		X	X	X
FINCH			X	X
OSTRICH				X
BEE			X	

Table 8: A matrix of animals and binary attributes.

Different researchers are using FCA lattices in areas such as lexicography (Janssen, 2002) or information extraction and ontology generation (Cigarrán et a. 2004; Cigarrán et a., 2005; Pedraza, 2007, among others). There is a certain correspondence between Sowa's Concept Graph and Ganter & Wille's Concept Lattices, in the sense that, supposedly, every Concept Graph can be translated into a Concept Lattice (Priss, 2006). Nevertheless, Sowa's formalism seems more appropriate to encode narratives, or the contents of an episodic memory (or assertional network, which would be Sowa's second type of networks), while a concept lattice is more suited to the organization of knowledge, like a semantic memory.

4.5 Conceptual Spaces

Conceptual Spaces, as presented by Gärdenfors & Williams (2001), is a framework that shows a certain degree of resemblance with FCA. However, it stresses the importance of categorization as a fundamental cognitive activity. It is a geometrical representation designed for modeling and managing concepts, with notable psychological plausibility. It involves the use of algorithms developed in computational geometry and Region Connection Calculus, which is a region-based spatial reasoning framework.

This formalism defines concepts as regions of objects that have similar attributes. One of the most interesting features of this formalism is that it demonstrates how it can be applied to problem solving such as to assign an object to a class. Given a matrix with objects as individuals and attributes as dimensions, multidimensional scaling techniques can be used to plot objects in a two dimensional space (only for visualization, because the formalism does not require a space of two dimensions) having as a result groups of objects with shared attributes. The smaller the distance between two objects in the conceptual space, the more similar these objects are. Concepts, thus, are represented as a region of that space where there are concentrations of objects. The prototypical object of a given concept is in the center of such region.

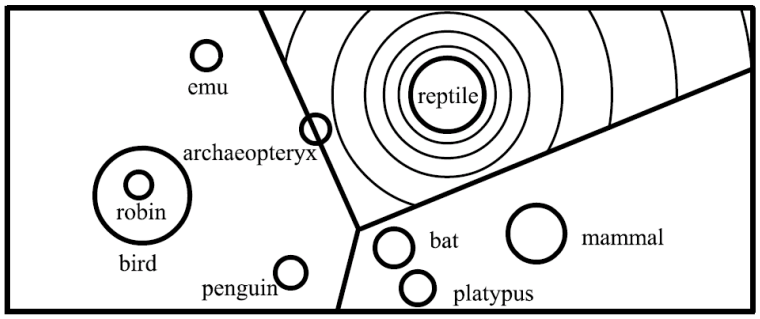

Figure 9: A concept space of animals (Gärdenfors & Williams, 2001).

Boundaries are not distinguished. Thus, regions may or may not overlap. Classifying an object using prototypes is accomplished by determining its similarity to a prototype. The representation of Gärdenfors & Williams' formalism is non-monotonic, since the progressive introduction of more objects may have the effect of a change in the categorization due to a re-localization of the prototype.

Another possible change in the representation is what the authors call *crisping*. It is the effect of blurring of a category, consisting of a progressive degradation in the distance measure. More degradation has the effect of more exemplars captured by the category, and at the same time, the effect of a successive blurring of the category. Consider Figure 9, for example. There is a conceptual space of animals with the three main categories of birds, reptile and mammals, as well as some exemplars and prototypes. The concentric circles on the category of reptiles represent different degrees of crisping. Wider circles would include more objects that are less similar to the prototype, as the 'bird penguin' is less related to the prototype 'robin' or the 'platypus' or 'bat' to the prototype 'mammal', and the presence of these words blurs the concepts.

4.6 Concept Maps

A concept map is a graph that represents the relationships between concepts. The nodes of these graphs are labeled with terms that denote concepts and the connections between nodes are labeled with predicates that relate different concepts (Novak & Cañas, 2006). The authors define a concept as a perceived regularity in the representation of events or objects, or records of events or objects, which can be designated with a label. A *concept map* is a special case of the semantic network, even if it comes from a framework unconnected to Artificial Intelligence proposals. Concept maps have become the object of growing interest as pedagogical tools (Kommers & Lanzig, 2000). They consist of the drawing of nets of

interdependent nodes labeled by terms, often nouns or noun phrases, connected by connections also labeled, but in their case with predicates, verbs or verb phrases. “Concepts are represented in a hierarchical fashion with the most inclusive, most general concepts at the top of the map and the more specific, less general concepts arranged hierarchically below” (Novak & Cañas, 2006, online).

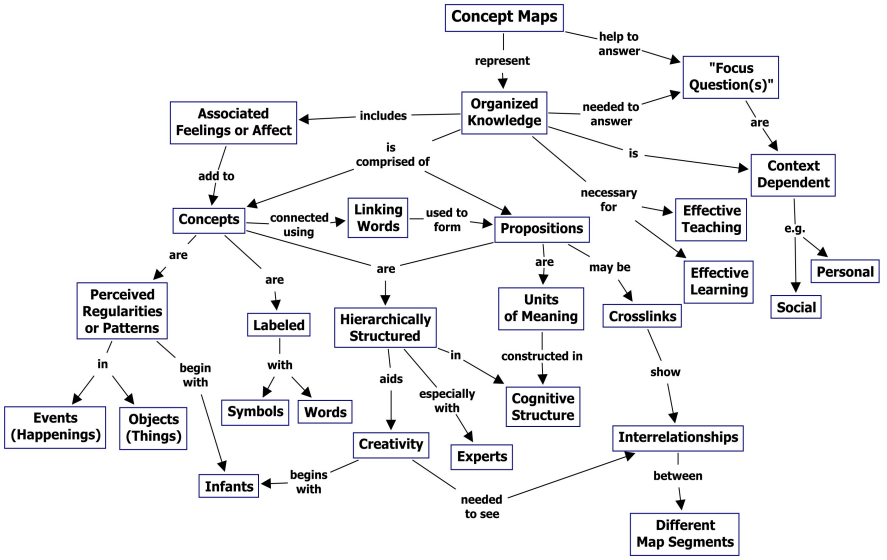

Figure 10: The concept map of the concept maps

Concept maps were conceived as a methodology for pedagogical support; students draw them manually to facilitate comprehension after reading. Concept Maps appear to be an effective method for learning and examination because of the similarity with the declarative long-term memory. This is the semantic memory, the human's structured deposit of knowledge, which plays a fundamental role in the learning process. When humans learn, they relate new elements to this memory, as a hierarchical tree with multiple entries and associations. This structure is developed by the interactive work of independent models of memory, for the interaction of short and long term memories, as it is explained in the authors own words:

> *All incoming information is organized and processed in the working memory by interaction with knowledge in long-term memory. The limiting feature here is that working memory can process only a relatively small number (five to nine) of psychological units at any one moment. This means that relationships among two or three concepts are about the limit of working memory processing capacity. Therefore, to structure large bodies of knowledge requires an orderly sequence of iterations between working memory and long-term memory as new*

> *knowledge is being received. [...] There is still relatively little known about memory processes and how knowledge finally gets incorporated into our brain, but it seems evident from diverse sources of research that our brain works to organize knowledge in hierarchical frameworks and that learning approaches that facilitate this process significantly enhance the learning capability of all learners.* (Novak & Cañas, 2006, online).

4.7 Topic Maps

Although not sharing bibliographic references, Novak's proposal shows some similarities with the "topic maps" framework (Rath, 1999; Park & Hunting, 2003), another net of nodes interconnected by conceptual relations. The difference is that the Topic Maps is a formalism born as a technological initiative to integrate and share information through a common index, which implies pretty much the same goal as the Semantic Web has (Shadbolt et al., 2006). The Semantic Web aims to add metadata to documents in order to facilitate the process of management of information in web searches. To do this, authors of web pages must add special tags, following the guidelines of the WWW Consortium[16]. Assuming that the Internet community adopts the formalism, just like it did with HTML in the nineties, the result will be an abstract network, also conceived of nets of interdependent concepts.

The Topic Maps framework was also created for the purpose of indexing collections of documents in order to facilitate Information Retrieval by semantic principles. Topic maps are similar to other formalisms because they are also networks where there are nodes or topics and arcs that establish different types of association between them. Each node is a pointer to different occurrences of this topic (especially, the term that serves as a label for a node) within other documents.

[16]http://www.w3.org [accessed June 2010].

```
<?xml version="1.0" encoding="UTF-8"?>
<topicMap
xmlns="http://www.topicmaps.org/xtm/1.0/"
xmlns:xlink="http://www.w3.org/1999/xlink">
<!-- 1st topic starts here -->
    <topic id="topic_1">
    <baseName>
    <baseNameString>Java
    Island</baseNameString>
    </baseName>
    </topic>
<!-- 1st topic ends here, 2nd topic starts -->
    <topic id="topic_2">
    <baseName>
    <baseNameString>Jakarta</baseNameStrin
    g>
    </baseName>
    <!-- Occurrence of 2nd topic: a HTML page
    -->
    <occurrence>
    <resourceRef
    xlink:href="http://www.greatestcities.com/
    Asia/Indonesia/Jakarta/introduction/"/>
    </occurrence>
    </topic>
<!-- 3rd topic starts here: the association type
"is located on" -->
    <topic id="topic_at_1">
    <baseName>
    <baseNameString>is located on
    </baseNameString>
    </baseName>
    </topic>
<!-- The association starts here, connecting the
two initial topics -->
    <association id="association_1">
    <!-- Pointer to the association type -->
    <instanceOf>
    <topicRef xlink:href="#topic_at_1"/>
    </instanceOf>
<!-- Pointers to the two topics -->
    <member>
    <topicRef xlink:href="#topic_1"/>
    </member>
    <member>
    <topicRef xlink:href="#topic_2"/>
    </member>
    </association>
</topicMap>
```

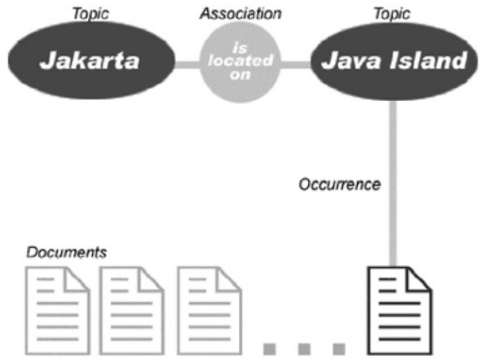

Table 9: a Topic Map with its respective source code (reproduced from Köhler et al., 2004).

Topic Maps were imposed as standard in the year 2000 (ISO/ICE 13250) being a formalism to express information content, specifying the XML characteristics in which topic maps should be encoded, in order to make information of different origin interchangeable and capable of being assimilated in major information structures. As an example, Table 9 shows two concepts, *Jakarta* and *Java Island*, which are associated by a connection also labeled as *is located on*, and each node is a collection of documents where the topic is treated.

4.8 Graphical Representations of Knowledge

In the early days of the Internet, the re-encoding of text into concept structures was already seen as the most natural way to browse the web and there were in fact some attempts to develop software for at least a manual construction of concept maps. These maps would then be integrated with web browsers (Gaines & Shaw, 1995). However, mainstream browsers and search engines did not incorporate this philosophy. There have been a great number of alternative proposals, though. Hearst (1999) provides a catalog of systems that offer different types of information visualization. Most often, these are graphical presentations in a two or three dimensional space of the results of document clustering algorithms. This means that their units of analysis are usually documents and their relations with query terms, instead of relations between terms, like was the case in previous work on hypernymy extraction (Hearst, 1992).

With respect to the general idea of encoding knowledge into graphs of nodes that represent concepts and connections that specify relations between them, in recent years a variety of proposals which show that the idea is already in people's mind have appeared. One of those examples is the Visual Thesaurus, basically a visual interface to WordNet (they both seem to output the same information) commercialized as a tool to "power one's vocabulary" (Figure 11).

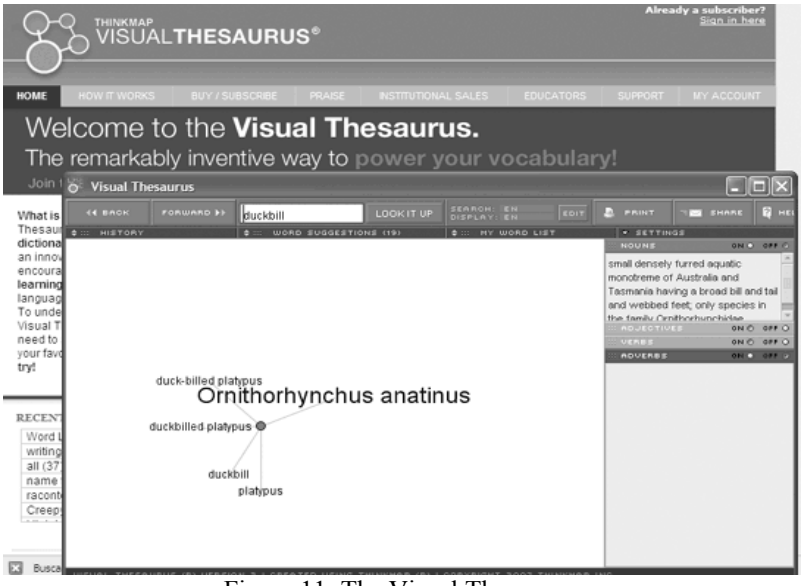

Figure 11: The Visual Thesaurus.

Shapiro (2001) shows another proposal, a very interesting visual version of Google, with a graph depicted by URL hyperlinking (Figure 12). Dodge (2004), on the other hand, offers a collection of network graphs for

the representation of information, some of them of such complexity that they exceed human readability (Figure 13). The work of Lima (2005) offers a complementary collection (Figure 14), intended to be a "unified resource space for anyone interested in the visualization of complex networks"[17].

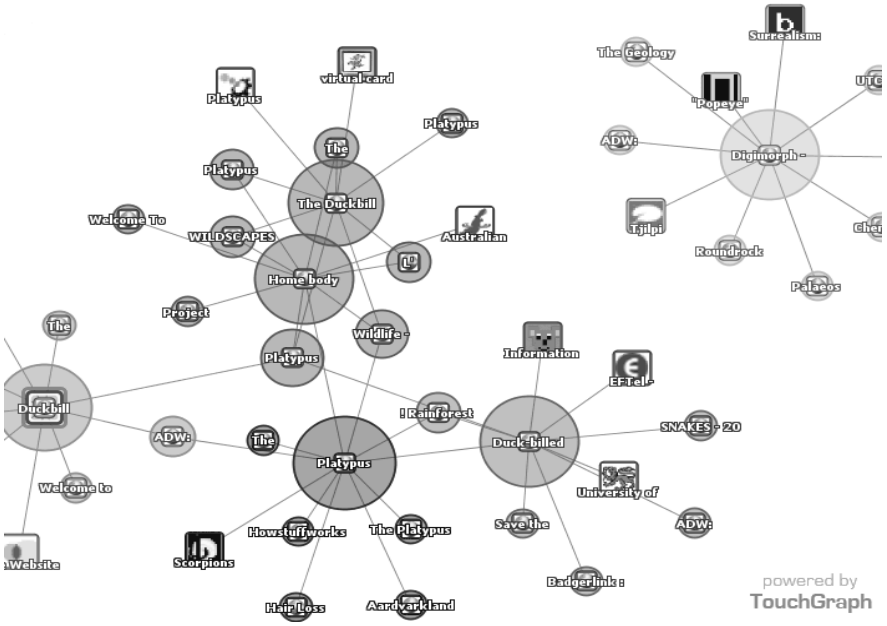

Figure 12: Shapiro's (2001) version of Google

[17]http://www.visualcomplexity.com/vc/about.cfm [accessed June 2010].

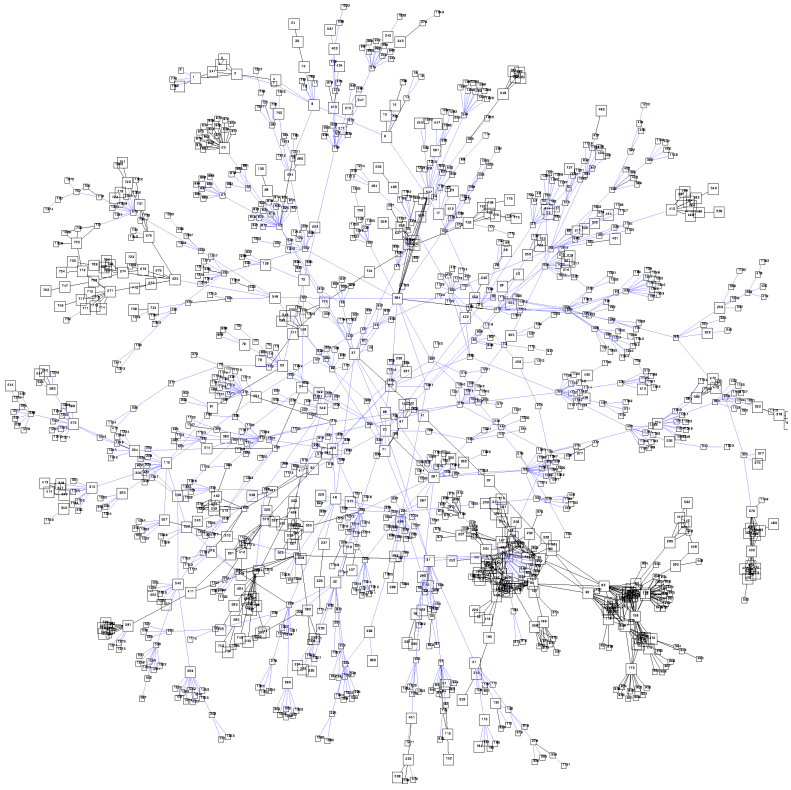

Figure 13: A conceptual topology (Dodge, 2005)

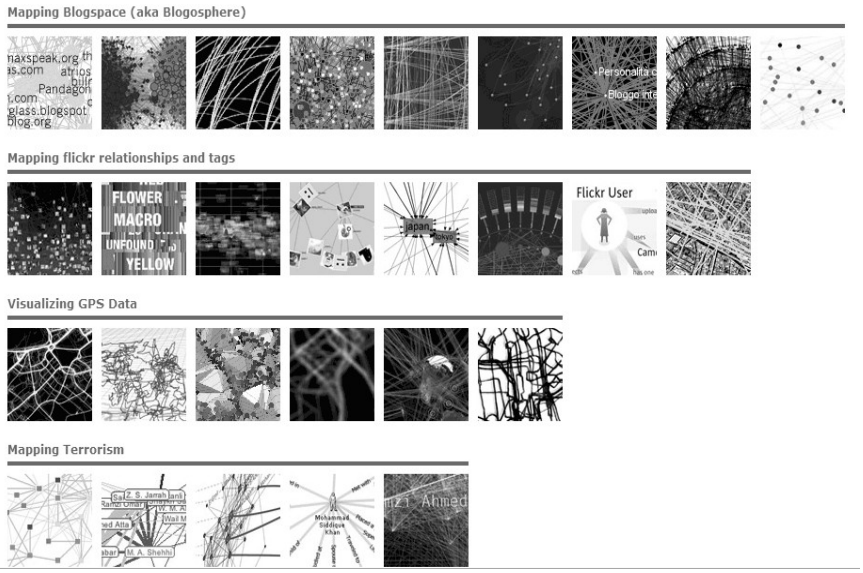

Figure 14: Lima (2005) shows a collection of knowledge networks.

Chapter 5: Automatic Extraction of Conceptual Representations

This chapter offers a summary of a variety of attempts to grasp the concepts denoted by terms by automatic means. First, the chapter examines the reasons for the efforts in this direction. Then, it comments on the main branches of research in the field and finally a more detailed examination of the symbolic and the statistical approaches. With respect to the latter, there is another distinction between authors who choose to analyze syntagmatic relations and those who choose to analyze paradigmatic relations.

5.1 Established Strategies

The studies that are most related to this thesis are those that propose the automatic generation of conceptual structures from texts. As Morin & Jacquemin (1999) point out, in this field, two families of algorithms can be distinguished: those related to terminology extraction and those related to the extraction of conceptual relations between terms. In the case of terminology extraction, it is not the case that authors interested in this field (Daille, 1994; Jacquemin; 1997; Vivaldi, 2001, among others) are also interested in the conceptual representation of a domain, it is rather the other way around: those who are interested in the conceptual representation will have to perform some kind of terminology extraction. Among the authors interested in the extraction of conceptual relations, one can find, on the one hand, some who perform an extraction of conceptual relations between terms from text based on linguistic knowledge, using algorithms with explicitly coded rules (Hearst, 1992; Lloréns & Astudillo, 2002; Feliu, 2004; Buitelaar et al., 2005; among others) and, on the other hand, those who undertake the same task from a quantitative point of view. In the last ten years, an important body of work in quantitative linguistics has been produced, especially in relation to the extraction of networks of semantic or conceptual relations between terms using statistics of lexical co-occurrence. This line of research resulted in attempts to extract thesauri by automatic means (Grefenstette, 1994; Schütze & Pedersen; 1997; Kageura et al., 2000; Curran, 2004); it has also started the study on the dynamics of the senses (or uses) of terms (Kilgarriff et al., 2004; Widdows, 2004; Hoey, 2004, 2005; Lancia, 2007) and has been applied in systems for automatic word sense disambiguation as well (Véronis, 2004; Quasthoff, 2006, etc.), including studies on the degree of polysemy of the terms within a domain (Bertels et al., 2007).

The use of statistics in the study of language is linked to the phenomenon of repetition in language. In general, authors who have shown interest in the phenomenon of repetition and the importance of its cognitive function represent a marginal trend inside linguistics. Statistical methods are often aimed at solving practical problems, but not to bolster cognitive or linguistic theories. In this sense, the present thesis has few precedents. This is also the case in the sense of its practical applications, because the specific examples of application that are shown in this thesis are not reported in previous studies. One of the main particularities of the present thesis, in comparison with related studies, is that it avoids the formulation of explicit rules of the type: *if condition X is met, then action Y is performed*. With the goal of automatically generating a graph of conceptual relations, it is possible to write a program with explicit rules that check for the presence of linguistic markers, terms or expressions, such as *...X is a type of Z...* or *...X, Y, and other Z...* Such markers (Hearst, 1992) indicate the existence of a hypernymic relation between the elements *X* and *Z*. With respect to this approach, Potrich & Pianta (2008) show, within a corpus of the domain of patents on Optical Recording Devices, very informative tables with different grammatical patterns and their reliability as predictors of conceptual relations between terms. Among a large number of other patterns, one can see that the mean precision of a pattern like (*HYPERNONYM) such as (HYPONIM)* is 61.9%, *such (HYPERNONYM) as (HYPONIM)* is 54.4%, (*HYPONIM) is a (HYPERNONYM)* is 30%, and so on.

Algorithms for the extraction of conceptual representations must include some technique for the automatic extraction of terminology, since concepts are represented by terms. The work of Kageura & Umino (1996) and Cabré et al. (2001) may serve as an introduction into the field of terminology extraction. The field is divided upon linguistically based strategies, which rely on morphological or syntactic patterns (Ananiadou, 1994; Jaquemin, 1997; among others), statistically driven strategies, usually by association of multiword units (Daille, 1994; Pantel & Lin, 2001) graph models (Aizawa & Kageura, 2001), supervised learning on manually trained data (Patry & Langlais, 2005), reference corpora of general language in order to identify *rareness* in analyzed technical literature (Drouin, 2003) and hybrid models which combine statistical models with semantic databases (Maynard & Ananiadou, 2000; Vivaldi, 2001; Sheremetyeva, 2009). There are also term extractors directly applied to bilingual terminology extraction (Ha et al.; 2008).

Leaving aside terminology extraction, the construction of concept networks requires further steps, since in a concept network terms are associated with others in order to convey information in the form of propositions. Two major trends can be seen in the literature extraction of the conceptual relations between terms. In one of them, which can be

called the *symbolic approach*, authors attempt to extract semantic relations by searching for grammatical patterns. The other approach concerns the extraction of association between terms by statistical means. This includes first order co-occurrence, also known as syntagmatic co-occurence or *contingency analysis* (Popping, 2000) and second order or paradigmatic co-occurrence, which puts terms in relation by their similarity of profiles of co-occurrence. The second approach is therefore able to relate words that do not necessarily occur in the presence of each other.

Systems which are based on statistics have an "implicit" sort of coding because the knowledge they use is not readily available at the time of implementation, but is present in the same data that is going to be analyzed. In the case of the patterns mentioned above, in an implicitly coded system it is not possible to know in advance which are the patterns to look for, but they can be discovered by statistical inference. This approach benefits from great numbers and does not need previous knowledge of the language nor of the domain in the form of dictionaries or grammar rules. Strictly speaking, explicit and implicit codings are strategies that cannot be compared. The first type of strategy is useful to extract information even when the corpus consists of a single document, while with the second it is possible to extract a representation of the consolidated common knowledge from a collection of texts.

The strategies reviewed in this chapter are, as it has already been stated, either linguistically or statistically oriented, although in recent years hybrid language-specific approaches have also begun to appear. These kind of studies are still not numerous enough, however, to devote an entire section to them. An example of this kind of hybrid approach is the work of Specia & Motta (2006), who combine different kinds of linguistic and domain specific knowledge with unsupervised corpus-based techniques for extracting semantic relations from text. Given a set of Subject-Verb-Object *seed* patterns, taken from lexical databases or manually defined, the set is enriched with new patterns found in the corpus based on similarity metrics. Among other examples of hybrid approaches are Saggion & Gaizauskas (2004), who combine the use of lexico-syntactic patterns with statistics of co-occurrence. In this case, the goal is the development of a Question Answering System, which can be considered a problem of concept analysis. The typical case they study are questions of the type "what is aspirin", and the system must be able to extract definition knowledge from text. Using a collection of lexico-syntactic patterns *à la* Hearst (1992), they are able to extract definitional as well as non definitional contexts, thus the problem is to select instances of the first type and not the other. For instance, for the question "what is aspirin", they are able to parse the question (using full linguistic knowledge, as they proceed with named entity recognition, part-of-speech

tagging and parsing) to extract the term *aspirin*, which is the term to be defined. Using the patterns, they assemble query expressions such as "aspirin is a", but these patterns will match definitional contexts such as "aspirin is a weak monotropic acid" as well as non definitional contexts, such as "aspirin is a great choice for active people". This is where the authors resort to co-occurrence statistics, following a line of reasoning similar to this thesis. They report that it is possible to select the best contexts by extracting what they call "secondary terms", which are terms that frequently co-occur with the definiendum in different sources such as Encyclopedia Britannica, WordNet or Google. Obviously, this is not the only way in which statistics can be applied to Question Answering. See Torres-Moreno et al. (2009) for an example of a different approach. These authors report high performance in a Question Answering system coupled by an Automatic Summarization System, only hybrid with respect to language-dependent pre-processing of the text, including dictionary based lemmatization. The assumption is that the correct answer of a question is in the summary of a document relevant to the question. It is, actually, a problem of categorization: given a set of sentences from the text, the one that contains the answer to the question must be selected. Leaving aside information deduced from the structure of the document (to determine if a sentence is a title or a subtitle), different statistical metrics are used, such as term frequency measures, entropy -as the probability of occurrence of a word in a text- and vector comparison of the contexts of occurrence of words, using Hamming distance. This latter metric will highlight words which are semantically related or synonyms, and if they co-occur in the same sentence, this sentence will have a higher score.

There also exist techniques which are neither related to language-specific approaches nor to quantitative corpus analysis. For instance, those which exploit hyperlinking among documents on the Internet offer a simple and straightforward way of extracting and analyzing social networks and represent an implicit form of knowledge extraction (Baeza-Yates & Castillo 2001). However, the center of interest of this chapter is in the strategies that go beyond clues such as hyperlinking and try the automatic extraction of knowledge from texts.

5.1.1 Symbolic Approaches

Attempts to extract a conceptual representation from free text using computational linguistics tools have been gaining attention since the early days in computational linguistics, in parallel with the manual effort in the elaboration of ontologies and lexical databases (Nazar & Janssen, 2010). The first efforts were centered in the automatic extraction of taxonomies from dictionaries, which started soon after the availability of the first

copies of machine readable dictionaries in the late seventies and eighties. A second trend is the attempt to extract taxonomies directly from corpus, which started in the nineties during the arrival of corpus linguistics and exploded with the advent of the World Wide Web.

The extraction of taxonomies from free text (both dictionaries or corpora) encompasses a wide range of studies from authors of different background and philosophy. Some of them imply different degrees of knowledge of a particular language, starting from strategies which use sets of lexico-syntactic patterns which typically express hyperonymy relations (e.g. "X is a kind of Y") (Hearst, 1992; Potrich & Pianta, 2008; Auger & Barrière, 2008), including combinations of different strategies. Some of them need rich semantic systems of symbolic grammar rules that use deep knowledge of a language and others even complex systems for conceptual representation like Formal Concept Analysis (Buitelaar et al., 2005; Cimiano et al., 2005).

Authors who are more frequently credited as pioneers in the field are Cazolari (1977) and Amsler (1981). Calzolari was interested in the study of the cases of circularity in definitions, which occurs when a loop is created with two definitions referring mutually or more definitions indirectly to one another by means of others. Circularity is a phenomenon that, according to Calzolari, is not a defect but a necessity in lexicography. For this objective, she was not only interested in taxonomy, but also in synonyms, since dictionaries provide synonymic definitions as well as definitions per *genus et differentia*. By iteratively linking words in the *definiendum* with the first noun extracted from the definitions, she could extract both open chains as well as chains that end in loops.

She reports an empirical study of the cases of circularity in definitions of 50,240 nouns in the Italian Machine Dictionary, and the latter with the analysis of the definitions 24,000 nouns and 11,000 verbs from the Merriam-Webster Pocket Dictionary. She extracted one word per definition, being this word the first noun found in the definitions. This word is often a synonym of the *definiendum*, in the case of synonymic definitions, or the *genus,* in the case of the definitions per *genus et differentia*. The procedure is then iterated using the extracted word as input, to find either a circularity or a definition that does not provide a noun, as in definitions of the type *who + verb phrase*. This study enables her to compare the number of open chains vs. close chains of different number of links in the dictionary.

Amsler is instead focused specifically on taxonomy extraction and presents the different types of taxonomies and the problems that such task imposes. One of the greatest difficulties in this task is how to deal with polysemous words. In fact, there is not just one problem with polysemy

but a number of different problems related with words that have more than one sense. The first problem is what happens when the word itself is polysemous and a further problem is when the hyperonym is polysemous. In the case of Amsler, he extracted chains by linking the *definiendum* by hand-disambiguating the kernel terms in the definitions before connecting the terms in taxonomic lattices. These lattices are described as transitive acyclic digraphs, or *tangled hierarchies*. This means that such lattices can be resolved in different "traces", which describe alternative paths from bottom to top nodes. For instance, it may occur that a node has two hyperonyms that do not engage in hyperonymic relations themselves, but are both hyperonyms of a third node upper in the lattice.

By these lattices, Amsler can study the structure of the dictionary and extract the semantic primitives, which are the most general elements of the lexicon and the most frequently used to define other words, thus corroborating the phenomenon of taxonomies finishing in "dictionary circles" or "loops". He also identifies other problems, such as the fact that not always the definitions are structured per *genus et differentia*, since many times nouns cannot be defined as being a type of anything and are better described as "part-whole" relations. Following a somewhat similar approach, Chodorow et al. (1985) process definitions of 40,000 nouns and 8,000 verbs to extract the genus terms to grow taxonomic trees from a specified root node. In contrast to Amsler, these authors attempt the automatic processing of genus extraction and disambiguation from the Webster Seventh dictionary. The genus term, they find, is usually the head of the defining phrase. The head can be extracted not via a full syntactic parsing of the definition but via some heuristics. In the case of verbs, the head is usually the first element following the element *to*. In the case of nouns, the head is usually the first noun in the definition after a relative pronoun, a preposition, a prep-conjunction-prep. configuration or a present participle following a noun, if the first noun is not an "empty-head" such as *one*, *any*, *class*, *kind* , among others.

These early attempts were followed by great interest of the research community and a large number of studies were presented, proposing different methods to parse the definitions of the dictionaries now not only to extract hyperonymic relations but a variety of semantic information such as *Part-of*, *Object-of*, *Location*, *Purpose*, *Manner*, *Size*, *Time*, *Agent*, *Act-of*, *Set-of*, *Inhabitant-of*, *Follower-of*, among many other relations. In the vast majority of the cases, authors applied different degrees of syntactic parsing and semantic analysis of a single dictionary using a variety of semantic formalisms (Fox et al.. 1988; Alshawi, 1989; Boguraev, 1991; Calzolari, 1992; Fontenelle, 1992; Meijs, 1992; Sanfilippo & Poznanski, 1992; Dolan et al., 1993; Barrière & Popowich, 1996).

The work of Fox et al. (1988) is a representative example of this trend. This work was aimed at the automatic construction of a thesaurus to help Information Retrieval Systems. Even though their strategy can be categorized as another symbolic approach, when comparing rule-based inference models against path-based inference they favor the second approach as probably the most appropriate approach for the use of semantic networks.

> *Path-based inference* [...] *relies only on the existence of paths (that is, concatenations of arcs) from one node to another. Conceptually, rule-based inference represents conscious reasoning from principles, while path-based inference represents (unconscious) reasoning by traversing associational links. The most natural implementation for path-based reasoning is hierarchical inheritance, but it can be applied more generally, for example to locate synonyms. In choosing related terms for expanding retrieval queries, it turns out that path-based reasoning is by far the more important. Path-based inference is more efficient than rule-based inference, because given a starting point it eliminates the need to search the network for unifying matches. That is, where path-based inference is possible the system does not have to look for rules which might apply; it needs only to traverse a very limited subgraph from a given starting point along a limited set of predefined paths.* (Fox et al, 1988, p. 105).

The methodology of these authors is to translate the information initially encoded in natural language, as it is presented in dictionaries, into a semantic network. Dictionaries may provide explicit mention of relations such as synonymy and antonymy and even some others such as hypernymy, meronymy, relations of degree, sequence (such as the days of the week), scale, as *hot* - *warm* - *cool* - *cold*, and collocations, which the authors define solely as pairs of words that co-occur frequently, such as *dig* and *hole*. This information is extracted and encoded in terms of triplets consisting of words or phrases linked by a labeled connection representing the relation. As an example they provide a semantic network for the unit *sheep.* Looking up first the definition in the *Oxford Advanced Learner's Dictionary of Current English,* they find the following data:

> **sheep**: n (pl unchanged): grass-eating animal kept for its flesh as food (mutton) and its wool.
> **wool**: n [U]1: soft hair of sheep, goats, and some other animals...
> **ram:** n I: uncastrated male sheep.
> **ewe:** n: female sheep.
> **lamb**: n 1 [C]: young of the sheep . . .

From these definitions Fox et al. extract triplets of lexical relations such as:

sheep TAX animal;
sheep TFOOD grass;
wool PART sheep;
wool PART goat;
wool TAX hair;
ram MALE sheep;
ewe FEMALE sheep;
lamb CHILD sheep;
etc.

These triplets, in turn, form the basis for the construction of a semantic network such as the one presented in Figure 15. Arcs between nodes have two labels because they have two directions, e.g. in one direction, *ram* is the male form of *sheep*, in the opposite direction *sheep* is the unmarked form of *ram,* and so on.

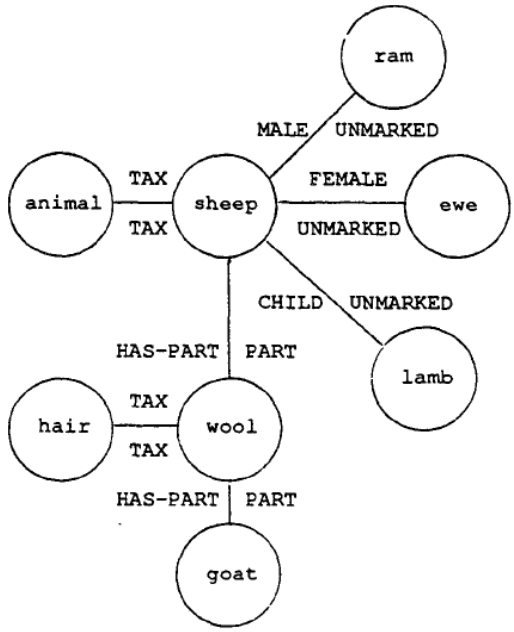

Figure 15: Conceptual representation of "sheep" (Fox et. al, 1988).

After a peak of enthusiasm on the field of MRD analysis, interest of the community began to diminish, partly as a consequence of the already mentioned explosion of corpus analysis, but also because of a series of rather pessimistic assessments of the successive efforts (Véronis & Ide, 1991; Ide & Véronis, 1993; Ide & Véronis, 1995), concluding that a fully automatic procedure for the extraction of semantic information from dictionaries with an acceptable quality was beyond the reach of the

computational linguistics of the day. In spite of this, there is a resurgence on the interest on taxonomic extraction from dictionaries in different languages (Chang et al., 1998; Renau & Battaner, 2008; Oliveira et al., 2009). Some authors have started to combine information extracted from the dictionaries with information extracted from corpus (Briscoe, 2001), and often using the web as a corpus (Velardi et al., 2007; Granitzer et al., 2009). However, it can be seen that the vast majority of the revised literature has concentrated in the extraction of semantic relations working on a single-dictionary basis, ignoring the fact that noise in the extraction can be reduced and certainty in the semantic relations can be magnified when taking into account the frequency effect of a relation that is found in more than one dictionary. Few authors have contemplated such possibility (Véronis & Ide, 1991; Sanfilippo & Poznanski, 1992; Nazar & Janssen, 2010), however, no author has attempted a completely statistical, dictionary and language independent approach.

In the early nineties, corpus linguistics became popular and a paradigm shift occurred. A trend in the extraction of semantic or conceptual information from corpora was mostly represented by the already mentioned study of Hearst (1992), perhaps one of the most cited references in the corpus of taxonomy extraction by symbolic approaches. One of the main differences with previous strategies is that Hearst does not use lexical resources but extracts the information directly from text using a limited number of grammatical patterns. Consider the example proposed by this author:

> *Surprisingly useful information can be found with only a very simple understanding of a text. Consider the following sentence:*
>
> 1) The bow lute, such as the Bambara ndang,
> is plucked and has an individual curved neck
> for each string.
>
> *Most fluent readers of English who have never before encountered the term "Bambara ndang" will nevertheless from this sentence infer that a "Bambara ndang" is a kind of "bow lute".* (Hearst, 1992, p.1).

Grammatical patterns like "such as" occur frequently across different genres and are usually not ambiguous. There are a number of other patterns that indicate hypernymy relations without the need for a detailed knowledge of the grammar of the language under consideration. Hearst provides more examples:

...works by such authors as Herrick Goldsmith, and Shakespeare.
Bruises, wounds, broken bones or other injuries...
...temples, treasuries, and other important civic buildings.
All common -law countries, including Canada and England...
...most European countries, especially France, England, and Spain.

The technique has been applied by many other authors to different languages (e.g., Rydin, 2002 for Swedish or Llorens & Astudillo, 2002 for Portuguese). There are a number of markers and grammatical relations that can be exploited. For instance, in the text: *Mazda is a kind of car*, then it can be inferred that between *car* and *Mazda* an instance-of relation holds. The same idea can be extended to identify other relations between terms. In order to do so, one has to collect a list of patterns, like *be a kind of*; *be a type of*; *be a part of*; *be equivalent to*. According to Llorens & Astudillo, these patterns can be used as a starting point for more complex domain generation procedures, and domain engineers will correct eventual errors in the taxonomies provided by the computer.

An approach that uses a greater variety of techniques and a deeper discourse analysis is Gaizauskas & Wilks (1998), within the Information Extraction (IE) framework. They developed a system called LaSIE, for the MUC-6 Conference, which was finally incorporated into the GATE System[18]. It is relevant because it is able to extract a certain form of conceptual representation from running text, known as a *template* in the IE community. Their system is complex and involves many steps of analysis:

- lexical preprocessing of the input documents
  - tokenisation
  - parts-of-speech tagging
  - morphological analysis
  - phrasal matching against lists of proper names
- parsing and semantic interpretation
  - lexical and phrasal chart edges in a feature-based formalism
    - two pass chart parsing:
      - pass one: with a special named entity grammar
      - pass two: with a general grammar
  - predicate-argument representation from sentences
- discourse interpretation from the predicate-argument representation
  - hierarchically structured semantic net with a world model
  - co-referrence resolution between new and existing instances

A notable feature to comment on, even if it is a minor feature in comparison to the rest of the system, is their strategy for named entity recognition. They use "trigger words", that is, they recognize a certain type of words that occur inside multiword proper names and give clues about what class of proper name these words denote: "*Wing and Prayer Airlines* is almost certainly a company, given the presence of the word

[18]GATE, A General Architecture for Text Engineering - http://gate.ac.uk/ [accessed June 2010].

Airlines; *Bay of Pigs* is almost certainly a location given the word Bay", (Gaizauskas & Wilks, 1998, p. 37).

Gaizauskas and Wilks use a manually constructed context-free grammar rules pertaining to named entities to recognize multiword structures. A logical form is assigned to a semantic interpretation carried out in parallel, assigning a form in predicate argument notation to each phrase identified by the grammar. For instance, the sentence *Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm* is parsed and given a logical form as shown in Figure 16.

```
person(e21), name(e21, 'Donald Wright')
name(e22), lobj2(e22, e23)
title(e23, 'executive vice president')
firm(e24), det(e24, this)
```

Figure 16: Semantic interpretation from parsing (Gaizauskas & Wilks, 1998).

The final discourse processing module integrates the semantic representation of each sentence into a model of the text. This model relies on an underlying world model which is an ontology that contains general conceptual knowledge of the domain of the text. Associated with the ontology is an attribute value structure, such as *animate:yes*. Figure 17 shows an example of what these authors call a discourse representation. According to Sowa's terminology (see Section 4.1), the representation proposed by Gaizauskas and Wilks is closer to the assertoric type of network rather than to a definitional one. In cognitive terms, it is a representation of an episodic memory rather than of a semantic memory.

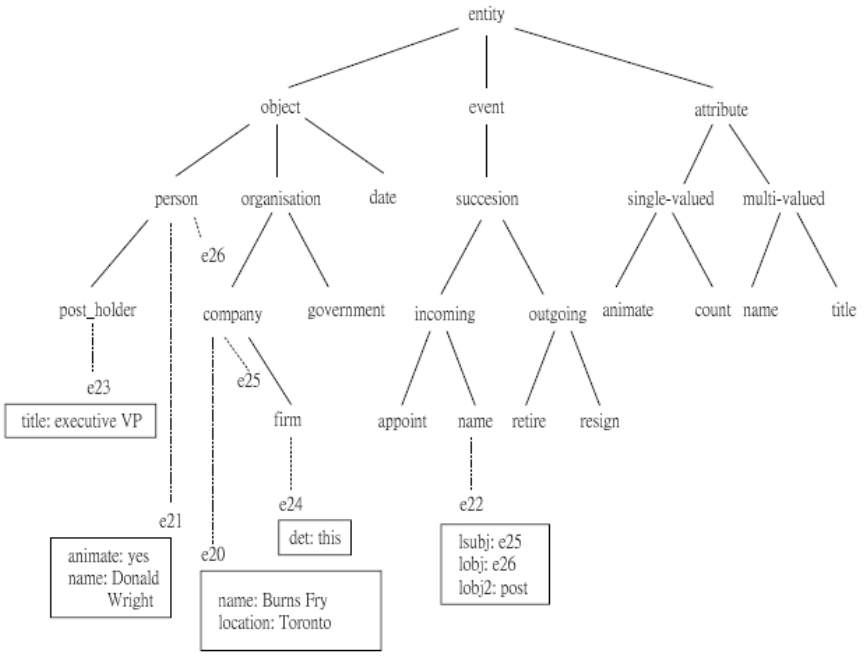

Figure 17: A discourse representation automatically generated from running text (Gaizauskas & Wilks, 1998).

There is a large number of strategies to extract conceptual relations between terms exploiting the hierarchical structure that lays beneath multiword terms with strategies similar to the one followed by Gaizauskas and Wilks. For instance, from the term *Artificial Intelligence*, one can infer that it is a kind of intelligence. This idea is used by Godby et al. (1999) and by Ibekwe-SanJuan & SanJuan (2004a; 2004b). These latter authors refer to the work of Jacquemin (1997) and Daille (1994), adding that terminology extraction can be combined with the identification of the relations among terms. Inside a collection of documents on the same topic, noun phrases and multiword terms share many lexical units. This phenomenon is referred to as "lexical overlap", and is represented in a graph structure that reproduces the schema of topics in the subject domain. Thus, given the terms *root hair*, *deformed root hair* and *root hair deformation*, in a concept hierarchy the first term would govern the other two.

The TermWatch system uses the same strategy (Ibekwe-SanJuan & SanJuan, 2004a). This system, presented as an unsupervised text-mining

algorithm used for information retrieval, is designed to offer clusters of expressions which are supposed to be semantically related. It returns a graph of interrelated nodes which are labeled with such expressions. The authors claim that the system fills a vacuum between linguistic and statistic strategies because it can capture phenomena of low frequency of occurrence. They say the system goes beyond the simple detection of association due to co-occurrence because it is able to recognize morphological similarities between multiword terms, clustering them on the basis of their shared components. TermWatch is capable of generating clusters like: *communication technology; communication technology in Nigerian print medium; communication technology in sub-Saharan Africa; communication technology infrastructure.* The cluster is governed by the head term *communication technology*.

TermWatch is able to recognize when different expressions refer to the same concept. The system offers several strategies to cope with this problem, based on patterns that give clues about the kind of relation that holds between different terms. Compounds like *wheat flour fractionation* can be paraphrased in terms of a prepositional phrase, like *fractionation of wheat flour.* In order to capture the relation between these two variants, a classification of terminological variation akin to Jacquemin's (1997) is used. There can be orthographic variations, such as *online web access / on line web access / on-line web access,* syntactic variations: *information retrieval / retrieval of information / efficient retrieval of information* and also semantic variations: *information retrieval / data access.* The system is also provided with a set of rules to relate the elements of this classification. An example of the application of such rules is the following: in a term like *access structure for similarity based fuzzy database*, the head is identified as the first noun before a preposition, that is *structure*. The rest are the left and right modifiers or expansions. The program clusters terms according to these features, offering a graph as output.

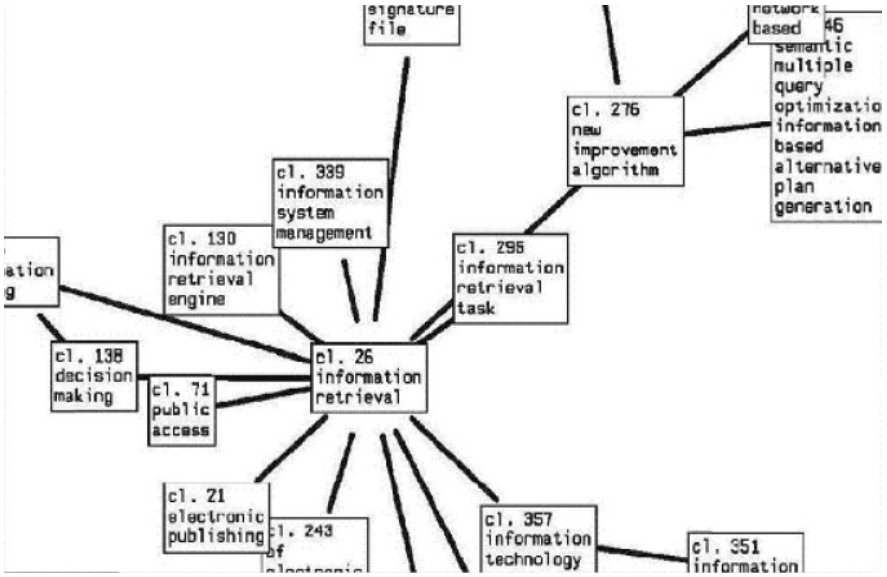

Figure 18: A screenshot from the TermWatch system (Ibekwe-SanJuan & SanJuan, 2004)

Figure 18 shows one of these graphs, with the node information retrieval occupying a central position and surrounded by other terms such as information retrieval task or information retrieval engine. TermWatch shows some points in common with the work of Morin & Jacquemin (1999), particularly in taking advantage of the implicit dependency underlying the structure of multiword terms. Morin & Jacquemin claim that their work is intended to fill the gap that exists between terminology extraction technologies and the structuring of the extracted terms as conceptual structures or taxonomies. They propose to perform a corpus-based acquisition of lexico-syntactic patterns that instantiate a specific conceptual relation (hypernym, in this case), a methodology inspired by Hearst (1992). Such patterns may be, for instance, [NP find in NP such as LIST] or [NP such as LIST]. One of the special features of their analysis is what they called Semantic Term Normalization, which is defined as the recognition of semantic variation between terms, a technique derived from previous work of Jacquemin (1997). With such a technique, Morin and Jacquemin are able to identify, given a hypernymy link between, for instance, fruit and apple, the same type of link between apple juice and fruit juice. According to Morin and Jacquemin, two terms are semantic variants of each other if they meet the following three conditions: 1) they must be semantically isomorphic, that is to say, the correlated words must occupy the same syntactic positions, being both head-words or arguments; 2) there must be a unitary semantic relationship between words, meaning that words must have similar meanings in both terms and 3) there must be a holistic semantic relationship that verifies that the global meanings of

the terms are close. They offer, as an example of variation, the pair of terms fruit composition and chemical compounds of the seed, which meet these conditions even though they do not exactly refer to the same entity, because the first one logically includes chemical compounds that may be found in the fruit but not necessarily in the seed of the fruit.

Once the relationships between terms are found, they are projected onto a hierarchy. There are two ways of doing this. The first is the *link mode:* the relation between two terms expressed with different form (synonyms or generic/specific links). The second is the *class mode*, in which semantically related terms are grouped into classes (hyponyms sharing a common hypernym). Figure 19 shows an example of a hierarchic projection.

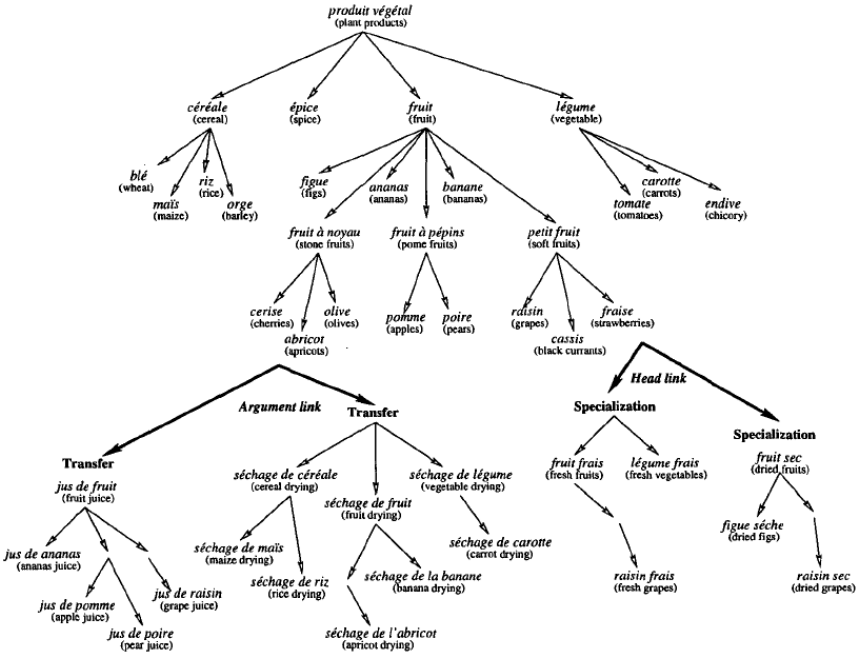

Figure 19: Projected Hierarchy of multiword terms, from Morin & Jacquemin (1999).

There are several authors whose strategies can be related to the identification of grammatical patterns. Girju (2002), for example, devotes her thesis to the identification of different grammatical patterns in which the conceptual relation of "consequence" is expressed. The work of Sierra et al. (2006) explores more general patterns of definitional contexts. A definitional context is a fragment of text in which a term is instantiated and defined. There are a series of morpho-syntactic similarities between definitional contexts in different genres and domains. A trivial example of

this kind is the pattern *defined as* or *consisting of*, which are useful to extract terms and their definitions from a corpus.

In the case of Feliu (Feliu, 2004; Feliu et al. 2006), the problem of the detection of conceptual relations is restricted to verbs. She presents a study of feasibility for a semi-automatic system for the extraction of conceptual relations from the Genomics domain, based on grammatical patterns found in the IULA's Technical Corpus. She has a fixed typology of conceptual relations hierarchically, organized in a first level in terms of: similarity, inclusion, sequence, causality, instrumentality, meronymy and association. Each one of these classes is divided into different subtypes.

Feliu argues that these basic relations can be identified by a closed list of specific markers. Thus, for the relation of similarity, there can be obvious markers such as *to be similar to*, or negative, *to be different from,* for inclusion, *to be a type of,* for spatial sequentiality, *to be in / to be in front of / to be behind,* for causality, *being the cause of / being the effect of,* for *meronymy, being a part of / an element of*. The enumeration of the different markers and their mapping to the conceptual relations goes beyond a hundred. Each marker has a distinct degree of reliability as a vehicle for the supposed conceptual relation, as some of them are ambiguous. With respect to the reliability of such patterns, recall the more recent study by Potrich & Pianta (2008), already commented at the beginning of this section.

Popping (2000) shows another classification of conceptual relations that allows him to do what he calls *cognitive mapping*, the extraction of the mental model of the individual who produced a text. Mental models are similarly presented as a conceptual network:

> *Within a cognitive map, the meaning of a concept is the aggregate set of relations it has to all other concepts that make up a conceptual network. Mental models are dynamic structures that are constructed and expanded as individuals make inferences and gather information.* (Popping, 2000: 33).

Relations between concepts can be expressed by natural language in a typical Subject-Verb-Object sentence, but there are many other possible constructions. Table 10 shows the list of semantic relations (called primitives) that Popping proposes.

- Directionality: a relation can be unidirectional or bidirectional. Transitive verbs are unidirectional.
- Strength: a relation can imply intensity, certainty, frequency, among others. The verb *to be* is stronger than *to appear,* as well as some adverbs might introduce uncertainty into it.
- Sign: can be positive or negative.
- Sense: the content of the relation.
    1. Symmetrical relations: { similarity, associations}
    2. Asymmetrical relations: { causality, hierarchy; order; evaluation; realization }
- Classification:

1.Transitive: { *it is a kind of* }
2.Asymmetrical: { *it is a property of* }
3.Symmetrical: { *inconsistent with; distinct of* }

- Structure: { *is a part of* }
- Affection: { subjective evaluations }

Table 10: Popping's (2000) table of semantic primitives.

The symbolic or pattern-based approach to concept relation extraction is still an active field (cf. the special issue of the Terminology Journal, Auger & Barrière, 2008). It must be said, though, that the strategy has limitations. The automatic extraction of taxonomies based on grammatical patterns may be less reliable when dealing with polysemous terms. Consider instances such as *viruses are programs*, *programs such as viruses* or *viruses and other programs*. Following a symbolic approach, these fragments would yield that a virus is a type of program, a conclusion that would only apply to the domain of computer science. Therefore, a necessary condition for symbolic approaches to be reliable is that the corpus from which the information is gathered is coherent and, ideally, when there is no polysemy in the terminology.

5.1.2 Quantitative Approaches

5.1.2.1 Quantitative Methods in Semantics

Lexical studies from the distributional point of view have until recently enjoyed only narrow interest in linguistics, most of the times associated with the work of Harris (1954, 1968), Firth (1957), Herdan, (1964); Halliday (1966; 1978), Sinclair (1966; 1991), Phillips (1985), Church & Hanks (1991) and Hoey (1991; 2004; 2005), among others. The scenario is now changing and many more studies on similar subjects and with similar methodology are appearing, undoubtedly due to technological advances on statistical analysis of corpora (Baroni & Lenci, 2008, for instance). Considering the technological limitations of the fifties or sixties, Harris or Halliday's accuracy of thought is now surprising:

> *If we intersect* [...] *the high frequency collocates of* moon [and sun] *we get a set, whose members include* bright, shine*, and* light, *with slightly greater generality. That is to say,* bright*,* shine *and* light *are being grouped together because they display a similar potentiality of occurrence, this being now defined as potentiality of occurrence in the environment of* sun *and* moon*.* (Halliday, 1966, p. 158).

Nowadays, the possibility of processing massive amounts of linguistic data has been complemented by new insights in the study of language. This was not fully acknowledged until recently. Gladkij & Mel'čuk (1972) declared that mathematical linguistics is not quantitative and statistics plays merely a peripheral role, whereas the term *mathematical linguistics* should only be used in the context of formal grammars, in contrast to opposite claims by Herdan (1964).

In the case of generative semanticists, they show little interest in the importance of the repetition of the verbal stimuli. Generative semanticists often associate repetition to the behaviorist approach to language. Jackendoff (2007), for instance, criticizes a paradigm based on repetition, such as the connectionist models of language (McClelland & Rumelhart, 1985). One of Jackendoff's main arguments against the connectionist program is that repetition is not an important feature in human language understanding.

> *In neural networks, long-term memories are encoded in terms of connection strengths among units in the network, acquired through thousands of steps of training. This gives no account of one-time learning of combinatorial structures, such as the meaning of 'I'll meet you for lunch at noon', a single utterance of which can be sufficient to cause the hearer to show up for lunch.* (Jackendoff, 2007, p. 25).

The comments of Jackendoff motivated an immediate reply from those he was criticizing (McClelland & Bybee, 2007), and I believe it is worthwhile to quote this reply extensively:

> [...] *Jackendoff defines productive processes as those governed by rules containing a variable—that is, symbolic rules* [...]. *Our approaches (either the connectionist approach or an exemplar plus associative network approach as in Bybee 1985, 1995, 2001) do not rely on symbolic rules, but rather postulate that productive patterns are built up from experience with exemplars of multiple types.* [...] *Jackendoff mistakes the hypothesis that symbolic rules are not needed for the claim that there*

> *are no combinatorial mechanisms at all. Jackendoff's claim* [...] *that compositionality requires symbolic rules and constraints with algebra-like variables is just a claim – one that depends on the unsupported belief that compositionality is categorical in nature. As with productivity, we view compositionality as a continuum. The evidence cited in our paper is synchronic, diachronic and experimental. There is now considerable literature finding that usage factors such as relative and absolute frequency affect compositionality (Bybee and Scheibman 1999; Haiman 1994; Hay 2001). Such continua again support the idea of a single mechanism for dealing with the whole range of phenomena.* (McClelland & Bybee, 2007, 4-7).

When McClelland & Bybee refer to productive patterns, they use the term *productivity* in a technical sense, as the potential of a pattern to apply to new forms, a form of learning that is a matter of degree rather than categorical, as in symbolic rules. The key notion in the discussion is *Graded compositionality,* the connection of the events (such as utterances *of lunch at noon*) of yesterday and the events of tomorrow, which are very similar.

According to McClelland & Bybee, the lack of understanding between the two research programs is a consequence of the history of the field: "The Chomskyan revolution of 50 years ago appeared to sweep away any basis for appealing to general purpose mechanisms, and the habits of mind set in place by Chomsky's insistence on abstract rules – and the need he and others felt to capture them using symbolic forms of computation – have been very hard to overcome" (McClelland & Bybee, 2007, p. 8).

The supposed freedom in the combination of linguistic units was the reason for Chomsky (1957, 1965) to concentrate on the study of grammar, since he believed that it is impossible to study texts because of their extreme variety. In principle, there can be as many as one can imagine. From this point of view, language may be seen as a chaotic multiplicity of local interactions of the participants of acts of communication, but when inspected from a statistical perspective, ordered and meaningful structures begin to emerge. Even when there is a potentially infinite number of utterances, there is also a grouping that is typical or normal with respect to specific concepts. Thus, it is appropriate to pose the problem in terms of quantitative linguistics:

> *...since statistical phenomena are those which result in the perceptions of mass regularities arising from large numbers of individual choices, the concept appears to apply naturally to text. Indeed, the*

> *fundamental problem of linguistic analysis is how to make general statements about language which can only be based upon finite samples of use.* (Phillips, 1985, p. 38).

Fifty years ago, a mathematician like Mandelbrot (1961) was still in position to put in question the scope of the term *linguistics*. According to Mandelbrot, there are linguists that would be better called grammarians. Linguistics, says Mandelbrot, should not be circumscribed exclusively on the determination of what is or is not grammatical, nor to elucidate which grammar distinguishes well-formed from ill-formed utterances. Reconsidering linguistics as sketched by Saussure (1916), it is the emitter who has an introspective idea of what is or is not grammatical, and that is language (*langue*), while discourse (*parole*) is the same person's view of messages encountered. Thus, there can be different grammars for every dialect, while statistical properties like frequency relationships are structural properties of all communication systems, not under control of the specific or isolated facts related to the language under consideration or the topic of discourse.

Perhaps Mandelbrot's argumentation is to be interpreted as an indirect attack on Chomsky (who is not cited) and it should be taken with certain precaution. As linguists know, grammars naturally do exist and are a legitimate object of the study of linguistics, being not a mere subject of prescriptive linguistics, at least, in the form of a consensus in language, and there are, as a result, grammatical and ungrammatical utterances.

Statistical analysis of language, without being studied in mainstream linguistics, has been an active field in the 20th century, especially by mathematicians, such as Markov (1913); Yule (1944); Shannon (1948); Zipf (1949); Mandelbrot (1957, 1961, 1983); Herdan (1964) and Muller (1973). Even though many of them are not linguists, their work is of much importance in linguistics. In specific areas such as Forensic Linguistics, particularly in stylometry and authorship attribution, statistical analysis was fundamental (Mosteller & Wallace, 1964). Interest in statistical methods in linguistics has continued particularly since Charniak (1993), and currently, statisticians have gained a firm position inside the linguistic field, as shown by the citation impact[19] of Manning & Schütze (1999). Yet modern tendencies are not exclusively statistical, and computational linguists prefer *hybrid* systems that combine statistical and symbolic approaches. As, for instance, Wanner et al. (2006) point out, in the case of collocation extraction, knowledge-poor techniques such as statistics based on frequency of co-occurrence do not provide enough information needed for collocation extraction. It is necessary to take account of the different semantic classes of collocations as expressed in terms of the lexical

[19] http://citeseer.ist.psu.edu/635422.html [accessed June 2010].

functions (Mel'čuk, 1996) of the word combinations. In a similar way, Vivaldi (2001) states that statistics based on the frequency of occurrence of terms cannot be the only source of information for the extraction of specialized terminology, rather it is better to strengthen knowledge from the combination of heterogeneous heuristics, including lexicographic knowledge, in order to decide whether a lexical unit is or is not terminological.

It should be obvious that the purpose of the research in this thesis is not to ignore qualitative insights of the work of linguists or lexicographers. The purpose of this research is, instead, to show how quantitative research, such as the statistics of lexical co-occurrence, can help us to achieve new qualitative insights. It is possible that for a specific application, other sources of knowledge would improve performance, but for a variety of problems, statistic or "knowledge poor" techniques have many advocates. Examples in the field of collocation extraction are the collocational networks based on syntagmatic co-occurrence (Phillips, 1985; Williams, 1998; Magnusson and Vanharanta, 2003; Widows, 2004), collocation extraction (Church and Hanks, 1991; Manning and Schütze, 1999; Evert, 2004; Kilgarriff et al., 2004) and the automatic thesaurus discovery based on paradigmatic similarity -under the distributional assumption that similar words appear in similar contexts- (Grefenstette, 1994; Schütze and Pedersen, 1997; Curran, 2004). Some consider that "knowledge enriched" techniques as, for instance, access to word senses collected in machine readable dictionaries, can be even detrimental in tasks like word sense disambiguation (Yarowsky, 1995; Widdows & Dorow, 2002; Véronis, 2004; Sahlgren, 2008). Authors that regard disambiguation as a purely mathematical problem are now common particularly since an influential study by Schütze (1998) in word sense disambiguation using a trainable vector-based word categorization algorithm which dispensed of external knowledge resources.

5.1.2.2 Syntagmatic Statistics

The work of the distributionalist (Harris, 1954; 1968) as well as the Firthian tradition (Firth, 1957) represent the earliest attempts to study syntagmatics with a quantitative intuition. The application of syntagmatic statistics to specific tasks such as terminology extraction has already been introduced, while also commenting on, the work of authors such as Daille (1994) and Vivaldi (2001). Measures of lexical co-occourrence are useful in general to quantify the joint probability of occurrence of two events and determine, as a consequence, if they are related events or if their co-occurrence can be attributed to chance. In the context of terminology extraction, they are often used to calculate the strength of association of the different components of a multiword lexical unit. Two events that

occur independently and with high frequency have a great chance of appearing together without meaning that their co-occurrence is related. In the opposite extreme, two relatively infrequent events that often appear together have a greater probability of being related.

There are a number of measures of statistical association that can be used to tackle this problem. One of the most widely known is Pearson's chi-square test (Formula 1). Using this measure one can appreciate the deviation that exists between some observed data and some expectation or theoretical model (Muller, 1973; Church & Gale, 1991; Manning & Schütze, 1999; Evert, 2004). In this case it will not be necessary to formulate a null hypothesis about a random combination of the lexical units to determine if two terms are or not related. Instead, it is to be expected that there will be a continuum in which distinct magnitudes register the associational strength between two units.

$$\chi^2 = \Sigma_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \begin{array}{l} O_{ij} = \text{Observed value in cell ij.} \\ E_{ij} = \text{Expected value in cell ij.} \end{array} \qquad (1)$$

	t_1	$\neg t_1$	
t_2	O_{11}	O_{12}	$= R_1$
$\neg t_2$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Table 11: A 2 by 2 contingency table.

A 2 by 2 contingency table can be used to apply this test to the study of lexical co-occurrence. In Table 11, O_{11} is the value observed in the upper left cell; it represents in this case the number of contexts where the term t_1 and the term t_2 both occur; O_{12} is the upper right cell, the number of contexts where t_2 occurs but not t_1; O_{21} is the lower left cell, and represents the number of contexts where appears t_1 but not t_2; and finally, O_{22}, the lower right cell, is the number of contexts where neither of them occurs. N is the total number of contexts and cells R and C are the marginal frequencies, from which expected frequencies are to be calculated. This is just the sum of the frequencies of the columns and rows (Evert, 2004). The expected frequencies, in order to be compared with the observed frequencies, must be computed as shown in Table 12. After this calculation, the chi-square formula is applied to the observed values, as shown in Formula 2.

	t_1	$\neg t_1$
t_2	$E_{11} = (R_1 C_{1)} / N$	$E_{12} = (R_1 C_{2)} / N$
$\neg t_2$	$E_{21} = (R_2 C_{1)} / N$	$E_{22} = (R_2 C_{2)} / N$

Table 12: Marginal Frequencies.

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \quad (2)$$

Another statistical measure of association known as Mutual Information (Church & Hanks, 1991; Manning & Schütze, 1999; Evert, 2004) is a by-product of Information Theory used to determine how informative the occurrence of an event *i* is once another event *j* has occurred (Formula 3).

$$MI(i,j) = \log_2 \frac{P(i,j)}{P(i)P(j)} \quad (3)$$

In an extreme case, the highest mutual information would be that *i* always occurs with *j*. In the opposite, that *i* occurs on average the same number of times irrespective of the instances of *j*. As with most statistical measures, its scores are not reliable when the absolute frequency of its collocates is very low (under 5).

There is a great variety of statistical alternatives to measure the associational strength of two lexical units. These statistics have their neural correlate: in discourse, the occurrence of a word provides more or less information about the occurrence of others. Associational strength has been correlated with experiments on semantic priming that psycholinguists have been reporting on for decades (Anderson, 1983; Aitchinson, 1987; Plaut, 1995; Loritz, 1999; among others). That is, in the context of the word *Barcelona*, the word *Spain* is also likely to be found, among other words that are conceptually related as well. It can be said that the word *Barcelona* activates the word *Spain* in my memory, but there are other words that do not have this property. Some words do not evoke or predict others; for example function words or predicates that appear frequently and in a great variety of contexts. This is the reason why these units of the vocabulary that are said to be uninformative.

By definition, statistics on Mutual Information are symmetric. That is, the event *i* has the same information with respect to *j* than the other way around, which does not happen with other measures, such as conditional probability. However, as it will be shown later, conceptual relations

between terms are not necessarily symmetric; an example of this is the relation of hypernymy between terms.

As already stated, it is also possible to find second order symmetric associations. That is, among units that have a similar profile of occurrence, such as the words that are synonyms or that engage in paradigmatic relations (Harris, 1954; Halliday, 1966; Grefenstette; 1994; Schütze & Pedersen; 1997; Curran, 2004; Lancia, 2007; Sahlgren, 2008). An opposition of this type would be the pair *oculist-ophthalmologist*, because these units share part of their profile of co-occurrence. It is a second order relation because these units do not necessarily appear in the same syntagm.

Another application of the statistical apparatus to distributional studies that is relevant to this thesis is the work of Phillips (1985), who argues in favor of a dynamic conception of lexicography and against stored lexical resources, which entail a static idea of meaning. According to Phillips, the meaning of a lexical unit is shaped by the text and context where it is instantiated. The key concept of Phillips' work is the "aboutness". The aboutness is the essence of the content that the reader keeps in mind after reading. This concept reminds us of Greimas' (1966) *semantic isotopy*, as a countenance of recurrent semantically related categories, although Phillips dos not cite him explicitly. Phillips refers instead to the concept of macrostructure, as described in van Dijk (1983), who studied Greimas extensively. Phillips is very critical of the work of van Dijk, stressing the idealism of his discourse analysis:

> *It is clear that although the approach raises pertinent questions, it is ill-conceived as a means of answering them. I accept that the basic insight into macrostructure is valid but draw the conclusion that any attempt to investigate it must look for evidence in the concrete patterning of linguistic substance rather than hope for insights to be generated by the abstractions of text theoretical apparatuses.* (Phillips, 1985, p. 20).

Phillips advocates data-driven quantitative text analysis, and he focuses on the statistics of the lexis, which are, according to him, essentially syntagmatic. Through co-occurrence of items separated by a small number of words, he extracts groups of lexical items that accomplish a structural cognitive function in text. Therefore, his model is presented as a way to extract semantic structures from the text's linear sequence. The concept of *circuit*, according to Phillips, is the environment consisting of concepts such as *amplifier, components, use, input, voltage, transistor,* and so on. Phillips does not intend to substitute dictionaries with this methodology, however it is clear that collocational data, defined in his way, can provide valuable information to complement dictionaries.

Phillips uses methodological categories such as interval, unit and span. The interval is the space in which two lexical units are supposed to co-occur. The interval should not be the whole text, especially in his case, because he conducts his statistics on a text-by-text basis. An interval is, ideally, a paragraph, but can be an arbitrary number of words. If the interval were too short, then it would overlap with the category of span. In his methodology, a span is a distance of four words (defined on an empirical basis). In a single interval of text, there should be more than one span. The units under analysis are single words, avoiding the treatment of multiword terminology. Collocations are defined as words that co-occur within the same span of text. Phillips is aware of the ambiguity problem of the term collocation and attempts to clarify his use:

> *...it should be noticed that I am using the term 'collocation' in a more abstract sense than is often the case. 'Collocational meaning' is not infrequently used to refer to that aspect of meaning which arises from the idiosyncratic association of words in stock phrases consecrated by the language. A typical example would be the fact that 'old people's home' is acceptable while, in normal circumstances, 'old people's house' is not. I am not, however, interested in such ad hoc facts about English. In this study the term 'collocation' is used to refer to the general event of co-occurrence within a specified span.* (Phillips, 1985, p. 63).

Phillips did not consider sentence boundaries relevant to the co-occurrence with the argument that collocational patterning transcends them. Word order and the syntactic structure of the sentence is also ignored. He does little preprocessing, but includes a stoplist consisting of cohesion elements and other elements of the vocabulary which are not semantically focused, such as *however*, *furthermore*, *nevertheless*, etc. He includes lemmatization but without word sense disambiguation, because disambiguation will emerge naturally on the basis of distributional features.

Collocational data is represented by a matrix that contains each lemma as a node in the rows and again each lemma but this time as collocate in the columns. The content in the cell in row *i* and column *j* is the frequency of the collocation of *i* and *j*, normalized to the range 0-1.

$$w(i,j)=\frac{f_{ij}}{\sqrt{f_i f_j}} \qquad (4)$$

The coefficient presented in Formula 4, where f_i is the frequency of a word *i* and f_j is the frequency of lexical unit *j* prevents the possibility of having the low frequency collocations swept by elements of the

vocabulary which have a high frequency. The matrix is supposed to be symmetrical above the leading diagonal, but then Phillips notices that the symmetrical property may be misleading since, as it was already stated in this thesis, collocational data can be directional or asymmetric. Another problem of the matrix, due to its quadratic complexity, is the cost of memory and speed of processing (especially in the eighties). This limitation lead him to use a matrix of only 200 x 200 items. Considering that in a text of 60.000 running words one may have vocabulary that ranges from 3000 or 7000 units, this means that some selection process must be undertaken. Phillips chooses lexical items by random sampling, hoping that the frequency distribution in the sample will represent the frequency distribution of the population from which it is drawn.

With this matrix at hand, Phillips does different kinds of analyses, most importantly clustering of the nodes with the collocates as features. Clustering is expressed in dendograms, but he transposes them into directed acyclic graphs, having as nodes of these graphs the nodes of the initial matrix, and connecting them according to the same similarity metric shown in Formula 4. Because he is interested in texts, he wants to represent their macrostructure (in the sense of van Dijk, 1983) in terms of the graphs, and can also show the relations that hold between different chapters and even detect how coherent a text is.

Phillips has had a notable influence on later studies such as Grefenstette (1994), Kilgarriff, (1997), Kilgarriff et al. (2004) and particularly Williams (1998) and Magnusson & Vanharanta (2003). Williams (1998) proposes collocational networks with the same motivation, to offer a dynamic view of language in which the meaning of a lexical unit is simply the product of a context of use, that is, a specific lexical environment. The notion of collocation with which he constructs these networks is the same as in Phillips, that is, the occurrence of two units at short distances in text. One of the difference is that he uses Mutual Information like Church & Hanks (1991). According to Williams, collocational data, presented as "chains" of words, can reflect the underlying conceptual framework of a domain. Following these collocational chains is a way "to isolate the frame of reference of a given item within the lexis of a given discourse community". (Williams, 1998, p. 156). Another study in the same line, but from a diachronic perspective, is that of Magnusson & Vanharanta (2003). These latter authors also register patterns of lexical co-occurrence but with the goal of finding transformations in the concepts across a chronologically ordered corpus. In this case, they consider the co-occurrence of two words appearing together in the same documents within a group of financial reports of the Ericsson company. Collocational networks allow them to detect patterns of conceptual variations across a temporal axis of documents faster than plain reading and without human

effort. These collocational networks (Figure 20) are supposed to be useful for experts in the domain of the corpus.

> *This approach gives us an opportunity to visualize the concepts that are emphasised in a particular text. These concepts are reflected through the words that constitute the nodes of the network. This approach also gives us a possibility to examine which concepts are more frequently linked to each other.* (Magnusson & Vanharanta, 2003, pág. 278).

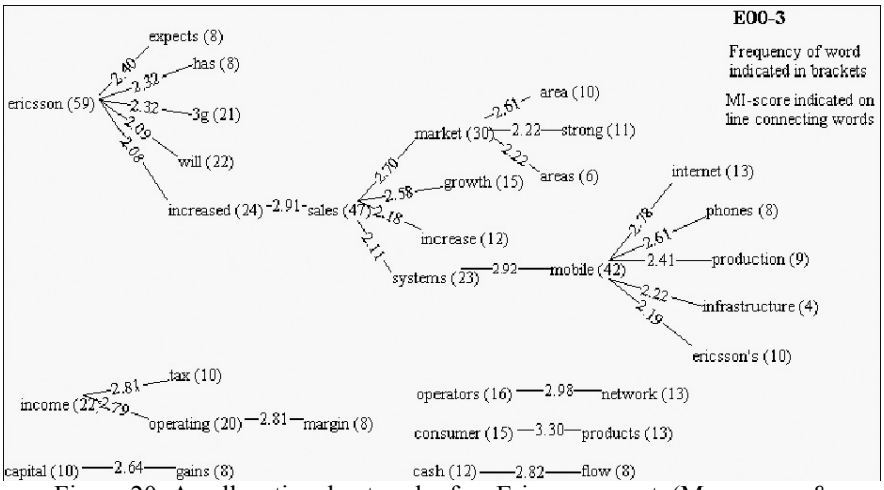

Figure 20: A collocational network of an Ericsson report (Magnusson & Vanharanta, 2003).

Attempts that follow a similar philosophy and techniques are those of Böhm et al. (2004) and Maicher (2004). In their case, co-occurrence data is used with the goal of automatic generation of topic maps from unstructured collections of documents. Both detect topics and their respective relations in the texts that feed their systems. Terminology extraction is their first phase. It is done with the help of a user through a process called relevance feedback: when a user selects a set of terms as relevant, the system will consider that terms similar to the selected ones are also relevant. The similarity measure is based on character trigrams. For example, a rare trigram like "cyt" appears often in specialized terms of a medical corpus, like *thrombocytopenia* or *cytomegalovirus*. If these terms have been selected as relevant, new terms with this trigram will be considered relevant, as well. In the case of multiword terms, the strategy is to inspect whether or not the term presents a typical terminological syntactic pattern, or if these terms share a component with others selected as relevant.

Maicher (2004) also reports extraction of terminology with statistics of co-occurrence: he identifies multiword terms if they occur together with a significant frequency inside a corpus. Maicher used frequency data from a reference corpus of general language that provides the expected frequency of a word with non-specialized meaning. In contrast to the reference corpus, if two terms show a significant co-occurrence in the text under analysis, they are assigned related nodes in the topic map. This association between terms is drawn from the co-occurrence in both a sentence and a document scale. The fact that the design involves a reference corpus implies that the method can be used even with small collections of documents. According to his reports, the minimum corpus size needed to obtain results, is of 17,000 words.

It should be noticed that most of the authors reviewed so far in this section do not seem interested in the identification of the types of semantic relations between nodes. In the case of Maicher, for instance, the goal is limited to the indexation of terms with the goal of mechanically assisting a user to merge topic maps created from different sources, or enriching them with running text.

The work of Köhler (2004) shows characteristics similar to Maicher's, stating that significant terms show a recognizable distribution in collections of documents. He exposes the following sequence of procedures: 1) indexation of texts; 2) extraction of multiword terminology; 3) weighting of each term; 4) search for association between terms by co-occurrence. Köhler refers to the work of Attar & Fraenkel (1977). These authors were, apparently, the first to notice that synonyms can be defined as pairs of terms that often occur in the same documents. Note that this is different than to relate synonyms as words that share part of their profile of co-occurrence. Attar & Fraenkel's explanation for this phenomenon is stylistic correction. Depending on the genre of the text, authors tend to avoid the repetition of words at short distances in a text. As a consequence, a document that contains the word *automobile*, is likely to contain also the word *car*. Co-occurrence statistics are done, again, with term by term matrices, like the fragment presented in Table 13. The terms *map* and *representation* are highly associated, thus, it may be inferred that their meanings are correlated.

	topic	map	knowledge
map	0,53		
knowledge	0,21	0,14	
representation	0,06	0,67	0,32

Table 13: Another example of a text x term matrix.

The same trend is followed by Quasthoff et al. (2006) in the Wortschatz system[20], which also offers collocational data in the form of networks of lexical co-occurrence of most European languages. Texts from the Wortschatz corpus are not linguistically preprocessed. However, just from the co-occurrence data, it is evident that nodes are semantically related and the topics represented by them are relevant. Figure 21 shows an example with the entry *Khartoum*.

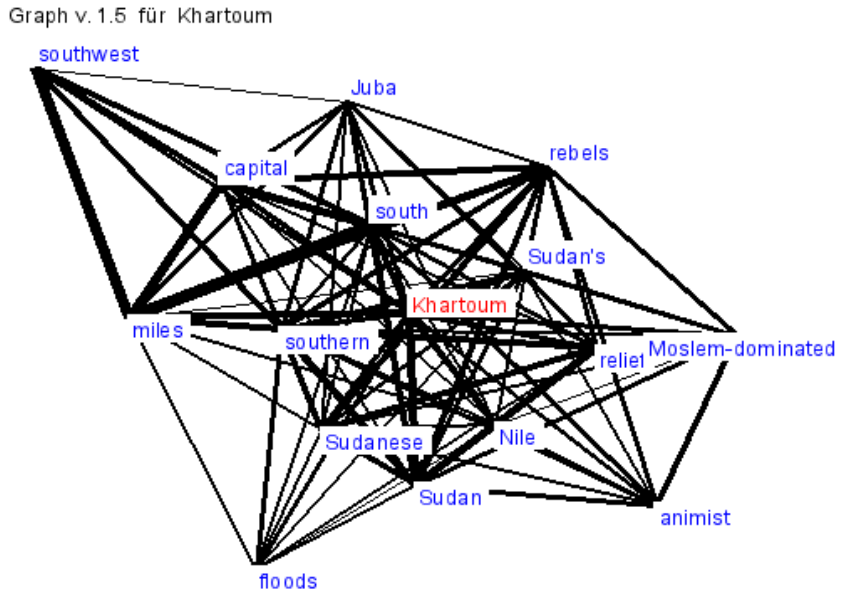

Figure 21: The collocational network of "Khartoum", according to the Wortschatz system.

It would have been interesting to include multiword terms in the model of Quasthoff et al. (2006), especially from the point of view of terminology. This adjustment could be a useful optimization, at the cost of the exponential vocabulary growth. Another variable that would surely benefit the results would be to incorporate a threshold of frequency in their statistics, because they appear to be too sensitive to low frequency data. In practice, this means that one or two contexts of occurrence of an expression can have a decisive impact on the resulting network. Finally, morphologically similar words are not identified or associated (*Sudan*, *Sudan's*, *Sudanese*), yet this would benefit the frequency counts.

Quasthoff's networks shows some resemblance to Chen's automatic generation of *Concept Spaces* (Chen et al., 1996). Concept spaces are clusters of semantically related terms, computed from a coefficient of co-

[20]http://corpora.uni-leipzig.de [accessed June 2010].

occurrence, with the goal of representing the state of knowledge as a network of propositions:

> *People remember not the exact wording of verbal communication, but the meaning underlying it. The smallest unit of knowledge that can stand as an assertion bearing meaning is the proposition. Memory, then, is represented as a network of such propositions. The strength of the association paths leading to that piece of information contributes to the level of activation being spread. This theory of spreading activation has influenced the design of many semantic network based information retrieval systems.* (Chen et al, 1996: 2).

Chen et al. show an implementation of Hopfield's (1982) neural network that is used as a memory of contents. The information in this network is stored in the form of nodes interconnected by weighted connections. Each term occupies the place of a node or neuron; the relations between these neurons are unidirectional. With each iteration, if two nodes are activated, the relation between them is strengthened while non-activated relations are weakened. This is called the *damping effect*, as in the spreading activation model of cognition (Collins & Loftus, 1975). After a number of iterations, there is a point of convergence where there are a few nodes strongly associated while the majority are in a periphery where the activation gradually fades away.

The methodology is to identify first each term in each document, then to count their frequency and to weight it using the inverse document frequency algorithm. There is, on the one hand, the tf_{ij} frequency, that represents the number of occurrences of the term j in document i, and on the other hand, df_j frequency that represents the number of documents inside a collection of n documents where the term j occurs. Instead of context windows around a term, matrices of the vocabulary of the whole document are used, at the cost of increased computational effort.

There is an extensive list of authors using co-occurrence data for specific tasks instead of hand coded lexical resources (Riloff & Shepherd, 1997; Roark & Charniak. 1998; Caraballo, 1999; Kageura et al., 2000; Widdows & Dorow, 2002; Widdows 2005). Widdows explains that to analyze the occurrences of *apple* in a corpus with the help of WordNet would be misleading because *apple* only appears in WordNet as a fruit and not as the name of a Company. In his work, Widdows advocates for a purely geometrical solution to the problem of word sense disambiguation. He uses co-occurrence graphs, where nodes are words and the connections between nodes are strengthened when the words that those nodes represent appear in the relation expressed by the operators “and” and “or”. Thus, when two words appear in a corpus connected by this operator, they occupy positions in the graph.

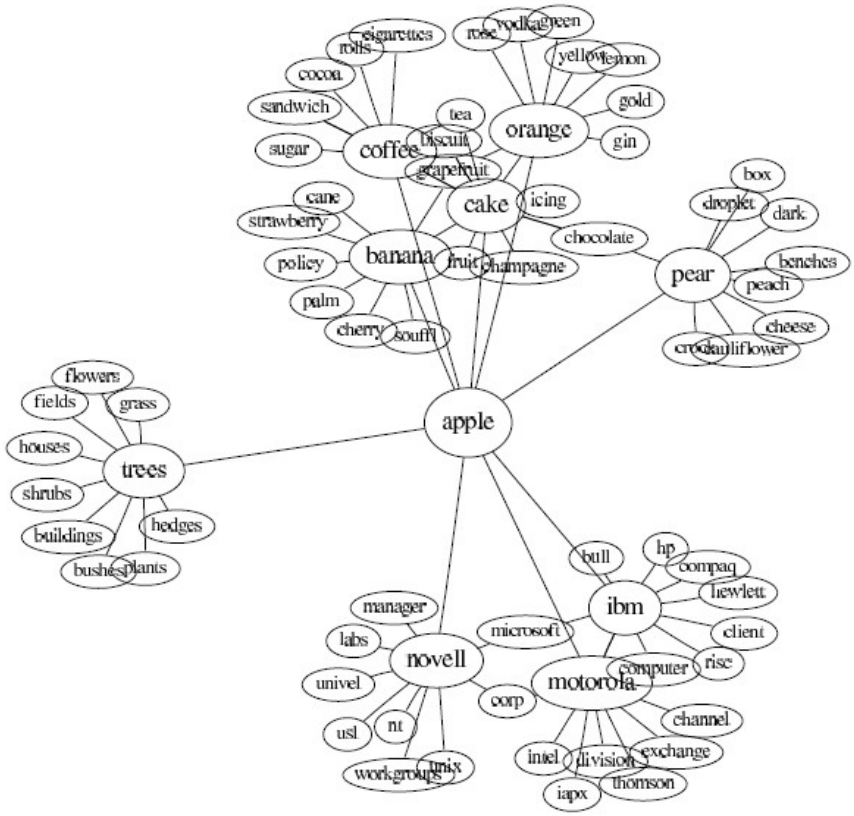

Figure 22: Word sense disambiguation using co-occurrence graphs (Widdows & Dorow, 2002).

The weight of the connection between nodes also depends on the frequency of the co-occurrence of such nodes. Figure 22 shows an example with the word *apple*, with different hubs of nodes, or clusters, that represent the different uses that the word *apple* has in the corpus.

Véronis (2004) presents an algorithm called *HyperLex*, which resembles the work of Widdows & Dorow (2002) and Widdows (2004). HyperLex also makes use of the specific properties of word co-occurrence graphs, which are capable of automatically determining word uses in a corpus without recourse to dictionaries. For example, depending on the search engine used, a query on the French word *barrage*, may return pages on dams, play-offs, barriers, roadblocks, police cordons, barricades, etc.

Véronis also shares with Widdows the belief that co-occurrence graphs can be better than dictionaries for specific tasks such as word sense disambiguation (WSD), an operation needed in most applications of natural language processing. According to Véronis, dictionaries lack an essential information for the characterization of a lexical unit, such as its distributional constraints with respect to the different uses that this lexical unit can have, and which can include number and types of complements, kinds of prepositions, selectional restrictions, etc. He was inspired by studies such as Schütze & Pedersen (1995) and Schütze (1998) on word sense disambiguation without using dictionaries. Véronis is critical, however, on the excessive computational cost of the vector space model presented in such papers and the fact that the model is only useful to discriminate word uses when they are few in number, clearly individualized and more or less equiprobable, such as *plant*, *train*, *vessel*, and so on. Therefore, Véronis was motivated to develop a less frequency-sensitive method, and in doing so he claims to have found in co-occurrence graphs a tool that supersedes the vector space models used by Schütze & Pedersen.

The graphs of Véronis share some features with the graph model presented in this thesis. Nodes in both graphs are words that co-occur with a target word that serves as query in a window of the size of a sentence or a paragraph. The connections between two nodes *A* and *B* strengthen whenever these words co-occur in the window. Figure 23 shows a fragment of a graph created from the target French word *barrage.* There, it can be seen that the nodes *production* and *électricité* are more connected to each other than other nodes that correspond to a different use of the word.

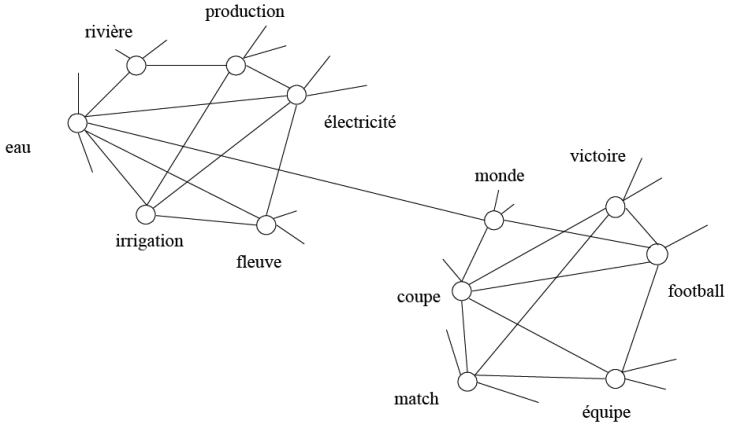

Figure 23: Fragment of a co-occurrence graph for the French word "Barrage"

As Widdows, Véronis is interested in the properties of what in graph theory is called *small worlds*. This is a graph whose connections do not follow a random pattern. Instead, there is a Zipfean like distribution of connections per node, meaning that few nodes, the hubs, concentrate the vast majority of the connections. This geometrical property means that while the number of nodes increases exponentially, the number of nodes needed to travel from one node to another only increases linearly. There are two measures for characterizing small worlds: the *characteristic path length* (*L*) and the *clustering coefficient* (*C*). *L* is the mean length of the shortest path between two nodes of the graph and *C* can be defined locally and globally: locally for a node *i* as the proportion of connections that *i* has with neighbors and globally as the mean of the local coefficients. This phenomenon was first noticed as the "Erdös number" (Widdows, 2004). Paul Erdös published around 1400 mathematical papers, mostly with co-authors. The Erdös number is, thus, the distance one has to Erdös as an author. A person that co-authored a paper with Erdös has an Erdös number of 1. A person that co-authored a paper with someone that co-authored a paper with Erdös has an Erdös number of 2, and so on. The astonishing fact is that there is no mathematician with an Erdös number higher than six. This was later known in popular culture as *the six degrees of Kevin Bacon*: if the population of Hollywood actors is represented in a graph, meaning that each actor occupies a node and the connections with other actors represent that they acted together in the same movie, then there is no actor in Hollywood that has a path length of more than six nodes in relation to Mr. Bacon (or to anybody else, for that matter). See Watts & Strogatz (1998) for an introduction on the subject of small world graphs and their function in self-organizing systems.

The methodology followed by Véronis consists of extracting a corpus from the web and analyzing the vocabulary defined as single nouns and adjectives that appear at least ten times. He found that verbs were not useful as nodes because verbs have more general uses and are, therefore, more independent of the content of the text. As consequence, they create false connections between different hubs. Véronis acknowledges that this treatment is unsatisfactory and that the problem needs more research. The problem is that not all verbs behave in this same way, and there are many verbs that are semantically focused or that are essentially related to the content of the text, as are some predicates in specialized terminology.

In Véronis' approach, function words (determiners, prepositions, etc.) are also eliminated, as well as other words found in a stoplist, including words related to the web itself, such as (*menu, home, link, http,* etc.) that have the same devastating effect as common verbs, connecting hubs which are unconnected from the point of view of content. The process of generating the graph involves a sequence of steps in which words are assigned to nodes and nodes are connected to each other. The connections

are weighted with a measure of distance that decreases (it would increase if a similarity metric had been used) as the association between words in the connected nodes increases. This is shown in Formula 5, where $p(A|B)$ is the conditional probability of observing A in a given context, knowing that the context contains B and, inversely, $p(B|A)$ is the probability of observing B in a given context knowing that it contains A.

$$w(A,B) = 1 - \max[\, p\,(A \mid B), p(B \mid A)\,] \qquad (5)$$

This coefficient expresses thus the semantic distance between two nodes. It is interesting to note how this coefficient solves the problem of the asymmetry of semantic relations.

As a result, the graph offers a spontaneous clustering of word senses. Again with the example of *barrage*, when it is used in the sense of hydraulic dam it is more probable that it will co-occur with nodes such as *eau* (water)*, ouvrage* (work)*, rivière* (river)*, crue* (flood)*, irrigation* (irrigation)*, production* (production)*, électricité* (electricity), etc., and these nodes will be highly interconnected, forming a hub.

5.1.2.3 Paradigmatic Statistics

Possibly the first antecedent of paradigmatic statistics is the work of Harris (1954), who exemplifies a paradigmatic relation, in this case between synonyms, with the pair *ophthalmologist* and *oculist*, two words that share the same profile of co-occurrence. Following this basic idea, Grefenstette (1994) undertakes the task of the automatic extraction of synonyms. In his book, Grefenstette advocates what he calls "weak" or "knowledge-poor" techniques. His claim is that they may be sufficient to advert structural regularities in text that offer enough clues about similarity between words.

The intuition underlying this approach is that when a human reader encounters a word whose meaning is unknown or unclear, what he or she unconsciously does is to compare the contexts in which words appear. The context allows the reader to judge if an expression is similar to a previously known one. With this intuition and the availability of massive amounts of electronic text as data, Grefenstette uses simple techniques and tools to manage and structure textual knowledge.

Grefenstette offers several motivations for the use of knowledge poor techniques. One of them is that manual approaches cannot keep pace with the ever growing amount of specialized knowledge. His approach does not need domain-specific information nor manually-assigned semantic tags. The semantic information is extracted in a totally automatic manner from

unrestricted large corpora. His proposal is a software called SEXTANT, which discovers similarities between words based on the distributional hypothesis that words that are used in similar contexts in a corpus are, supposedly, semantically similar. The methodology involves the parsing and extraction of syntactic contexts. For example, given a noun, the system will find which nouns and adjectives modify it, as well as verbs of which the noun is the subject or object. The words that occupy these relations with respect to a given word are considered the attributes of this word, in order to use a similarity measure between vectors to compare two words. Thus, those words that share a great number of attributes are considered similar. The shallow parsing part can be done with little knowledge about the language under consideration, Grefenstette claims that a technique that uses this syntactic analysis outperforms those that only use co-occurrence counts within a window of words, like Phillips' (1985).

The preprocessing of the text is almost identical to other approaches commented upon earlier. Grefenstette's method is slightly more language-specific, because it uses tokenization with treatment of English contractions. As others, he avoids the treatment of multiword terminology, so spaces and punctuation are word delimiters, except in the case of acronyms, where there is a letter-period-letter-period pattern. Proper nouns are defined as sequences of words beginning with uppercase letters not appearing after punctuation marks. Then, a morphological analyzer runs through the corpus under analysis. If a word is not found in the dictionary, then it is assigned a UKN (unknown) category and is treated as a noun. SEXTANT uses Brill's (1992) tagger, on whose output, he applies noun and verb phrase chunking.

$$\text{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{6}$$

The result of this process is a feature vector for each lexical unit. Grefenstette applies to these vectors a Jaccard similarity measure (Formula 6), where the numerator is the total count of attributes shared by object *X* and object *Y*, while the denominator counts the total number of unique attributes possessed by objects *X* and *Y*.

The attempt of Schütze & Pedersen (1997) shows some resemblance to Grefenstette, with the difference that Schütze and Pedersen do not use linguistic processing, but instead extract pairs of synonyms (or quasi-synonyms), applying a cosine similarity measure among word-vectors with their collocations as features. Thus, astronaut and cosmonaut appear to be very similar because they share many of their collocates, like Moon, space, spaceship, etc. Schütze and Pedersen's methodology is to extract

first the pattern of co-occurrence for each term, what they call the thesaurus-vectors, and then, the degree of semantic relation through the cosine similarity measure among vectors.

$$w(i,j)=\frac{f_{ij}}{\sqrt{f_i f_j}} \qquad \begin{array}{l} f_{ij} = \text{Number of contexts where } i \text{ and } j \text{ occur together.} \\ f_i = \text{Number of contexts where } i \text{ occurs} \\ f_j = \text{Number of contexts where } j \text{ occurs} \end{array} \qquad (7)$$

As in Phillips (1985), the purpose of the measure exposed in Formula 7 is to assign to each word a vector of local co-occurrences. This vector can be compared with other vectors to determine the degree of similarity. The computation is done with a symmetrical term by term matrix. The cell of the matrix corresponds to the pair of terms *i* and *j*, and the value is their co-occurrence in a context window of 40 words. The similarity between vectors is determined by the cosine coefficient, which yields as output a multidimensional scaling plot, necessary to reduce the complexity of an *n*-dimensional space to a two dimensional space that a human can perceive.

Vector based similarity coefficients between words have also been applied to Word Sense Discrimination by Schütze & Pedersen (1995) and Schütze (1998). This problem is only a part of Word Sense Disambiguation because it consists only in identifying clusters of similar contexts of occurrence of a polysemoys word. Word Disambiguation would be subsequet process of labeling each context of occurrence of such word with the correct sense. In addition, if senses are labeled as in lexicographical definitions, this would require some sort of external knowledge. The clustering of senses, in contrast, does not need any source of knowledge and is therefore purely automatic.

The idea of context-group discrimination is captured on Figure 24 -taken from Schütze (1998)- which depicts a vector representation of contexts of occurrence $\{c_1, \ldots c_8\}$ of the ambiguous word *suit*, which has different senses in legal and in clothing contexts. A first cluster, with its centroid labeled as sense 1, is closer to the *legal* vector because it shows higher cosine similarity, while the centroid of the second cluster, or sense 2, is closer to the *clothes* vector. The cosine similarity coefficient reflects the overlap between the neighbors of the words. Basically, the idea is that if the ambiguous word *suit* occurs in a context with other words related to the legal world, then it is likely that this is the context of the *legal* sense of the word.

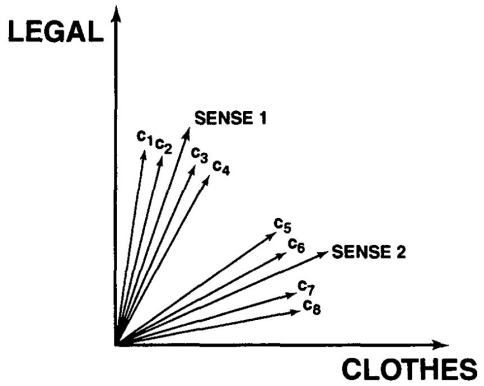

Figure 24: Vector-based clustering of contexts of the word "suit"

Vector comparison applied to paradigmatic linguistics is also the basic idea underlying *Latent Semantic Analysis* (LSA), which is also called *Latent Semantic Indexing* (LSI) when applied to Information Retrieval. LSA is a statistical technique specially designed for corpus linguistics. It is derived from Singular Value Decomposition (SVD) and Factor Analysis, two standard techniques in statistics used to manipulate matrices with high number of dimensions (Landauer & Dumais, 1997; Landauer et al., 1998). It can be seen as a method for multidimensional scaling. LSI extracts a co-occurrence matrix from a collection of documents. The matrix has terms as rows and documents (or contexts of occurrence) as columns. The cells of this initial matrix contain the frequency of occurrence of every term in each context. This means that terms or contexts can be compared as vectors using a similarity measure, that is, to relate words that do not necessarily co-occur but share the same profile of occurrence.

Although the model is presented at a pure theoretical level as a psychologically plausible theory of associational learning, LSA is also used as an IR technique. LSA can relate a query to a relevant document even when the document does not contain terms of this query, because it has a model of the type of context in which the terms of the query occur. Therefore, it can relate a particular document with this model of context and compute a similarity metric. The principle is transitivity: if a word or stimulus *X* is associated with stimulus *Y* (*X* and *Y* are frequently found in joint contiguity) and *Y* is associated with a word or stimulus *Z*, then *X* and *Z* are also related.

The basic idea underlying LSA is the same as that of the already mentioned studies such as Harris (1954), Grefenstette (1994) and Schütze & Pedersen (1997), among others. Landauer & Dumais (1997) are

however cautious when presenting their claims based on cognitive grounds:

> *We, of course, intend no claim that the mind or brain actually computes a singular value decomposition on a perfectly remembered event-by-context matrix of its lifetime experience using the mathematical machinery of complex sparse matrix manipulation algorithms. What we suppose is merely that the mind (or brain) stores and reprocesses its input in some manner that has approximately the same effect.* (Landauer & Dumais, 1997, p. 25).

Paradigmatic statistics has also been applied to the extraction of bilingual lexica. The strategy followed in the first attempts to extract bilingual dictionaries by statistical means, in the eighties and nineties, involved the exploitation of parallel corpora (Brown et al., 1990; Gale & Church, 1991; De Yzaguirre et al., 2000, among others). Using some measure of statistical association, such as chi-square of Mutual Information, one can obtain a reasonable accuracy (over 90%) depending on the size and quality of the corpus. The problem with this approach is, in the case of terminology, it is not always possible to obtain a parallel corpus of a given domain. Thus, the application of paradigmatic statistics is, when there is no parallel corpus available, is to obtain the bilingual dictionary from monolingual corpora. This trend is represented by Fung (1995; 1998); Fung & McKeown (1997); Rapp (1999) and Tanaka & Matsuo (1999), among others, whose strategy is to study and compare profiles of occurrence of both the term to translate and each of the equivalent candidates. Another strategy (Nagata et al., 2001; Guihong et al., 2007; Nazar, 2008; Nazar et al., 2008) is to concentrate on the contexts of occurrence of the source terms directly in the target language, because the current size of the Web has acquired such a critical mass that term equivalents in different languages are already statistically associated.

5.2 Conclusions from Related Work

This chapter has given an account of an important number of studies devoted to the general topic of automatic generation of conceptual representations from running text. The chapter shows that under this label a rather heterogeneous set of studies can be found. There is a major division between methods that are symbolic and those that are statistical as well as hybrid. Each approach has benefits and drawbacks. The first is more costly than the second because it needs a greater degree of knowledge of the language and is domain specific, while the second is useful even when this knowledge is not available. The first is capable of dealing with infrequent data, while the second needs a critical mass of text to yield results.

With respect to the statistical approach, mainly based on lexical co-occurrence, there is again a division between methods that exploit the relations between terms *in praesentia*, as well as those which study relations between terms *in absentia*. Each perspective allows us to observe different phenomena in language, in one case, syntagmatic relations and in the other paradigmatic relations.

At this point it is difficult to state if one approach is better than the other. It is difficult to compare the main approaches as well as the individual studies because of the incompatibility of their evaluation methods. A potential experiment that could be undertaken on the basis of the replication of all the described proposals would be an immense amount of work, yet undoubtedly fruitful and instructive. Only in an experiment like this there would be a common goal, common data, and then different algorithms would have measurable and compatible scores.

One of the most important commonalities in the reviewed literature is that the motivation is lexicographical, in contrast to this thesis' pursuit on a conceptual plane. Consider, for instance, the work of Kilgarriff et al. (2004), who attempt to summarize the typical grammatical relations of a lexical unit, including its collocational behavior and a distributional thesaurus. The present thesis, in contrast, is not headed towards the study of vocabulary but towards the study of referential expressions. Recall that referential expression is one of the most important terms in this thesis, and has been already defined and commented in the Introduction as well as in Chapter 2. An intuitive way to distinguish referential units from common words is that the latter are defined in dictionaries while referential units are included in encyclopedias. Words like *Paris* and *Socrates* are referential units, because they refer to specific entities, while words like *tables* and *chairs* as well as all the words that refer to general categories or that serve as predicates for other words are not referential units but common words, or functional units. However, and this has already been stated repeatedly in this thesis, the distinction between the two is sometimes not as clear as in the above examples, since it is entangled with historical events. Proper nouns become common nouns and this is also true the other way around. Furthermore, when a term designates a concept instead of a singular entity, as it is the case with most specialized terminology, it can still be said that such term is a referential expression. It is hard to tell exactly which units belong to the core of a language and which ones are part of the extra-linguistic reality.

This thesis is in part motivated by the fact that there is relatively small number of studies that operate with a linguistic perspective. Almost all of the studies reviewed at the moment exhibit a clear engineering or practical perspective, while the motivation of this thesis is to draw attention to the

importance that repetition has in language and conceptual learning. Of all the studies within this chapter, only those that are more psycho-linguistically oriented show a greater degree of resemblance to this thesis, in particular those which describe the phenomenon of semantic priming. It is this phenomenon that demonstrates the psychological plausibility of the approach offered in this thesis. More recently, this orientation has motivated some authors (Lenci, 2008; Sahlgren, 2008) to vindicate the place of distributional studies in linguistic theory and, more specifically, semantics, and at the same time to distance themselves from the "uncritical approach" of Information Retrieval engineers.

Among the topics that are less covered by the literature is asymmetrical priming. This has been object of study in psycholinguists but it is a phenomenon that has received little attention in the literature on quantitative linguistics, and there is still much to be said about this subject. Most authors that use measures of vector comparison ignore this problem and this may be misleading even from a practical point of view, for example when dealing with collocational data. As already mentioned in the introduction, terms can have asymmetrical relations like hypernymy, and this is reflected in collocational data. The terms *asthma* and *illness* are semantically related, but it is much more probable to encounter the hypernym when one has encountered the hyponym than the other way around.

In summary, a number of approaches have been commented, each of them with advantages as well as shortcomings. The rule-based algorithms to extract terms and conceptual relations that have been commented in this chapter show specific shortcomings, for instance in the treatment of polysemous terms. There is, thus, an evident need of more research and development of robust tools to treat polysemous terms to overcome problems derived from ambiguity. Another shortcoming of rule-based systems is that they are language and domain dependent, which means that to adapt these strategies to a particular language and domain may be potentially costly. Finally, the major shortcoming in all the literature commented upon, and not only of rule-based systems, is perhaps the lack of linguistic conclusions, because most of the authors reviewed concentrated on the solution of practical problems. What motivates this thesis is, thus, the need to fill this gap and shed light directly on the intrinsic relation that exists between conceptual structuring and lexical co-occurrence. Experimental solutions to practical problems will also be addressed, however the main motivation of the experiments of the thesis is to provide empirical evidence in support to this thesis' claims.

Chapter 6: This Thesis' Approach

The general idea of the approach is to analyze concepts by studying the co-occurrence of the term that is used to designate that concept. In other words, the idea is to see what is the lexical neighborhood(s) of the term(s) used to refer to a given entity, being this material or conceptual. It is possible to formulate this statement in a more general or abstract way by saying that it is possible to capture the recurrence of certain events in the same situations. In this study, an event is defined as the occurrence of a term in a context, and a situation as the context itself, a fragment of text of *n* words left and right from where this term occurs. In this way, it is possible to estimate a relationship between events by their significant co-occurrence. When the event is the occurrence of a given term, it is then possible to "represent" the concept referred by the such term by the selection of significant co-occurring terminology.

The methodology described in this chapter has been designed and implemented in order to serve as a proof of empirical sustainability and replicability. Thus, this chapter describes an algorithm which is applied to a series of experiments that are intended to serve as empirical evidence for the hypotheses presented in Chapter 2. The description provided by the present chapter targets the core algorithm, which is common to the methods used in each of the experiments that are shown in the next chapter. Each of those experiments has, in addition, specific variations on the methodology which will be explained in the corresponding sections of that chapter.

The methodology is expressed as a function that, given a query term and a collection of documents, constructs a network containing the query term itself and other terms that represent what is essential to the entry term's meaning(s). Assuming an operative definition of *term* as a word or a combination of up to *n* consecutive words with some restrictions that will be explained later (section 6.2), an overall description of the methodology would be:

1. Contexts of occurrence of the query term are extracted and the vocabulary items within these contexts are filtered on the basis of their statistically measured quantity of information in order to treat only the most significant units from the corpus.
2. The algorithm accepts as input the vocabulary that was selected as most informative from each context and each vocabulary item is assigned to a unique node in a network.
3. The arcs between the nodes of this network are strengthened each time the terms associated with the nodes appear together in the same context, while the rest of the connections are weakened.

As this process goes on, few nodes within the network are significantly associated while significant arcs are not found between the vast majority of the nodes, following a Zipfean distribution. At the end of the process, the group(s) of the most interconnected nodes shows an important overlap with the key terms that contribute to the meaning(s) of the query term.

6.1 The Graph Model

At the beginning of the chapter, it was stated that an event is defined in the scope of this thesis as the occurrence of a term in a fragment of text or sentence. Example 1 shows the occurrence of the term *Multiple Sclerosis* in a sentence.

1) *Multiple Sclerosis (MS) is an inflammatory, chronic, degenerative disorder that affects nerves in the brain and spinal cord.*

There is a set of concurrent events in this same situation: the occurrence of the terms *nerves*, *brain* and *spinal cord*, and others. The co-occurrence of these terms is not accidental. It is to be expected that these terms will appear together. This is because example (1) is an analytical statement, thus the predicate is included in the subject by definition. Example (2), instead, is a factual (or synthetic) statement which is based on empirical evidence (the results of experiments using magnetic resonance imaging).

2) *Regular magnetic resonance imaging evaluations show that only about half of patients with multiple sclerosis achieve and sustain a response to treatment.*

The facts related to the term *multiple sclerosis* will determine if the predicate of example (2) will be attached to the term *multiple sclerosis* as an essential attribute, that is, an attribute that is presupposed by the term. Again, this does not necessarily imply that this will make it a definitional feature of the concept *multiple sclerosis*. The vast majority of the contexts of occurrence of the term *multiple sclerosis*, however, will not have the function of predicating essential attributes of the term. Rather, they will be applied to a diversity of particular instances, as it is the case with example (3).

3) *This 47-year-old patient was diagnosed with multiple sclerosis in November 2000.*

Example 3 is a statement where the term *multiple sclerosis* occurs without being the subject of an analytical statement. It is a factual statement about the condition of an individual, it is not intended to predicate about the disease. Sentences such as (3) are a class of sentences that is characterized by its extreme diversity as a result of the multiplicity of individual cases. Although some of the terms in (3) may be repeated (i.e. *patients*, *diagnosed*), the probability of lexical co-occurrence of contingent (non

conceptual) information is sufficiently low to be discarded from the analysis.

Terms that are conceptually related form patterns of co-occurrence and these patterns can be studied using graphs. A graph model is inferred from text and represents the relations between terms as paths between nodes. Nodes represent the related terminology. The selection of the terms from the text is the result of *traveling many times over a small number of paths.* Every time a path is activated, the weight of its arc is strengthened while the rest of the arcs is weakened. The result, after a number of cycles, is a configuration of terms that are semantically associated with the query term. It is a natural selection because the terms converge in clusters that represent the canonical features of the concept designated by the term.

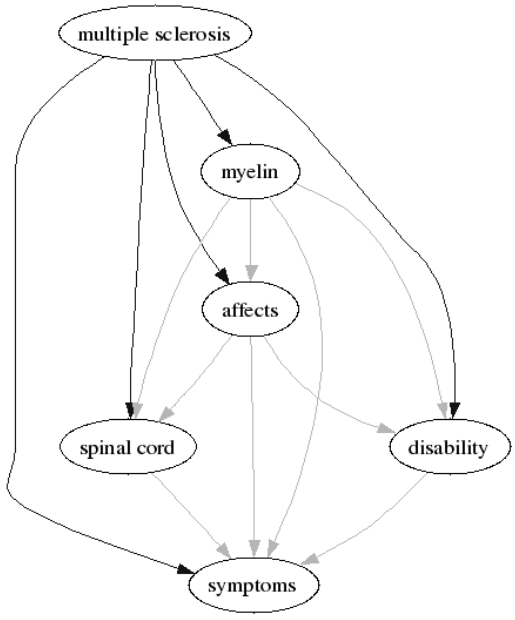

Figure 25: A small co-occurrence graph of *Multiple Sclerosis* (already shown in chapter 1).

What follows now is a more detailed explanation of how the nodes and arcs are drawn, yet it is already possible to show, in Figure 25 (already shown as Figure 1), an example of a co-occurrence graph of the term Multiple Sclerosis. This chapter deals only with the graphs for the study of syntagmatic relations, that is, of first order co-occurrence. Paradigmatic relations, or second order co-occurrence will be used in Chapter 7, where different experiments are carried out comparing the graphs of different terms.

1. **The Nodes:** As it was already stated at the beginning of this chapter, only the most informative terms from the corpus will be assigned to nodes. This selection of the vocabulary is performed using Mutual Information (section 5.2.1.2). Nodes are, thus, defined as strings of *n* consecutive orthographic words which have a value of Mutual Information in relation to the query term above a threshold. The application of the Mutual Information coefficient requires a model of the language of the analyzed documents. This language model informs the expectation of a certain word in a message. A minimum useful model can be a plain frequency list of words from a corpus of general language. The justification for this procedure is that the only thing needed from such a model is the set of high and mid-frequency words, because this range covers the core vocabulary of a language, which is the least informative.
2. **The Arcs:** The arcs between nodes in the graph can be categorized as primary and secondary arcs. This distinction is needed because the graph is built starting from a query term. Therefore, the category of primary arcs refers to the arc between the query term and the rest of the nodes, while a secondary arc is the one that connects two nodes when neither of them is the query term. Primary and secondary arcs between two nodes are strengthened every time these nodes are activated, that is, when they co-occur in a context window. Given a query term *t,* the weight of the arc between a node *i* and a node *j* that appear in the context of *t* is defined in Formula 8.

$$R(i,j,t)=\log\left(\frac{F_{ij}(t)}{N}\right) \quad (8)$$

t = query term
N = the sum of the contexts where *t* occurs.
$F_{ij}(t)$= the number of contexts of occurrence of term *t* where nodes *i* and node *j* occur.
i, *j* = nodes in the context of *t.*

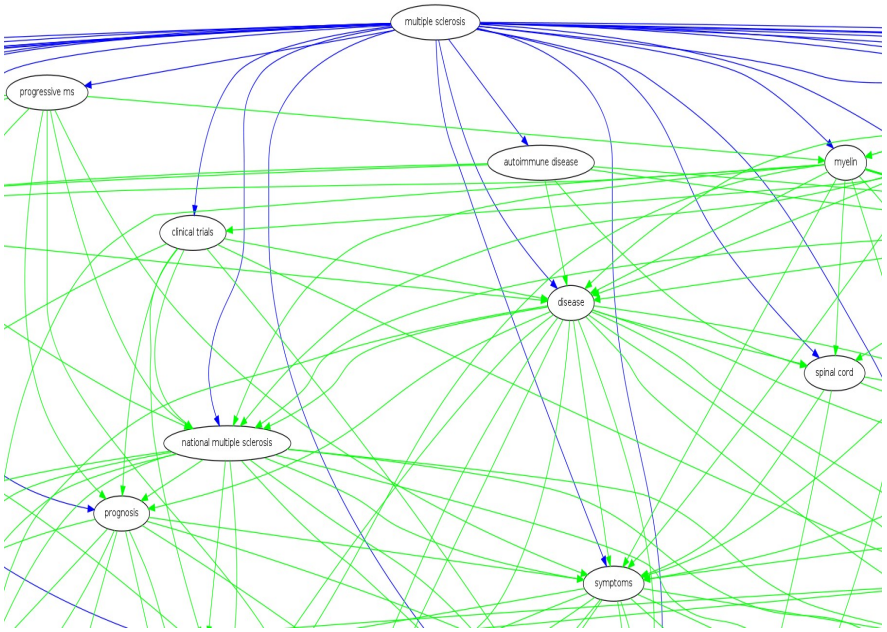

Figure 26: A more complex graph for "multiple sclerosis"

Figure 26 shows another graph for the initial query term *multiple sclerosis,* represented earlier in Figure 25, this time with a lower threshold for the selection of the vocabulary and, thus, with a more complex graph. This will depend on the needs of every specific study. For instance, in the type of experiments shown in Chapter 7 on the extraction of taxonomies, the graphs that are generated are complex to include more information. In a more complex graph like the one in Figure 26, it is possible to observe that the central nodes are the hypernyms *autonomus disease* and *disease*, an information that was not present in the initial phases of the graph in Figure 25. A critical aspect of this process is, thus, to determine when to stop growing the graph, otherwise it can continue indefinitely growing in complexity. This can be done in several ways. One of them is to take advantage of a ubiquitous property of language which is the Zipfean distribution. After a number of cycles, one can see a pattern of interconnection in the graphs where there are a few hubs with nodes that are strongly connected while the majority is dispersed and weakly inter-connected. There are at least three options: the first and simplest option is to eliminate the nodes whose arcs have a weight under a certain threshold. A second possibility is to keep only the first *n* most connected nodes. The third is to automatically determine the threshold from looking for a discontinuity in the curve representing the graphed values of the arcs: ordering the weights of the arcs by decreasing value, then it is possible to compare each value with the previous one and select the most different pair. If that difference is greater than a given threshold, then a pair of nodes will represent a discontinuity. The solution that offered better

results and was finally implemented was to automatically increase the constraints until a desired number of nodes was reached, that is, the second option.

6.2 Description of Graph Model's Parameters and Functions

While the previous section offered the basic idea of the kind of process that is taking place, the rest of this chapter describes the graph model and how texts are analyzed, with the details that are important for replication of experiments and evaluation of empirical evidence. The implementation presented here is not intended to be the only possible application of the theoretical issues discussed in this thesis. Different possibilities of application will be discussed in Chapter 9. This implementation serves only as the basis for an empirical validation of the hypotheses.

6.2.1 Input Parameters

The description of the algorithm begins with the definition of the parameters it uses. Parameters are defined as sets of attribute – value pairs. The list of parameters that will be explained in this section is the following: 1. Input to the Process, 2. Language, 3. Other Input Terms, 4. Type of Document to be Analyzed, 5. Threshold of Mutual Information, 6. Maximum Number of Nodes in the Graph, 7. Size of the Context Window, 8. Minimum Frequency, 9. Number Filtering and 10. Filtering of Outliers.

1. **Input to the Process:** The input consists of a query which is the term that will be subject to analysis. It can be a single word or a multiword expression. The only condition is that the expression must be expected to appear in documents in the exact form that it was typed.
2. **Language:** The language of the documents to be retrieved are limited to available corpora, although there is no inherent restriction with respect to the language that can be selected. Experiments with this algorithm have already been undertaken with languages such as Spanish, English, Catalan, French, German, Dutch, Italian, Turkish and Danish. To be able to process documents in these languages, a minimum knowledge of them is needed. There are two kinds of language models in the experiments undertaken in this thesis: a simplified model and a complex one. Details on the construction of these language models are offered in section 6.3.
3. **Other Input Terms:** Other terms can be added as a separate group attached to the query, simultaneously or afterwards. These

terms are added to the query that will be submitted to the search engine, with the only purpose to narrow down searches in case the user knows that the expression first introduced is ambiguous and wishes to disambiguate it.

4. **Type of Document to be Analyzed:** The choice of the types of document formats to be retrieved included .pdf files; Word .doc documents; .ps postscript files or general .html formats. It becomes quickly evident that specialized scientific literature often uses .pdf or .ps format. The selection of the type of document may have a dramatic impact on the quality of the downloaded collection, but creating many difficulties for its conversion to txt.
5. **Threshold of Mutual Information:** The purpose of the application of the measure of Mutual Information is to determine how significant the co-occurrence of a pair of terms is, and to reject collocations that may appear just by chance, that is, because their collocates have high independent frequencies of occurrence. The threshold for this measures can be defined by hand, although it is automatically adjusted by default to meet the best conditions. These conditions are set on the basis of the number of nodes that populate a graph. If the number of nodes is above a certain parameter, then the minimum association score will increase until the threshold of number of nodes is reached, because fixing a high threshold will result in fewer terms selected as significant in relation to the occurrences of the query expression.
6. **Maximum Number of Nodes in the Graph:** The number of nodes, at this stage, can be set arbitrarily, and it is possible to adjust it in subsequent stages of the execution, according to the results obtained. By default, this number is set to 30 nodes, so the program eliminates nodes with the weakest connections until the desired number of nodes is reached.
7. **Size of the Context Window:** There is no sentence boundary detection in the processing of text. The selection of the contexts of the instances of the query expression is done on the basis of a number of orthographic words at each side of the query. This number may be set to 5, 10, 15, 20 or 30 words, 15 being the default value. The theoretical basis for such a decision, beside practical considerations, is the assumption that semantic association transcends sentence boundaries (as suggested by Phillips, 1985).
8. **Minimum Frequency:** As already discussed, it is a widely known fact that vocabulary size grows exponentially as one takes into consideration words of fewer occurrences. For practical reasons, there must be a limit in the size of the vocabulary to be treated because uncontrolled size would be detrimental to run time. However the reasons are not only practical, since the model presented in the thesis is intrinsically related to the frequency of

co-occurrence. For this reason, taking into consideration the elements of low frequency would be self-contradictory. There is a lower threshold that is adjusted by default in relation to the size of the vocabulary set, with 4 as a minimum.

9. **Number Filtering:** This parameter, which carries a binary value, specifies if numbers within the text are to be included in the analysis. It is often difficult to treat numbers because they do not exist in the language model. As a consequence, it is not possible to assign an expected probability of occurrence to them. By default this switch is on, that means, it ignores numbers. Nevertheless, it would be important to turn it off in cases where the user thinks numbers are indeed important, for example if years, dates or dimensions of objects are considered to be important. Numbers are also present in specialized terminology, for instance in terms such as *trisomy 18* or *alpha-1-antitrypsin*.
10. **Filtering of Outliers:** Also a binary parameter, with value 1 it is set to ignore terms that appear in only one document of the sample, regardless of the frequency of occurrence. This is consistent with the theoretical measure of relevancy stated so far, which associates the importance of a term with the degree in which it is distributed among a corpus of specialized literature. Terms that come from only one source are not to be considered important[21]. By default, the status is on.

6.2.2 Step by Step Simulation of the Construction of the Graph

Now that the parameters have been defined, this subsection offers a description of the steps that are undertaken during the construction of the graph. Let us assume a definition of the parameters such as those in Table 14. Then the process would have the nine steps explained in this section.

1. Input	duckbill platypus
2. Language	English
3. Other Terms	100 documents
4. Type of Document	any
5. Threshold of MI	9 bits
6. Number of Nodes	30
7. Size of Context	15 words at each side
8. Minimum Frequency	4 occurrences
9. Number Filtering	off
10. Filtering of Outliers	on

Table 14: A sample of a parameter setting.

[21]This is also relative to the total number of documents in the corpus. In a corpus of thirty or fewer documents, the occurrence of a term in only one of them is more important than the occurrence in one document when the total amount of documents is a hundred or more.

1. **Selection of the Corpus:** With the defined parameters, the process begins by conforming a corpus and that is achieved by querying an Internet search engine. It may be the case that the search engine does not find any document with the query. In such a case, the process is aborted. The same occurs if the number of documents retrieved is less than a minimum threshold (three documents).
2. **Retrieval of Documents:** Once the URLs have been stored in the previous step, the download process begins. Sometimes problems occur, for example if the page is not found or if the server does not allow robots. In this case, the error is registered and the process continues. Once the documents have been downloaded, they must be converted from their original formats to text file format. In the case of web pages, html code, as well as Javascript code and Cascade Style Sheets are eliminated. In the case of the rest of the formats, the conversion to text is done with the aid of external programs that are invoked automatically. Pdf (Portable Document Format) files are transformed into ASCII text using the "pdftotext" (Noonburg, 2004); Ps (PostScript) files, using "pstotext" (Birrell et al., 1995) and .doc (Microsoft Word) documents are transformed using "antiword" (van Os, 2003). These programs entail some error-ratio and a corresponding data loss. All text is transformed into UTF-8 character encoding.
3. **Extraction of Concordances:** This step consists of extracting of the contexts of occurrence of the original query expression. As stated earlier, these context windows are defined by a number of orthographic words at each side of the query term. In order to perform this task properly, a minimum text-handling of the texts must be undertaken. This is mainly the separation of words from punctuation marks, like the following: . , : ; ¿ ? ¡ ! " ' : and other symbols like ( ) [ ] {} * + - _ / \ and newline characters. Concordances, once extracted, are allocated in a data structure. A fragment of such structure is shown in Table 15.

document reference	left context	query	right context
doc3	rubbish bin because it had a worm in it .	**duckbill platypus**	likes eating apples in warm water ponds in the middle of the night because
doc3	noon because he was part bug , part man .	**duckbill platypus**	was hiding in his burrow EATING AN APPLE today at 3 o clock because
doc3	eating an apple at midnight because it tastes yummy .	**duckbill platypus**	is our travel buddy who has gone to Canada. He is sadly missed
doc4	Platypuses have a bill which looks like a duck sand so the name	**duckbill platypus**	. But this is no ordinary bill It has an electronic radar system that
doc5	649656 2000 . PMID Ornithorhynchus anatinus Defensinlike peptide 1 , Ornithorhynchus anatinus	**duckbill platypus**	Ornithorhynchus anatinus

Table 15: Fragment of a context matrix.

4. **N-grams Extraction:** the strategy to select terms that form multiword expressions is to detect patterns of recurrent word combinations in the text, or word *n*-grams. The list of *n*-grams is filtered, eliminating those that begin or end with a member of the stoplist, or with numbers, in case the *number filtering* parameter is active. Of course, members of a stoplist can occur inside a trigram or larger *n*-gram without meaning that the unit is not terminologically valuable, like it is the case of the preposition *of* in English occurring inside multiword terms such as *analysis of variance*. The remaining terms form what will be called the *vocabulary*. This vocabulary, however, is subject to further deletions in subsequent filtering steps.
5. **Detection of Orthographic Similarity:** Since there is no lemmatization nor part of speech (POS) tagging in this procedure, a process of "pseudo-lemmatization" is undertaken in order to recognize and merge different forms of the same word. This facilitates word frequency counts and avoids the presentation of similar and redundant nodes in the results. No POS tagging has been used to avoid introducing language specific knowledge into the model. Instead of lemmatization and tagging, the system performs a pairwise comparison of all units of the vocabulary to find orthographic similarities. Two word forms are assumed to be forms of the same lemma if their orthographic similarity is above a threshold. As a consequence, their frequency counts are merged under the form of the most frequent word. It is certainly a matter of debate if this procedure is actually language independent, since some languages present inflectional systems that would surely

pose a problem for this simple procedure. In addition, an even less language-independent decision was made in order to increase accuracy as well as to speed up the procedure: only the words with the first three characters in common are compared. This will avoid confusions between forms such as *military* and *paramilitary*, but of course will only be possible in languages with inflectional derivation taking place at the final part of the word. Empirical evaluation shows that this ersatz renders satisfactory results for purposes of this thesis. Although *Australian* and *Australia* are surely not the same word, their semantic relatedness is considered enough to be considered as the same node in the context of these experiments. *Australian* and *Australia* are related in the context of the query search, which, in the case of this example, is the term *platypus*, and these two word forms have been signaled as collocates. Again, they are not clustered together just because they have an orthographic similarity, this happens only after those units have been identified as semantically associated with the query term. A very different case would have been to consider *intention* and *intension* as the same word just because they look similar. Nevertheless, the probability of having an error such as this is negligible because *intention* and *intension* are unlikely to be found in the same contexts, being not semantically and therefore not statistically associated.

Term A	Term B	Dice Cf.
Australian	Australia	0.933
mammal	mammals	0.889
animals	animal	0.889

Table 16: Different word forms merged under the same root.

In order to compare every word form against each other, all items of the vocabulary are decomposed into vectors with trigrams of characters as features. Table 16 shows some examples of orthographic similarity values. The similarity is calculated in the following way: when comparing the word *animals* with the word *animal*, for example, both words are decomposed into trigrams of characters such as *ani nim ima mal als*. It is easy to verify that both words share most features. The similarity coefficient used is Dice (Formula 9), and an arbitrary threshold of 0.75 was selected empirically for the detection of inflectional derivation.

$$\mathrm{sim}(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{9}$$

6. **Deletion of Outliers:** All lexical units that only occur in a single document are erased in this stage.
7. **Term Weighting:** For the reasons discussed in Chapter 2, raw frequency data of lexical co-occurrence is not enough to calculate how informative a term is. The co-occurrence of two elements that have also a high individual frequency is likely to be due to chance. As an example, the term *duckbill platypus* occurs frequently with *figure* as with *mammal* in documents retrieved from the web, thus it is necessary to ascribe a significance score to each co-occurrence. As already mentioned, a model of the language is needed to apply this measure. The model of English used here is a plain list of frequencies of words and common bigrams and trigrams. The fact that the word *figure* has a higher expectation is derived from the high frequency it shows in this model of English. In a similar way, the word *mammal*, being less frequent in that model, is less expected, and therefore, more informative. The column MI in Tables 17 and 18 indicates the mutual information of each unit (each row) with the query term *platypus*, on the basis of the downloaded collection combined with the reference language model. The forms *mammals* and *figure*, two examples of word above and below the MI threshold, are marked with arrows.

Unit	MI
mammals	12.8825
animals	10.9738
fur	10.9231
species	10.6076
sydney	10.5446
habitat	10.5446
evil	10.5446
swimming	9.9231
item	9.6601
links	9.5446
unique	9.2816
text	9.2094
length	9.1747
brain	9.1021

Table 17: Units above the threshold

Unit	MI
page	8.9887
facts	8.6905
eating	8.6905
document	8.6402
eat	8.4377
title	8.3871
lay	8.2451
figure	8.2331
obviously	8.1747
unusual	8.1576
types	8.1576
angry	7.9413
pictures	7.9413
features	7.9231
...	

Table 18: Units below the threshold

Candidates with an MI score below this threshold (9 bits by default) will be deleted. Of course, some of the terms in the

extracted vocabulary do not exist in the model, therefore it is not possible to calculate a Mutual Information score. As a result, in the case of multiword expressions, MI is the mean of the MI of each individual unit, because they are not expected to occur on the reference corpus that is used as a language model. At the moment, it is enough to say that the MI of a bigram like *lay eggs* is the mean of the MI of the unit *lay* and the unit *eggs*. One important exception is done in case the multiword unit includes, as a component, a unit of high frequency such as *of*, because it would distort the estimation. In the context of the experiments presented in this thesis, a unit is considered to have a high frequency if it has a frequency greater than 2000 in a reference corpus of an extension of two million words.

8. **Extraction of Extensional Nodes:** Frequently, and particularly in technical literature, when two terms have a hyperonymy relation, it happens that a hyponym is formed as a compound form of the hypernym. The hypernym is presented as a longer sequence with new elements modifying the head. These modifiers express a specification of the basic type and represent a subcategory or instance. This is the case, for instance, with the form *ulcerative keratitis* as a type of *keratitis.* Nodes that have this specific type of relation with respect to the query term are called *extensional nodes*. Any node in the graph is considered an extensional nodes if it includes the query term as one of its components. The problem consists in the refinement of such list of candidates, because there are also compositional fragments of text (simple *n*-grams) among them that include the query term as a component but cannot be considered extensional nodes, such as *patients with keratitis*, which is simply a compositional fragment. The strategy for the refinement is to eliminate those candidates that include components that can be considered uninformative from a statistical point of view. This is a form of statistical noise reduction. It is performed by finding patterns that are typical in certain domains, as elements that keep appearing in the context of any query term, which is the typical behavior of an uninformative unit. Let us consider the example of *keratitis* again: the trigram *patients with keratitis* is common considering that it is typical to find the pattern *patients with* *. This means that with independence of the query term that is used, if it is a disease, there will be *patients with* that disease. The strategy, thus, to extract extensional terms will be to extract all the sequences that include the query term as a component except those that have uninformative candidates at the beginning or at the end of the term.

9. **Generation of the Network:** The next step in this process is to build a network that represents the co-occurrences of the vocabulary terms. A co-occurrence graph is a data structure drawn directly from the concordances. Consider, for instance, the following context:

*...amongst the world's most unique and unusual animals and for good reason too . The **duckbill platypus**' are mammals because the females produce milk and nurse their young however they are...*

The central node is the query term, in this case *duckbill platypus.* From this node, an arc to each of the occurrence of other terms in the vocabulary is drawn. Having registered that *unique*, *animals*, *mammal*, etc. are significant and are stored in the vocabulary, then they are connected with the central node and also with each other. The strength of the arc is the ratio of the co-occurrences and the amount of contexts found. This means that, after context 1, the weight of the arc *duckbill platypus* and *mammal* (as well as the others) is 1/1. If in the next context the pair appears again, like in this other context:

*Yet, evolutionists do not propose that **mammals** evolved from birds with the **duckbill platypus** representing a transitional form between these two groups. With respect to RTB's model...*

...then the weight of the arc *mammal – duckbill platypus* has a new value of 2/2, while the others, such as the arc between *unique* and *animals*, have lowered to 1/2, as well as the arc between *animal* and *mammal,* because mammal has occurred twice but only one time with *animal*, and so on. This process continues until all contexts have been processed. At the end, nodes that are not well interconnected with the rest are eliminated from the network.

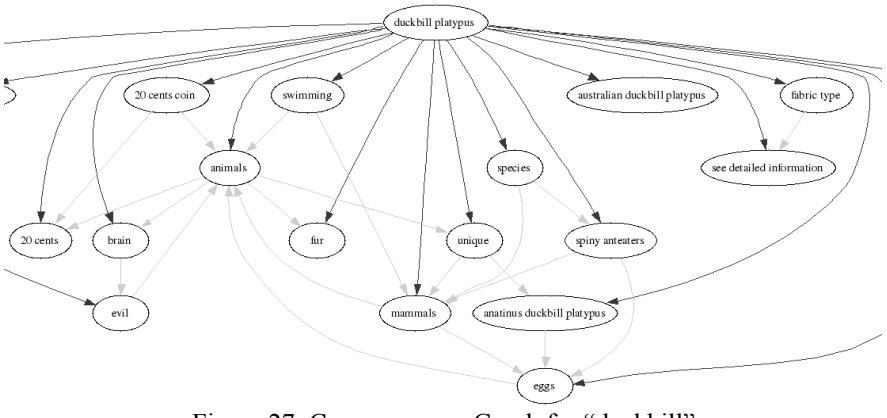

Figure 27: Co-occurrence Graph for "duckbill"

A fragment of the resulting graph is shown in Figure 27. The node *Ornithorhynchus anatinus*, the Latin name of *duckbill platypus*, is present in the graph but omitted from the figure due to lack of space. Notice that hypernym nodes in Figure 27, such as *animals* and *mammals,* are central nodes, with a large number of connection with other nodes. Some other nodes, such as *swimming*, *fur*, *unique* and *eggs*, are related to some of the distinguishing features of this species. It would be possible to write a satisfactory definition of the term using only some of the terms that this networks offers:

> Duckbill platypus: (ornithorhynchus anatinus) furred swimming animal of Australia, one of the unique species of mammals that lay eggs, along with the spiny anteaters.

This is the definition provided by WordNet:

> Duckbill platypus: small densely furred aquatic monotreme of Australia and Tasmania having a broad bill and tail and webbed feet; only species in the family Ornithorhynchidae.

There are, nevertheless, some strange objects such as *20 cents coin.* The presence of this element is explained by the fact that there is a 20 cents coin in Australia that has the Queen on one side and the figure of a duckbill in the other. This element disappears when the "ignore-numbers" parameter is active.

6.3 Details on *N*-gram Language Modeling

The purpose of this section is to explain the methodology for language modeling that was briefly introduced earlier. Language modeling is based on corpus statistics over large samples of documents. The first two questions that arise at this point are: 1) How to acquire such corpus and 2) How big must it be.

With respect to the first question, there is already a variety of corpora available on the web. The Wikipedia[22], the Wortschatz project[23] and the CT-IULA[24] are mere examples of large collections of text from different languages, genres, registers and topics. It is always possible to compile new corpora from off-the-shelf search engines. In this case the procedure is to automatically query these search engines with random combinations of high and mid-frequency words. Non-content words have the property of appearing in documents with independence of the topic of the documents, and this is what lets us gather a collection of general language that should not be biased in terms of vocabulary.

[22]http://www.wikipedia.org [accessed June 2010].
[23]http://corpora.informatik.uni-leipzig.de/ [accessed June 2010].
[24]http://bwananet.iula.upf.edu [accessed June 2010].

The other important issue at stake here is the size of the sample. This is not the place to discuss the problem of sample representativeness and the variety of techniques that are available to determine which is the minimum size of a sample to be considered representative. However, there is a possibility that there is no threshold. The philosophy in this thesis has been that bigger corpora simply yield better models. Therefore, in the cases where a more precise language model is needed, for instance in experiments on terminology extraction or terminology translation (see Chapter 7), a bigger model of the language can be used. An empirically determined minimum size for the language model a corpus of two million words.

The samples of text for the model should not be genre-biased, i.e., they must represent general language instead of technical or specialized literature. The diachronic factor and the evolution of language is also a source of problems for the model of the language. When dealing with corpora extracted from the web, such problems are, for instance, the cascading of words that occur very often in web pages but not very often in the corpus that was used to build the model (e.g., *page*, *search*, *edit*, *contact us*).

Leaving aside the problem of the acquisition of the corpus, there is the problem of how to proceed to develop the model. There are different possibilities for language modeling. The simplest one probably being to collect a corpus of general language and order the units of such corpus by decreasing frequency. Frequency would then be a rough measure of the probability of occurrence of such a term in a text, which is a way to determine the amount of information that the term has. This information forms a model of each language that is analyzed. This is only to assign a probability of occurrence to each word, and in fact the calculation is very approximate because the only thing that is needed to know is which are the most common words.

One of the uses of such language models is, for instance, the extraction of a stoplist from the language model. In a model build from a well-balanced general language corpus of two million words, a useful stoplist can be defined as the 100 highest frequency words. In addition, words with an absolute frequency of occurrence of less than six in the same corpus do not need to be included in the model. If one of such rare units appears frequently in a collection of documents under analysis, it will be considered significant. Given these characteristics, this version of the language models is simple and therefore easy replicate in other languages.

The more complex version of the language model, which is also more accurate, includes other factors besides just the frequency of occurrence of single units. One such factor is the distributional properties of the lexicon

on large corpora. There is a key measure, apart from frequency, that must be used in every model of language, and that is dispersion, the way in which the frequency is distributed across the corpus.

The process of language modeling involves the frequency listing of *n*-grams (n <= 5) of the whole corpus. *N*-grams sorting techniques were already explained in Section 6.2.2. An estimation of the heterogeneity of the distribution of a term in a corpus, and therefore, its weight or information, can be given by the simple ratio of term frequency over the frequency of such term in certain partitions of the corpus. These partitions can be documents or subsets of documents from the corpus of more or less the same extension. If a corpus is divided in domains and each partition is approximately of the same size, then a measure of dispersion is a way to determine when a pattern is independent of the topic of discourse. Predicates, in general, are expected to have a greater dispersion in comparison to referential expressions.

The coefficient proposed here for the modeling of units according to their distribution can be applied to a corpus with any arbitrary partition. In Formula 10, the dispersion of a unit *t* is equal to the quantity of subparts of a corpus with frequency 0 -or a minimum frequency indicated as a parameter- multiplied by the highest relative frequency of *t* in any subpart of the corpus, or year, in this case. The weight of *D* is not delimited between 0 and 1 but between 0 and infinity. The purpose of this measure is to highlight the units that have a prominent presence in one or few partitions of the corpus.

$$D(t) = \mathrm{Mf}(t) \,.\, \mathrm{Cr}(t) \qquad (10)$$

t = term under analysis.
D(*t*) = dispersion of *t*.
Mf(*t*) = maximum frequency of *t* in a partition.
Cr(*t*) = number of partitions with a frequency of *t* less than *k* (*k* ~ 6).

6.4 Conclusions from this Chapter

This Chapter has shown a detailed description of the methodology used for the experiments that will be described in the next chapter, which is mainly how to build a co-occurrence graph and the mathematical properties that these graphs have. This does not mean that each of the experiments uses the same methodology. The basic algorithm described in this chapter serves as the basis for different prototypes used in each experiment, with their variants and adaptations needed for every different scenario. Adaptations will be specifically described in detail in each case.

The variant of graph used in this thesis is simple and liberal, since it applies no restrictions to the interconnections between nodes. Each graph is dominated by a node, which is a term from a query. Potentially, each node in each graph could be used as a query term to generate its own independent graph. It can happen, then, that two nodes include each other in their respective graphs. There may also be asymmetrical relations, in which a node *A* has a strong relationship with a node *B*, but *B* does not show the same relation, which is what one usually observes in the case of hypernymy relations, at least when the hyponym is not precisely the prototypical case of the class.

Chapter 7: Experiments

The previous chapter offered a description of a methodology for the analysis of syntagmatic relations. This chapter will apply that methodology to the study of not only syntagmatic but also paradigmatic relations, in order to obtain empirical evidence in support of the hypotheses presented in Chapter 2. Recall the two hypotheses of this research: the first hypothesis states that co-occurrence analysis of a given term can lead us to discover the essential features of the concept that is referred to by such term. The second hypothesis asserts that, via the first hypothesis, it is possible to identify, from the vocabulary of a given corpus, the subset of single or multiword expressions whose function in discourse is referential. As explained in Chapter 2, expressions can be referential in the sense that they designate a particular entity or concept in a real or imaginary world, as long as this entity is socially perceived or forms part of the common knowledge of some discipline. According to these hypotheses, the distributional analysis can thus be directly applied to concept analysis.

This chapter intends to provide empirical evidence for the above hypotheses by means of four experiments which resemble basic types of operations in concept analysis: classification, denomination, dissociation and association. These four types of operations are mutually dependent. However, their dependence is neither directed nor sequential. It would be naive to regard only one as a condition for the others because each of them needs, to a certain degree, the remaining three. One could think, for instance, that an ascending order of complexity would be the following: denomination → association → dissociation → classification, as one first knows that something has a name or is referred to by some name, and then associates such an entity with other entities, then dissociates this entity from other entities and, finally, classifies the entities in a unique taxonomy. However, the very act of naming something is already an act of association, dissociation and classification. One discriminates a part of the world by giving it a name, and by doing so gathers different things under the same label, which is an act of classification.

The first of the four types of experiments undertaken in this chapter is an algorithm for classification, where by means of co-occurrence statistics, given input terms are automatically assigned a hypernymy chain, i.e., an operation of automatic taxonomy extraction from corpus. This experiment should be seen as empirical evidence of the first hypothesis because in most domains, and especially in scientific domains, hypernymy is a relevant if not essential conceptual feature of a term. This experiment is

presented in Section 7.1. under the title of "Automatic Taxonomy Extraction".

The second type of experiments delves into the distinction between sense and reference. This is the type of experiments that should be regarded as empirical evidence particularly of the second hypothesis because in this scenario the task consists in determining whether a certain term or group of terms (e.g., the vocabulary of a document) is or is not referential. These experiments are presented in synchronic and diachronic variants, in Section 7.2. under the title of "Distinction between Sense and Reference".

In the context of what was identified as the task of "dissociation", at the beginning of this Chapter, the third type of experiments is devoted to the problem of disambiguation. Here, starting from a set of polysemous terms, the task consists in determining which are the possible senses that each of these polysemous terms can have. The task in this case is to separate the different instances of a given term into groups, according to each of the senses of the term. This experiment is presented in Section 7.3. under the title of "Analysis of Polysemous Terms".

The fourth and last experiment deals with association, where the idea is to associate expressions that refer to the same concept. Probably the most intuitive scenario for an experiment of this nature would be an automatic extraction of synonymy. However, in this experiment a more interesting -and difficult- goal has been set: to find equivalents of a given term in different languages. Bilingual terminology extraction is seen, in this context, as a special case of synonymy extraction. The experimental design in that case is that, for any input term in a source language, a program will return ranked list of candidates for the correct translation into a target language. Section 7.4. presents this experiment under the title of "Bilingual Terminology Extraction".

All four types of experiments are based on the methodology described in Chapter 6. However, it must be said that they also include their own methodological variations since in all cases they go beyond syntagmatic analysis. Thus, some experiments are slightly more complicated than what is presented in Chapter 6. In some cases the inverse is also true. For instance, in the first experiment, the methodology of syntagmatic analysis has been simplified considerably for practical reasons, mainly to save computational cost.

With respect to the corpus used to build a language model for these experiments, in the case of Spanish, a corpus of two million words was extracted from the category of 'general language' of IULA's LSP corpus (Vivaldi, 2009). The rest of the language models were extracted from the Wortschatz Project (Quasthoff et al., 2006). In the case of English, the corpus consists of two million words from the Wall Street Journal in the period 1987- 1992, The Financial Times in the period 1991-1994 and The Associated Press, from 1988 to 2006.

7.1 First Experiment: Automatic Taxonomy Extraction

This experiment is intended to gather empirical data that would serve as evidence for the first hypothesis of this research. Once again, the hypothesis is that for every referential expression there is a set of other terms that are syntagmatically related to it (such that their frequency of co-occurrence can be statistically assessed) and this set has an important overlap with the set of terms that make up the basic conceptual features of the referent of the term under analysis. Terms have, for instance, a high probability of finding their hypernyms among their most significant co-occurring terminological units, which occurs in particular in technical and scientific domains.

Chapter 2 introduced the notions of analytical and synthetical statements, which are behind the observable phenomenon of the syntagmatic association of terms with their conceptual features. The intuition is, thus, that analytical statements are prominent when compared to synthetical statements. Being prominent does not mean that they are more frequent, since synthetical statements are far more frequent. Analytical statements, or, more precisely, the terms which are involved in analytical statements, conform a minority, but the cumulative effect of their repetition makes them noticeable. In other words, synthetical statements are much more frequent, but at the same time they are much more diverse and, therefore, their terms do not accumulate. This use of the two types of statements is in an unstable equilibrium as a consequence of pragmatics and communicative skills. Authors, in general, provide definitions and descriptions of the terms they introduce when they assume that the majority of their readers are not acquainted with the terminology. Otherwise, if they suppose that this is the case for only a part of their readers, they at least provide conceptual features of a given term in the neighborhood of the term's first instances. This can be done with the aid of appositions and other grammatical devices as already shown in Chapter 2, which help the author to avoid any explicit assumption on the limits of their reader's knowledge on the matter.

The aim of the experiment presented in this section is to discover the hypernymy chain of a set of terms. Hypernymy is, naturally, not the only kind of conceptual feature for a given term, however it is one of the most important. Canonical definitions are based on hypernyms, as well as on other characteristics that make the concepts different from other concepts of the same category. This is called the "genus-differentia" kind of definition. The construction of a taxonomy, however, is possible just on the basis of a set of pairings of the type *X is hypernym of Y*. As the next subsection explains in more detail, the statement about the supposed hypernymic relation of *X* to *Y* will be a consequence of the assessment of the asymmetric statistical association of *X* and *Y*.

From a quantitative point of view, the hypernymy relation has two characteristics that are of significance for the task of automatic taxonomy construction: it is asymmetric and it is transitive. Its asymmetry leads to an unequal distribution of hyponyms and hypernyms of a given term in a predefined context window, which means that for any given term it is more likely to find its hypernyms than its hyponyms. Given the transitive nature of the hypernymy relation, if a given term *X* is hypernym of *Y* and that *Y* is hypernym of *Z*, then *X* is hypernym of *Z* too, at least ruling out accidental semantic crossings from one chain to another due to polysemous terms.

A final remark on the introduction of this experiment concerns the conditions that terms must fulfill in order to be subject of an experiment of taxonomy extraction:

(i) they must have a specific reference in a real or imaginary world.
(ii) they must be apt to be included as a class node in a taxonomy.
(iii) they must be recurrent in the web, i.e., form part of the vocabulary of a discourse well-represented in the web.

Condition (i) states that only referential expressions can be subjected to this analysis. This does not imply that they are the only type of units that can be placed in a taxonomy. Verbs, for instance, which have a predicative instead of a referential function, have hypernyms and can be placed in a taxonomy as well. However, it would not be appropriate to try to treat with this methodology non-referential units such as verbs, because the statistical properties that are studied with these methods are restricted to referential units. Condition (ii) simply means that the units to be analyzed can be comfortably placed in a taxonomy because they are concepts which can be defined and decomposed in its constituent semantic elements (the hypernym being one of these constituents). It is certainly difficult to find hypernyms for very general or abstract notions such as *time*, *space*, *being*, *universe*, etc., but these account for the minority of the terms in question. Condition (iii) states that, given that in this case the construction of the taxonomy is supposed to be taken from a web corpus,

the analyzed units must be well represented in the web; that is, they must have a minimum frequency of occurrence. The exact number will vary with each experiment, but a minimum "critical mass" of data is always necessary to perform statistical analysis. In the case of this experiment, a minimum of 15 contexts of occurrence has been set as a threshold to output a result.

7.1.1 Experimental Set-up

The proposed strategy for the construction of a taxonomy is grounded in distributional semantics because the elements of hypernymy relations are assumed to co-occur in the corpus in a syntagmatic context window of *n* words from each other with greater frequency than expected by chance. Terms tend to appear in more than one hypernymy relation. As a consequence, hypernymy relations between terms can be extracted from a corpus as a directed graph in which hypernyms can be identified as hub nodes which are targets of the arcs that originate in a number of hyponym nodes.

The experiment starts with the selection of a set of "seed terms", which are those terms whose hypernym is expected to be found. This represents a first difference with respect to the methodology explained in Chapter 6, where the analysis was cast on a term-by-term basis. In this experiment, the idea is also to perform the analysis term by term, but the difference is that the algorithm of this experiment will analyze a set of terms because in this way it is possible to obtain more information. With a set of input terms, instead of just one, the algorithm will use the results of the analysis of each seed term to reinforce the certainty of the results obtained with each individual term with the results obtained for the other terms. For experimentation, the way to gather seed terms is by random sampling from the set of terms of a given domain (or more than one[25]). In a real scenario, the set of seed terms would be the terms extracted from a document, from a dictionary, or selected according some criterion. Any term that fulfills conditions (i), (ii) and (iii), as already mentioned, can be chosen to be a seed term. Let us denote the set of seed terms as $T = \{ t_1, t_2 \ldots t_n \}$. Each t_i ($i \in [1 \ldots n]$) is submitted to three stages of processing:

Stage 1. Analysis of the first order co-occurrence which accounts for the terms (first raw list of hypernyms candidates) $H_i = \{ h_{i,1}, h_{i,2}, \ldots, h_{i,k} \}$ in the context window of t_i, where k is the number of hypernym candidates.

[25]It is to be expected that mixing terms from completely different domains will be detrimental for the precision of the results, and furthermore there is no practical reason to proceed in such way other than testing the robustness of the algorithm. In a real scenario, one would expect to organize a taxonomy of terms of the same domain or subdomain.

Stage 2. Analysis of the second order co-occurrence which accounts for $j \in [\ 1 \ldots k\ ]$, the terms $h_{i,j,1}$, $h_{i,j,2}$,..., $h_{i,j,m}$ in the context window of $h_{i,j}$, where m is the number of hypernym candidates of $h_{i,j}$.

Stage 3. Connecting t_i to a taxonomy of depth d using the co-occurrence links (H) found in stages 1 and 2 for all the set T.

Before a more detailed description of each of the stages, first the overall procedure should be sketched to account for the order and manner in which each stage is applied, as well as the description of the procedure to compile the co-occurrence vectors which are essential to the strategy. Figure 28 outlines the general design of the algorithm, which uses the function "Co-occurrenceList" detailed in Figure 29.

Automatic Taxonomy Extraction

```
T ← {seed term list} // initialize the list of starting terms
LM ← Language model // as described in section 6.3
H ← ∅ // initialize the hyperonymy stack
d ← depth // initialize the desired depth of taxonomy
for each t_i ∈ T do
  Lft ←Co-occurrenceList(t_i, LM)
  for each h_j ∈ Lft do
      H ← H ∪ Co-occurrenceList(h_j, LM)
  endfor
endfor
BuildUpTaxonomy(T, H, d)
```

Figure 28: Pseudo-code for constructing a taxonomy from a corpus

```
function Co-occurrenceList(t, LM) {
    Ct ←CompileCorpus(t,Web)
    Frt ←CreateN-gramFrequencyList(Ct )
    Frt ←FilterUninformativeVocabulary( Frt ,LM)
    Frt ←MergeSimilarForms(Frt)
    Frt ←EliminateOverlappingTerms(Frt)
    return Frt
}
```

Figure 29: Detail on the *N*-grams list function

The first for-loop in Figure 28 implements the analysis of first-order co-occurrence of each term t_i, which is performed by the function "Co-occurrenceList". It provides, for each seed term t_i, a ranked list of hypernym candidates. The second loop (analysis of second order co-occurrence), which is carried out by the same function, provides for each of the terms in the hypernym candidate list, its own hypernym candidate list. The result of the first and second order analyses should result in a set of <X> ISA <Y> statements. In the last stage, from this set of isolated statements, a connected taxonomy is constructed. The compilation of the co-occurrence vector, sketched in Figure 29, is in turn composed of different functions which perform essentially as the algorithm described in Chapter 6, however with minor variations in order to simplify and speed up the process. Details on each of these functions follow.

Compilation of an analysis corpus: The first step in the analysis of the first order co-occurrence of t_i consists in the compilation of an analysis corpus for t_i. This step corresponds to steps 1-3 of Section 6.2.2. The most straightforward way to obtain such a corpus is to retrieve it from the web, querying with a standard search engine for t_i. From the retrieved document result set, the context windows of a predefined size of all occurrences of t_i are extracted. The size of the context windows may be varied, depending on the genre and domain. The extracted context windows constitute the analysis corpus of t_i. Note that it is essential to omit t_i itself from the corpus since its presence would unjustifiably affect the frequency listing of n-grams used later on.

Creation of an *n*-gram frequency list: In the second step of the first order analysis, the vocabulary from the analysis corpus is

sorted in the form of *n*-grams, again as explained in Chapter 6. In this case, $n \leq 3$ seems to be a reasonable trade-off between the desired coverage of multiword terms and the processing time. However, $n > 3$ may also be considered, for instance, because the language or the domain in question favors compounds with a higher number of words. This step corresponds to Step 4 of Section 6.2.2. First, for different values of *n*, separate vocabulary lists are created, which are then merged into one list. Terms with a term frequency or document frequency less than 3 are discarded. To account for the fact that terms with a lower *n* tend to be much more frequent than terms with a higher *n*, the frequency of all *n*-grams (n = 1, 2, 3) in the merged list is normalized in that the absolute frequency of each term is multiplied by *n*.

Filtering uninformative vocabulary: To eliminate uninformative terms from the merged frequency list obtained in the preceding function, a filtering procedure is applied. This procedure corresponds to Step 7 in Section 6.2.2. Every candidates receives a weight on the basis of a language model built from a reference corpus which is, ideally, a corpus that reflects standard word usage. The filtering is done in three stages. In the first stage, terms that consist of less than *c* characters and multiword terms that contain an element of a size less than *c* characters are eliminated (in this experiment $c = 4$ characters). The decision for this procedure is empirically motivated. There is a need to filter out acronyms that in specialized discourse often follow multiword terms, e.g., *recombinant bovine somatotropin (rBST)*, *recombinant bovine growth hormone (rBGH)*, etc. This has been done just for convenience, it is obvious that more refined methods for acronym extraction could have been applied. In addition, one must be aware that this procedure is less useful in sub-domains such as diseases[26], where one is used to finding terms such as *hemophilia A* or *trisomy 22*. In the second stage, the reference corpus language model is used to compare the observed frequency of a term in the analysis corpus with the frequency this term has in the reference corpus (henceforth expected frequency). Any single word with an expected frequency higher than a predefined threshold *P* (in this experiment $P = 1000$) is eliminated. The case of the *n*-grams with $n > 1$ is slightly more difficult since one cannot assess all word combinations in a reference corpus. The workaround is to eliminate all *n*-grams that begin or end with a word with a frequency higher

[26]This is an issue that will need further development in future versions of this experiment. In fact, this oversimplification of the segmentation of terms explains some of the errors made by the algorithm during the evaluation (see Section 7.1.3).

than *P*. Also, if all elements of an *n*-gram have a frequency higher than (P/10), this *n*-gram is eliminated too. In the third stage, a new series of term deletions from the frequency list is performed by applying a score calculated according to Formula 11.

$$w(h_j) = \log (f_o(h_j) / f_e(h_j)) \quad (11)$$

h_j = *n*-gram from frequency lists

$f_o(h_j)$ = observed relative frequency of h_j in the analyzed corpus

$f_e(h_j)$ = expected relative frequency

In Formula 11, if h_j is any *n*-gram in the frequency lists compiled by the previous function and $f_e(h_j) = 0$, a sort of smoothing $f_e(h_j)=1$ is introduced if h_j is a single unit. But if h_j is a multiword unit, then the product of the expected frequencies of the individual elements of h_j is calculated (should an element *e* of *n*-gram h_j have $f_e(e) = 0$, again $f_e(e) = 1$ is assumed). If the score $w(h_j)$ has a value lower than a predefined *k* (in this experiment $k = -11$), then h_j is eliminated as a candidate.

Merging of similar vocabulary forms: As the model is aimed to be language independent, it involves neither lemmatization nor POS-tagging. This step does not differ from the explanation of Step 5 given in Section 6.2.2. In order to capture, at least approximately, inflectional and spelling variations and recognize when two words are forms of the same lemma, the model includes a kind of pseudo-lemmatization or normalization of the forms in the corpus and subsequent merging of similar forms.

Elimination of overlapping term sequences: If an *n*-gram has approximately the same frequency as the (*n*+*i*)-gram of which it is a part, then there is no need to keep both units competing to be included into the ranking list of *t* because it is likely that both denote the same concept; cf., for instance, *adrenal* and *adrenal glands* for such a case. If these two units have the same frequency in a corpus, this means that every time *adrenal* appears in the analyzed corpus, it appears followed by *glands*. Therefore, if two competing units show a total overlap (i.e., if one forms part of the other) and a similar observed frequency, the *n*-gram with the lower *n* is deleted.

Stage 1. Analysis of First Order Co-occurrence

For simplicity of notation, and after the explanation of how the loops are applied, this explanation leaves aside the consideration of the loop that runs through the seed term list *T* and then consider *t* as each term in *T*. First order co-occurrence of a term *t* accounts for its immediate syntagmatics, i.e., *t*'s syntagmatic association with terms that occur in the neighborhood (more precisely, in a context window of a predefined size to the left and to the right) of *t*. The analysis of the first order co-occurrence implies the function Co-occurrenceList, explained above. The result of the first order co-occurrence analysis for each seed term *t* is a list of *n*-grams h_1, h_2,..., h_k, which are syntagmatically related to *t* and which tend to co-occur with *t* with a higher frequency than expected by chance. In accordance with the stated hypothesis, among h_1, h_2,..., h_k a certain number of hypernyms of *t* are expected to be found.

Stage 2: Analysis of Second Order Co-occurrence

The result of the second order co-occurrence analysis for *t* is a list of terms $h_{j,1}$, $h_{j,2}$,..., $h_{j,m}$ that are syntagmatically related to *t* and/or to other h_j $j \in [\ 1\ ...\ k\ ]$ resulting from Stage 1. For illustration, consider that seed term *t* is *catostomus ardens*. Applying first order co-occurrence analysis, the result is a hypernym candidate list for *catostomus ardens* that includes *Utah sucker*, *sucker*, *catostomus* and *catostomus bernardini*, among others. Now, in the second order co-occurrence analysis, the first order analysis is applied to each element in the candidate list -as if they were seed terms- in order to find out which terms are related to *Utah sucker*, *sucker*, *catostomus*, etc.

The procedure in stage 2 is almost the same as in stage 1, however stage 2 provides new information that is useful in order to have a better assessment of which candidates should be discarded and which should be kept for the construction of the taxonomy. This new information is achieved by three means: (I) the notion of term dispersion (*D*), (II) the notion of term weighted dispersion (wD), and (III) higher punctuation of overlapping sequences. Let $h_{j,l}$ ($l \in [\ 1\ ...\ m\ ]$) be one of the candidates obtained in stage 2.

> (I) The coefficient $D(h_{j,l})$ captures the ratio of the distribution of a hypernym candidate across the different hypernym candidate lists. This coefficient must not be confused with the Dispersion coefficient used in Formula 10 in Chapter 6. They both measure how dispersed a unit is among a series of partitions, but the context of application in this case is slightly different. In this case, the coefficient simply counts the recurrence of a candidate in a series of trials. For instance, in the analysis that corresponds to the seed term

catostomus ardens, there are 28 terms which have a significant (first-order) co-occurrence with it, such as *Utah sucker*, *sucker*, *catostomus*, etc. This means that there will be 28 trials in the second-order analysis. Subsequently, when inspecting which terms co-occur with each of these 28 terms, it is found that the term *fish* shows a significant co-occurrence with 12 of these 28 terms. Let m be the number of hypernym candidate lists (28 in the example) and d the number list where a candidate $h_{j,l,}$ is present (12 in the example), then $D(h_{j,l}) = d/m$ (0.42 in the example). This means that *fish* appears as a hypernym candidate for *catostomus ardens* but also for *Utah sucker*, *sucker*, *catostomus*, etc. The intuition is that a higher dispersion of a term implies a greater probability of it being the most generic hypernym.

(II) The weighted dispersion of a hypernym candidate, $\mathrm{wD}(h_{j,l})$, takes into account the overall observed frequency of $h_{j,l}$ in the analysis corpus (recall that high frequency and uninformative terms in the vocabulary have already been eliminated), multiplied by its dispersion, as expressed in Formula 12:

$$\mathrm{wD}(h_{j,l}) = \log(1 + f_o(h_{j,l}) \times D(h_{j,l})) \qquad \begin{array}{l} f_o(h_{j,l}) = \text{observed frequency of } h_{j,l} \text{ in the analyzed corpus} \\ D(h_{j,l}) = \text{dispersion of term } h_{j,l} \end{array} \qquad (12)$$

(III) Overlapping sequences[27] are assigned a higher score. Thus, if t and one of its hypernym candidates $h_{j,l}$ show a total overlap (if there is a full string matching between $h_{j,l}$ and t such that $h_{j,l}$ is completely contained in t, e.g., *catostomus* and *catostomus ardens*), then it is likely that, in fact, $h_{j,l}$ is a hypernym of t. As a consequence, the weighted dispersion score of $h_{j,l}$ is increased by 25 points—an empirically determined figure, which is approximately a quarter of the average score of the 15 best-scored hypernym candidates.

[27]As in the previous note, care must be taken not to confuse this case of overlapping term sequences with the previous case ("Elimination of overlapping term sequences") in stage 1, where forms of different size were collapsed together for being considered parts of a same term. This was done to prevent these parts of a same term for competing for the same positions in the rank. In this case, it is the analyzed term and those hypernym candidates which are being compared to find an overlap and to increase the punctuation of that pairing if the overlap exists.

As a result, each term $h_{j,l}$ is thus assigned two scores: (1) the overall score shown in Formula 13, (2) the dispersion score. With these two scores at hand, the final qualification of the terms in the hypernym candidate lists is carried out: If a term has a $D < 2$, it is removed from the candidate list (since it is unlikely to be a hypernym). The remaining high dispersion terms are ranked with respect to their score s; only the terms with the 10 highest s score are kept.

$$\mathrm{s}(h_{j,l}) = \mathrm{wD}(h_{j,l}) + B \qquad \begin{array}{l} B = \text{“overlapping bonus” of 25 if} \\ \quad t \text{ and } h_{j,l} \text{ show a total overlap} \\ \quad \text{(and of 0 otherwise)} \end{array} \tag{13}$$

Stage 3: Final Taxonomy Construction

After the first and second order analyses, the result is a set of first and second order weighted hypernym lists for each term in the seed terms list. With the processing of each additional seed term, a unified taxonomy, which has the form of a set of statements of the type <X> ISA <Y>, is developed as more statements are added after each experiment run. Each run also makes the taxonomy more reliable since the relations between the statements are reinforced. It is assumed that the bigger the sample of seed terms is, the more complete and reliable the taxonomy becomes.

From these lists of hypernymy statements, the taxonomy is derived by taking each seed term and reconstructing the ascending hypernymy chain. Starting from each seed term, two tasks need to be accomplished: (i) the establishment of the ISA-relation between pairs of terms, (ii) hierarchical ordering of the ISA-relations in order to obtain the taxonomy. For the first task, the feature of asymmetry of ISA-relations is exploited; for the second task, the feature of transitivity is exploited. The central criteria for the selection of a term h as a hypernym of the term t are: (i) that h is in the hypernym candidate list of t and (ii) that h is “general” enough. With respect to (i), it may also occur that in the final taxonomy t ends up as hyponym of a term that was not included in the candidate list thanks to the transitivity property. With respect to (ii), the generality of a term h is reflected by the number of times it appears in the hypernym candidate lists of other terms in the seed terms list. Thus, the best h for t is calculated by wT(t) in Formula 14.

$$w_T(t) = |H| + \sum_{i=1}^{|H|} w_T(h_i),\ h_i \in H, \tag{14}$$

In Formula 14, variable H is the set of hypernym candidates of a term t. The term with the highest wT with respect to t (henceforth the most generic hypernym candidate term of t) qualifies as hypernym of t. With the hypernym h of t determined, it can be assumed that all members of the candidate list of t that have h as the most generic term are also hypernyms of t. Path-based inference is thus the tool to use for the hierarchical ordering of the ISA-relations and thus the actual construction of the taxonomy. Starting from a number of seed terms T, the procedure is as follows:

1. introduce an ISA-link between each $t \in$ T and its dominant (most generic) term h
2. $\forall\ t' \in H_{\cup}$ (with $H_{\cup}$ being the set of the available hypernym candidate lists) with h as dominant term: introduce an ISA-link between t and t'
3. apply the transitivity feature of ISA: $\forall\ t$, t' and t'' with t ISA t' and t' ISA t'': introduce an indirect ISA-link t ISA t''
4. repeat from 1., with the established hypernyms of the ts $\in T$ as new T until the desired depth of the taxonomy has been reached.

Consider, for illustration, the construction of a taxonomy of depth 3 for the seed term *somatrem* displayed in Figure 30. The first move is to draw a link between *somatrem* and the most generic candidate which in this case is *hormone*. Then draw hypernymy links from *somatrem* to the rest of the candidates of the list which in turn have *hormone* as a hypernym candidate. These terms are: *somatropin*, *human growth hormone* and *growth hormone*. These nodes may also have hypernym candidates present in the graph, as it is the case with *somatropin* and *human growth hormone* with respect to *growth hormone*. The process can be iterated in this way until there are no more links to draw or until a desired depth d is achieved.

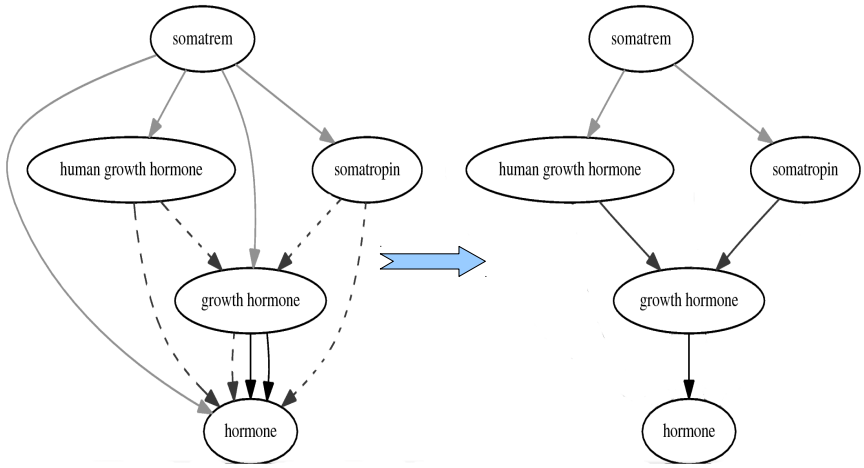

Figure 30: Automatically generated taxonomy for "somatrem"

The left hand side figure shows the connections which describe the kind of path based inference, that is, the number of times that a certain path between nodes have been traversed. This number is represented in Figure 30 by the number of arrows coming into a node. The node for *hormone* is in this case the one with the largest number of incoming arrows and is therefore selected as the most generic of the available terms. Differences in the layout of the arrows express the number of iterations or, in other words, the distance in nodes from the initial node. Thus, for *somatropin*, the algorithm detects that it has *growth hormone* and *hormone* as hypernyms, which are marked with light arrows. As a consequence, direct ISA-links are established between *somatropin* and these nodes. *Human growth hormone* has *growth hormone* and *hormone* as hypernyms, which is again reflected by the corresponding direct ISA-links. These arrows are drawn as dashed lines. The same is done with *growth hormone*, which has as only hypernym *hormone*, drawn in solid black lines. Since the term *hormone* does not show any hypernym in the sample, it remains the most general term. The node of the term *hormone*, which is the most generic hypernym in this chain, is precisely the one that has most incoming arrows. On the right hand side of Figure 30, there is the same taxonomy chain, but this time without the indirect links between nodes since at this point they are deemed redundant.

7.1.2 Results

In order to test the strategy, a number of experiments were carried out extracting hypernym relation instances from corpora downloaded from the web and constructing taxonomies from the extracted instances. This subsection shows the results obtained in each stage of the algorithm after

an experiment using a random sample of 100 seed terms from the SNOMED[28] medical ontology (see section 4.7) as input. The sample is divided in equal parts into the following four arbitrarily selected categories: *hormones*, *virus*, *fishes* and *diseases*. The results are shown for only one of the terms in that seed collection: *hybopsis gracilis,* which is a kind of fish. Comments on the overall performance will be given in the Evaluation, in section 7.1.3.

The reduced number of trials was chosen not only for the purpose of reducing computational effort and time of execution, but because it is necessary for a thorough qualitative result evaluation. However, it must be borne in mind that the results for any individual term depend on the results of the other seed terms. Therefore, there must be a balance between cost of processing and certainty of hypernymy links in the extracted taxonomy given that the certainty of the taxonomy increases as more seed terms are included in the sample.

In all experiments, the document collections for the compilation of corpora for the individual terms, from which the hypernyms of these terms are then distilled, have been obtained from the web using the Yahoo Boss Api[29]. The corpora consist of snippets provided by the search engine for each document in the retrieved document collection. The decision to use snippets instead of retrieving the context windows for each term from the documents themselves was motivated by two considerations. First, snippets already are context windows of the terms. Second, there are reasons to avoid the high pre-processing costs that come with the compilation of a clean web-based corpus: downloading, conversion of different document formats to plain text, character encoding detection and conversion, etc. For practical reasons (speed of processing), the sample was reduced to the first 100 snippets per term, although it is to be expected that higher amounts of information would result in better performance.

Results of Stage 1 (Analysis First Order Co-occurrence)

Table 19 shows an example of the analysis of first order co-occurrence for the term *hybopsis gracilis.* In this table, *n*-grams are ordered by

[28]In this experiment, the Spanish version of the SNOMED ontology, however it also has terms in English and it is only the English part the ontology the one that was used. This version only contains fully specified names (that is, terms that unambiguously identify concepts) leaving aside shortened terms as well as synonyms,

[29]The Yahoo Boss Api interface to Yahoo's Web index. http://developer.yahoo.com/search/boss/boss_guide/ [accessed June 2010].

decreasing frequency. The table also shows the weight of each *n*-gram, as defined in Formula 11 of Stage 1. Shadowed cells are eliminated candidates, in the first case for the detection of total overlapping with an already registered term with similar frequency (see section 7.1.1) and, in the second case, because of a low informational score, mainly due to the low information of components such as *North* and *American*. This is also an example of the kind of errors that can be made with the weighting measure, since "North American Cyprinid" has a weight below the threshold of -11 set for this experiment, however it is one of the relevant hypernyms of *hybopsis gracilis*. Leaving this case aside, some other candidates in the list are correct hypernyms, such as *fish, chub* and *cyprinid fish*.

rank	***n*-gram**	**frequency**	**weight**
1	flathead chub	60	4.09434
2	chub	58	4.06044
	flathead = flathead chub	60 >= 33 (29.7)	3.49651
3	**fish**	23	-6.45609
	north american cyprinid	21	-17.39015
4	gracilis	20	2.99573
5	platygobio gracilis	18	2.89037
6	hybopsis	15	2.70805
7	cyprinid fish	14	-2.15673
8	species	13	-4.09946
9	mus nat hist	12	2.48491
10	univ of kansas	12	-2.14007

Table 19: First Order Co-occurrence for "hybopsis gracilis"

Table 19 also shows some false candidates that were not eliminated in this stage of the analysis, such as the acronym of "Museum of Natural History" in position 9 of the ranking, or "University of Kansas" in the next position. The truncation of these forms makes them "rare words" from the perspective of the language model commented in Section 6.3.

Results of Stage 2 (Analysis Second Order Co-occurrence)

Table 20 shows the ranked hypernym candidate list offered by the algorithm again for the seed term *hybopsis gracilis*. In addition to the columns for frequency and weight, which were available in the tables of the previous stage, this table shows now the candidates ranked by the weighted dispersion 'wD' (Formula 12) in Stage 2. The information provided by this coefficient ameliorates the results obtained for the first order co-occurrence, given that there are less irrelevant candidates. Interestingly, one of the terms that has replaced them, *platygobio gracilis*,

is a synonym of the seed term[30]. It is not surprising, though, since Table 20 shows paradigmatic similarities while Table 19 shows syntagmatic relatedness.

rank	form	frequency	weight	dispersion	weighted dispersion
1	fish	30	-6.19038	3	10.302
2	flathead chub	60	4.09434	2	8.222
3	chub	58	4.06044	2	8.155
4	gracilis	20	2.99573	2	6.089
5	platygobio gracilis	18	2.89037	2	5.889
6	hybopsis	15	2.70805	2	5.545
7	cyprinid fish	14	-2.15673	2	5.416
8	species	13	-4.09946	2	5.278
9	hybopsis x punctata	12	0.00000	2	5.130

Table 20: Second Order Co-occurrence for "hybopsis gracilis"

Results of Stage 3 (Final Taxonomy Construction)

From the best ten hypernym candidates obtained in Stage 2, only one term is selected as the most generic hypernym[31]. As already explained, the selection is determined by the coefficient wT, which is the number of times that a given term has been appointed hypernym candidate in the complete set of trials triggered by the seed terms. As shown in Table 21, the most generic of the available term is *fish,* thus the taxonomy chain is built upon that hypernym. This is because *fish* has been appointed hypernym of some other term 181 times in total (taking into account the terms that co-occurr with the terms that co-occur with the seed terms, as already explained). Among the ten best candidates, there are terms which are semantically relevant to the term in question, *hybopsis gracilis,* but are not a hypernym. This is the case of *flathead chub*, which is the common name of this creature. The term *species* would have been an unfortunate selection since it is too general. The term *chub*, however, would have been a good selection.

[30]http://www.bio.txstate.edu/~tbonner/txfishes/platygobio%20gracilis.htm [accessed June 2010].

[31]As it will be explained in the conclusion of the experiment, no attempt to solve potential problems of polysemy were made at this point. Polysemy is treated more extensively in the third set of experiments of this Chapter.

#	candidates	wT
1	fish	181
2	species	32
3	chub	19
4	wildlife	18
5	cyprinid fish	1
6	flathead chub	1
7	hornyhead chub	
8	richardson	
9	hybopsis	

Table 21: Candidates of Stage 2 re-ranked according to coefficient wT

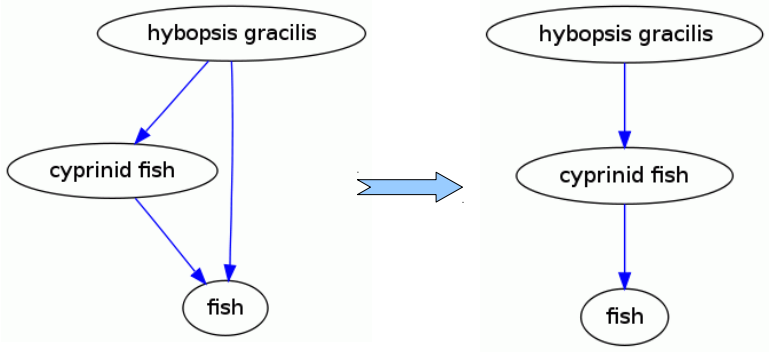

Figure 31: Taxonomy of two levels for "hybopsis gracilis"

The taxonomy chain in Figure 31 is a two level taxonomy having an intermediate node *cyprinid fish*, since it is a node in the candidate list, which itself has the node fish as a hypernym candidate itself. This is not the case for other candidates in Table 21. The units *wildlife* and *species*, for instance, never had *fish* as a hypernym candidate after all the trials triggered by the seed terms, and therefore they are not introduced in the taxonomy graph. This is unfortunate in some other cases. For instance, there are no links between *chub* and *fish* in the taxonomy built so far. This may be a problem of sampling, explained by the fact that there are not enough occurrences of *chub* in the selected sample of seed terms. In addition, *chub* is affected by the polysemy of its use[32]. This is clear indication of the influence that the selection of the seed terms can have on the results. It is likely that a selection of highly related seed terms have a positive impact on the quality of the resulting taxonomies. And it is also

[32]According to Wikipedia, a chub is "an overweight or obese gay male who identifies as being part of the related chubby culture" http://en.wikipedia.org/wiki/Chub_(gay_culture) [accessed June 2010].

true that a better selection of the analyzed corpus -instead of internet search engines- would prevent some cases of ambiguity.

7.1.3 Evaluation

The previous section offered an example of a result to show what the taxonomies look like. This section, in contrast, presents an assessment of the quality of the results of stages 1-3 described above. For convenience, the experiments have been limited to three links in the hypernymy chain. Note, however, that in principle there is no theoretical reason for not building deeper taxonomy chains.

Evaluation of Stage 1 (Analysis First Order Co-occurrence)

The goal of the first stage of these experiments was to estimate the probability of a specialized term co-occurring with its hypernyms. In order to do so, it is necessary to test how possible it is to replicate hypernyms of terms represented in off-the-shelf taxonomies in the area of medical terminology. The taxonomies, in this case, are WordNet and SNOMED.

Given that WordNet is a general purpose lexical database and SNOMED a medical terminology repository and ontology, it is to be expected that the latter is more specific and comprehensive with respect to medical terminology than the former. That is, while common terms such as *enzyme* or *protein* are likely to occur in both resources, but specialized terms (such as *nasogastric intubation for enteral infusion of concentrated nutritional substances*) are likely to be found only in SNOMED. As a consequence, it will be difficult to find exact matches among the hypernyms provided by this strategy, those provided by the SNOMED-taxonomy and those provided by the WordNet taxonomy.

Randomly selected samples of seed terms from WordNet and SNOMED were used in the experiments of automatic hypernym extraction. In the case of WordNet, they were terms related to the field of medicine. In the case of the SNOMED, the samples were from three arbitrarily selected categories: *organisms*, *substances*, and *disorders*. They are random selection of terms under these nodes in the hypernymy chain of SNOMED. In order to somewhat relax the evaluation criteria, two levels of hypernyms of the corresponding seed terms (i.e., mother and grandmother) were included in the generic term samples in the case of WordNet, and three levels of hypernyms in the case of SNOMED. Table 22 summarizes the size of the samples.

	WordNet	SNOMED		
		organisms	substances	disorders
Seed term sample (#)	447	95	91	81
Generic terms sample (#)	504	287	339	531

Table 22: Size of the samples in WordNet and SNOMED in Stage 1

Tables 23 shows the qualitative figures obtained in the different runs of Stage 1. There is a distinction between exact (or complete) and partial matches between seed term–hypernym tuples proposed by the algorithm and seed term–hypernym tuples from the reference taxonomies.

A match is a complete match if the seed term–hypernym tuples identified by the algorithm are exactly the same as one proposed by the ontology. A match is a partial match if the algorithm correctly identifies the base element of one or several of the generic multiword terms of a seed term as its possible hypernym(s). For instance, given the term *hybopsis gracilis*, SNOMED offers the following hypernyms: *family cyprinidae* and *carps and/or minnow* as 1st order hypernyms and *fish* as 2nd order hypernyms. If the algorithm includes *fish* into the hypernym candidate list, that is considered a complete match. If it includes *minnow* then there is a partial match because the term *carps and/or minnow*, found among the SNOMED hypernyms.

run	compl. Match (%)	compl. Match (mean rank)	part. Match (%)	part. Match (mean rank)
WordNet	49	4.94	20	5.33
SNOMED (organisms)	25	9.12	49	4.04
SNOMED (substances)	39	5.41	41	3.71
SNOMED (disorders)	30	5.08	81	3.92

Table 23: Results of Stage 1

The figures in Table 23 are to be read as follows: for the WordNet run, on the average, 49.0% of the first order co-occurrence terms show a total match with the generic terms from the reference taxonomy. The columns that refer to the average position in the ranking list are also important because they express how the first order co-occurrence terms were

weighted. If the mean rank is 4.94 (as for the terms from WordNet), it is that a hypernym is found, on average, around the fifth position of the rank. On the following column, the percentage of partial matching is 20%, meaning that only in 20% of the cases there was a partial match between some WordNet hypernym and some of the five most frequent co-occurrents of the input term. It is important to take into account that both percentages are independent, that is, there is no reason to expect that the sum of both should be less than 100, and in fact it is greater in some cases, as in the category of *disorders*. In that case, 30% of the trials returned full matches and 81% of them returned partial matches. This is because in a same trial it is possible to obtain both complete and partial matches, therefore percentages can show an overlap.

In order to compare the results obtained in stage 1 with the ones obtained in Stage 2 (shown later in Table 25), evaluation of stage 1 is repeated with a smaller sample which is also used in the evaluation of Stages 2 and 3. This is the sample of 100 terms already mentioned in the section 7.1.2, the one that is divided into *hormones*, *viruses*, *fishes* and *disorders*. Only 82 of the 100 seed terms are found in the search engine with sufficient data, and therefore evaluation is conducted only over the appearing subset. These results are shown in Table 24. The greater variability might be explained by the fact that the size of the sample is smaller.

SNOMED run	compl. Match (%)	compl. Match (mean rank)	part. Match (%)	part. Match (mean rank)
hormones	47	2	21	2.6
viruses	8	1	78	1.83
fishes	56	2.93	13	6.66
disorders	15	6	69	3
Average	31.5	2.98	45.25	3.52

Table 24: Results of First Order Co-occurrence on a small sample

The most important information that tables 23 and 24 provide is the percentage of cases in which, on average, one can expect to find a correct hypernym among the first five positions of the ranking of the first-order co-occurrence list. It is an estimation of how likely it is that a term and its hypernym will co-occur frequently in the same context (at least in scientific literature). However, it does not reveal anything about the quality of the overall results. These are, therefore, not to be regarded as figures of failure and success. The fact that there are cases where there is no frequent first order-coccurrence between term and hypernym does not

imply that the hypernym is not going to be found at some point later in the process.

Evaluation of Stage 2 (Analysis of Second Order Co-occurrence)

The goal of the experiments of Stage 2 is to assess how well second order co-occurrence analysis identifies hypernym candidates for terms within the hypernym candidate list of a seed term. As explained in Section 7.1.1, the algorithm starts from a sample of seed terms and performs first order analysis for each of them. After that procedure is accomplished, the algorithm performs the second order analysis for each term in the hypernym candidate list obtained from the first order analysis.

Consider Table 25, which shows the results obtained in the second order co-occurrence analysis, in comparison to Table 24, the cases of the first order co-occurrence analysis for the same sample of seed terms.

SNOMED run	compl. Match (%)	compl. Match (mean rank)	part. Match (%)	part. Match (mean rank)
hormones	47	2.09	21	2
viruses	8	1	78	2.38
fishes	69	2.18	8	8
disorders	23	3	92	3.41
Average	36.75	2.06	49.75	3.94

Table 25: Results of Second Order Co-occurrence of the small sample.

In general, values in the second-order analysis are slightly better than in the first-order analysis. The percentages of complete matching is the same in the cases of *hormones* and *viruses*, but it increases 13 points in the case of *fishes* and 8 points in the case of *disorders*. The rank of hypernyms is also better in the case of *fishes*, which advanced from the third to the second position in the rank, and in *disorders*, from the sixth position to the third. The same pattern is followed in the cases of partial matching. There is no difference in the percentage of retrieval of hypernyms in the first two categories, *hormones* and *viruses.* The case of *fishes* is the only one that shows a difference, because it shows a decrease from 13% in first-order co-occurrence to 8% in the second-order. In the case of *disorders*, however, there is a sharp increase from 69% to 92%. Manual examination of the results reveals that these great difference in the percentages of complete and partial matching between categories is due to the different nature of the terms of each sub-domain. The cases where there is low full match are those where there is more syntagmatic terminology. This is

most evident for instance in the comparison between *fishes* and *disorders*. In the case of *fishes*, multiword terminology is generally shorter in number of words than in *disorders*. In the case of *disorders* it is common to find terms of more than five components, while in the case of *fishes*, multiword terminology is mostly two words long. This explains why it is much more likely to find a full match in the name of a fish than in the name of a disorder.

As already mentioned in the evaluation of first-order analysis, these figures still do not show the rate of failure or success of the experiment, but only the percentages where hypernyms are found in second-order co-occurrence in individual cases. For instance, the fact that at this point in only 8% of the cases there is a complete match of hypernyms in the case of *virus* does not mean the taxonomy will not be built in 92% of the experiments with *viruses*. As has been mentioned already, it is from the whole set of seed-terms that the information for the taxonomies is extracted and not from the individual cases. The relations between terms in the whole set is stored in a single table of hypernym candidate pairs, mutually reinforcing the certainty of information obtained for individual cases.

Evaluation Stage 3 (Final Taxonomy Construction)

In the last stage of the experiments, the trial consisted in taking the outcome of the second stage experiments with the set of seed terms to construct an ascending hypernymy chain, i.e., to search for their direct hypernyms and then the hypernyms of the hypernyms from the ranked candidate lists obtained during the previous two stages.

In this stage, where the taxonomies are finally built, there is a remarkable improvement, showing a correct pairing with a hypernym of at least one level in 90% of the cases. A taxonomy of one level is a taxonomy where the input term is associated with only one hypernym. However, the generated taxonomies can have different extensions; i.e., they may have a different number of nodes in their chains. Figure 32 shows an example of a one level taxonomy, stating that *carpenter syndrom* IS_A *disease*. Figure 35, below, shows an example of a three level taxonomy. In 46.6% of the cases, the taxonomy chain had only one level; in 29.3% there was a chain of two levels; in 21.3% there was a taxonomy of three levels and more than three levels in the rest of the cases.

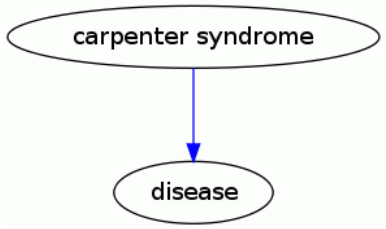

Figure 32: Taxonomy of one level for "carpenter syndrome"

The quality of the results of the overall algorithm for the construction of taxonomies can be evaluated in terms of precision and recall. Precision reflects the ratio of times that the algorithm was correct in assigning a hypernym to a term. Recall reflects the ratio of hypernyms provided by the reference ontology that were also provided by the algorithm, including terms that were correctly assigned as hypernyms by the algorithm and were registered as such by some authoritative source, not only the SNOMED-taxonomy[33]. With respect to this method of evaluation, there is an important distinction in the figures of precision and recall between two types of assigned hypernyms: the most generic hypernyms and the intermediate hypernyms. In each trial, there may be different assignments of hypernyms to seed terms and hypernyms assigned to these other hypernyms as well, however there is always one hypernym which is dominant (as said before, the dominant node is the most generic term available in the taxonomy chain). Consider again the taxonomy for *somatrem*, now repeated as Figure 33. In this case, the dominant or most generic node is *hormone*, which is the most generic of the available terms related to the seed term *somatrem*. The fact that *somatrem* is classified as a *hormone* is a successful trial (recall that only 25% of the sample of seed-terms are hormones). This is also a case where not only the most generic term but also the intermediate hypernym assignments are correct, thus the whole hypernymy chain is free of error (*somatrem* is a *somatropin* and is a *human growth hormone* which in turn are *growth hormones*). This is an important achievement considering that there is no prior knowledge about

[33]Some of these sources are the databases from the Smithsonian Institution (http://collections.si.edu/search/results.jsp/), the website of the Texas State University's Department of Biology, (http://www.bio.txstate.edu/), the Iowa Rivers Information System of the University of Iowa, (http://maps.gis.iastate.edu/iris/), the U.S. Food and Drug Administration (http://www.fda.gov/Drugs/default.htm) the e Therapeutics Initiative (TI) of University of British Columbia, the Department of Health and Ageing Therapeutic Goods Administration of the Australian Government (http://www.tga.gov.au/) and the Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/), among others [all accessed on June 2010].

what classes there are, or how many of them or what proportion of terms per class is to be expected. This is a completely different scenario from supervised categorization, where, for instance, a algorithm knows that there are four categories (say, *hormones*, *virus*, *fishes*, *diseases*) and each element has a probability of being correct just by chance. In this case, every word or *n*-gram has the same probability of being the most generic term, thus, the chances of having a correct hypernymy by random assignment are practically null.

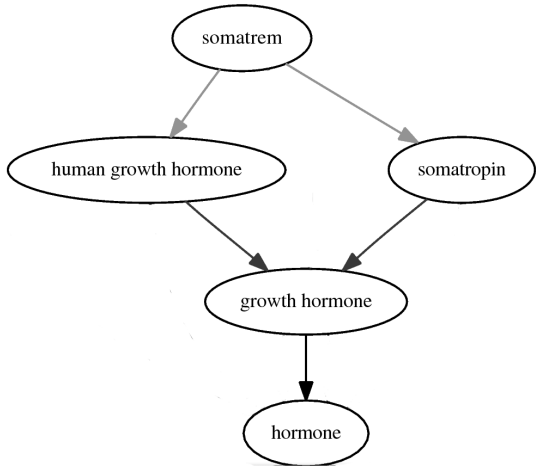

Figure 33: Automatically generated taxonomy for "somatrem"

Figures of recall are, however, much lower. This is to be expected, since the human-made taxonomies are more complex and include more nodes. Compare, for instance, the taxonomy made by the algorithm in Figure 33 with the taxonomy for the same term provided by the SNOMED-taxonomy, shown only partially in Figure 34, until the *hormone* node. The first and most important difference is that, in the case of SNOMED, the hypernymy structure is more complex and therefore with larger number of nodes and with multiple inheritance in the hypernymy chains. This is to be expected because there exists a normal level of description in scientific literature. Authors will provide some hypernyms of a term as descriptors when the term is introduced in discourse, however it is unlikely that authors will include hypernym for every level in the taxonomy chain in their descriptions of the concepts they are referring to. Another difference in the comparison between the taxonomy made by the algorithm and the one provided by SNOMED in the case of *somatrem* is the absence of the term *growth hormone* in SNOMED. It is, however, a correct hypernym, as *somatrem* is a synthetic type of growth hormone[34]. It is not an error of the SNOMED ontology, it is just a decision of their designers to include a more technical alternative like *pituitary hormone*.

[34]The Human Growth Hormone Glossary
(http://www.somatropin.net/glossary.htm [accessed June 2010]).

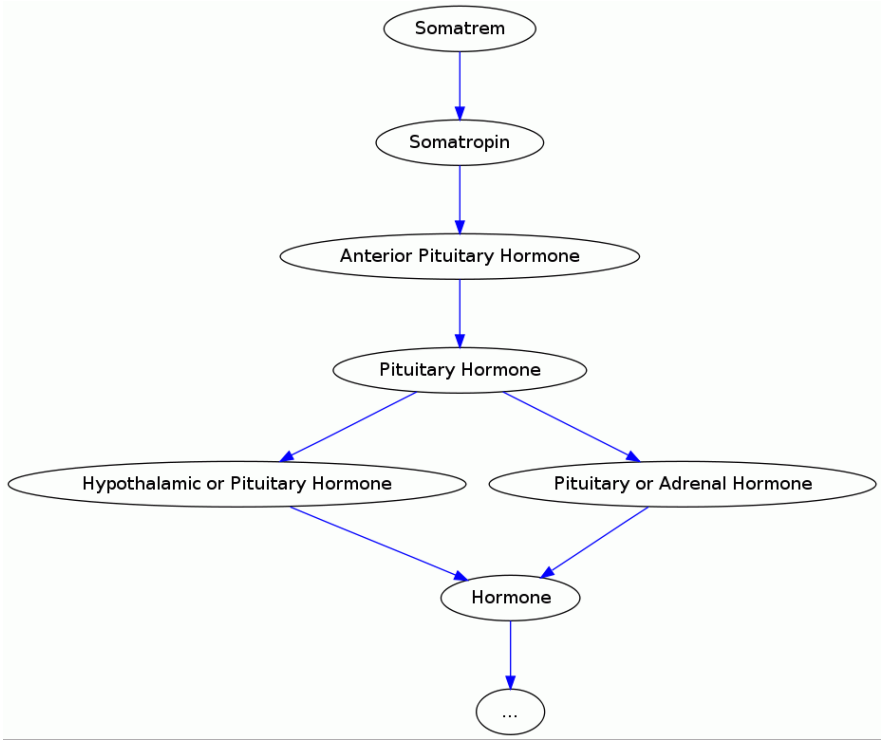

Figure 34: Partial taxonomy for “somatrem” by SNOMED

Consider another example with the term *monomeric prolactin*, whose taxonomy obtained with this algorithm is depicted in Figure 35. As it was shown before in Figure 30, the taxonomy chain is presented twice: at the left hand side of the figure showing the paths that are traversed from node to node as well as the number of times that each path is traversed. Light arrows represent the first iteration, dashed line arrows represent the second iteration and solid black line arrows represent the third and final iteration. For readability, the right hand side of the figure the taxonomy chain appears stripped from indirect links.

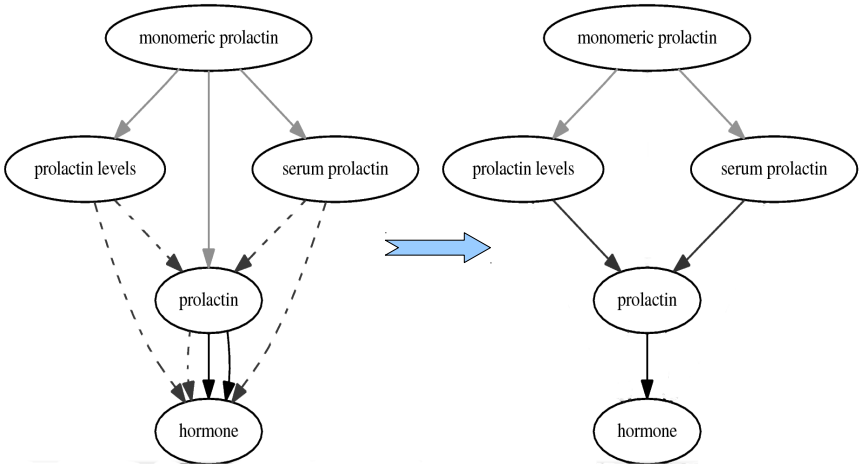

Figure 35: Three level taxonomy generated for "monomeric prolactin"

In the case of figure 35, the first two nodes, *prolactin levels* and *serum prolactin*, are wrong. Both have been originated by the same term, which is *serum prolactin levels*, which is not a hormone, but rather a procedure. This problem has occurred during the process of segmentation. Instead of having a unique trigram, *serum prolactin levels*, the sequence was treated as two independent bigrams. Possibly, with a better terminology extraction procedure, fewer errors of this type would have been committed. Despite the term extraction problems, the taxonomy assignment is in fact correct. It can be seen that the overall assignment of hypernymy to the seed term *monomeric prolactin* is correct because the most generic hypernyms *prolactin* and *hormone* are correct. For this reason, the most generic hypernymy assignments are counted separately from the rest of the assignments, given that it is more important to assign a correct generic hypernym than a middle level hypernym. That is, it is important that the seed term was identified as a type of hormone. It would have been much worse if the error were in this level of the hierarchy, for instance, if the algorithm had proposed that *monomeric prolactin* is something else, like a kind of fish, a procedure, etc. For this reason, in the evaluation there is, on one hand, a number of precision for the most generic hypernyms and, on the other hand, an average proportion of precision in the assignment of hypernyms inside each taxonomy. Again, in comparison with the taxonomy provided by SNOMED for the same term, shown in Figure 36, there are many hypernyms lacking.

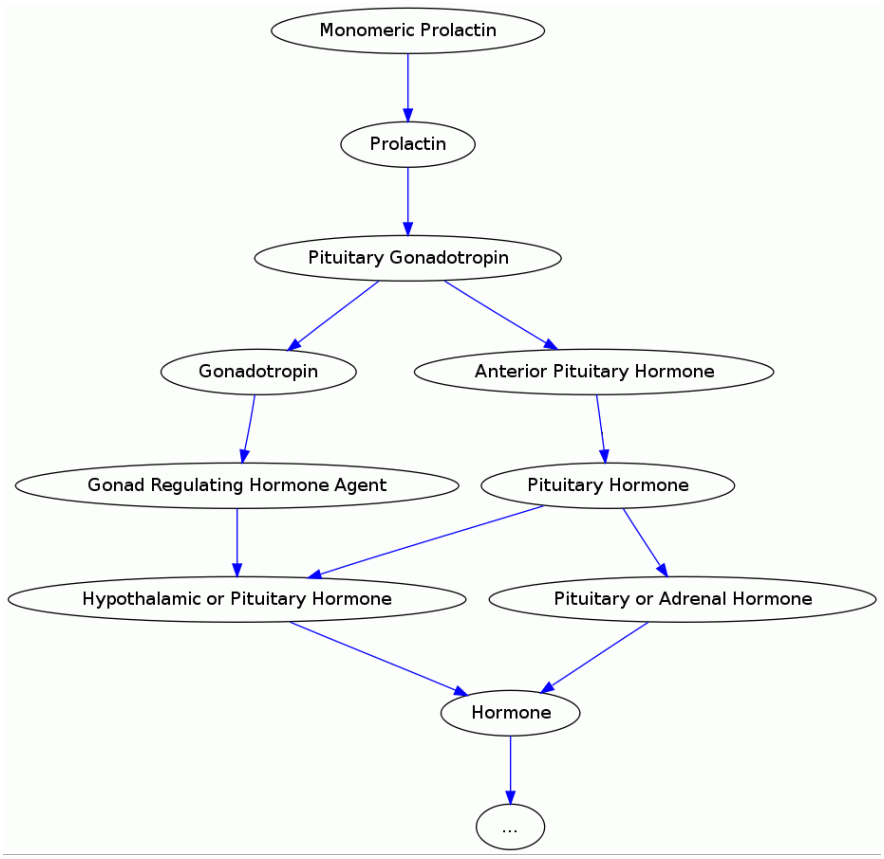

Figure 36: Snomed taxonomy for "monomeric prolactin"

Domain	Precision in classification (%)	Average precision per trial (%)	Recall (%)
hormones	47.61	48.82	19.33
viruses	100	87.26	34.6
fishes	100	98.21	35.94
disorders	75	72.91	11.79
total	81.33	76.24	26.89

Table 26: Overall precision and recall figures in a sample of 82 terms

Table 26 shows the figures of precision and recall for each seed term of the initial sample and their average in the last row. As stated above, the most important is the first column, because it is the one that classifies a given term as a determined type of entity. That is, the most generic hypernym. The second column accounts for the internal consistency of the

generated taxonomy. Its importance is given by the fact that, besides classifying a term assigning a hypernym to it, the taxonomy also provides internal hypernymy assignments that should also be correct. Table 26 shows generally good performance, with the exception of the group of seed terms pertaining to the class of hormones. The reasons for this are known mainly to be the interference produced by the seed terms pertaining to the field of diseases, a highly associated field (see 7.1.4). Figures of recall, on the other hand, are considerably lower. As already explained, this means that an important share of hypernyms that are present in the ontology is missing in the automatically generated taxonomies. Slightly higher recall is found in those subdomains where there are fewer hypernyms per term in the ontology, as in the cases of *viruses* and *fishes*. Low recall, in an experiment such as this one, is not really the problem if the goal is to construct a single taxonomy, given that limited pieces of information collected per each seed term can make up for the information missing in other trials.

The overall results that Table 26 shows are, at least with respect to figures of precision, quite remarkable in comparison to the results reported in the assignment of hypernymy relations by lexico-syntactic patterns. Consider, for instance, the 39% precision achieved by Davidson (1997), the 36% achieved by Snow et al. (2006), the 40% by Feliu et al. (2006) or the 62% reported by Potrich & Pianta (2008).

7.1.4 Discussion of the results of this Experiment

This experiment was intended to prove the reliability of a methodology for automatic taxonomy construction from unstructured large scale document collections (such as encountered, e.g., in the web) based on the first hypothesis presented in this thesis. This methodology does not use any knowledge of a given language or domain, relying only on statistics of lexical co-occurrence. The algorithm has no information about any of these classes, about their names or the number of them, and only by distributional analysis it is expected to return, for any input term, its correct category. As already said, it would be unrealistic to expect a complex multiple inheritance taxonomy of the quality of SNOMED with a small scale experiment like this one. Nevertheless, this experiment has a evident practical engineering interest for the problem of taxonomy acquisition from corpora. It is to be expected that with a higher number of seed words and documents, more complex and accurate taxonomies could be built. This practical outlook, however, should not obstruct a deeper outlook on a study that is relevant to theoretical linguistics as well: the study of distributional semantics of lexis in the context of text corpora. This experiment shows that the conceptual space of a term can be captured by terms that occur in its neighborhood.

With respect to the experiments on first order co-occurrence, the results presented in the evaluation of Stage 1 are informative because they serve as an estimation of the probability that a specialized term is syntagmatically related (i.e., co-occurs) with some of its closest hypernyms in text corpora. There is an evident relation between the figures of co-occurrence of the hypernyms and the ratio of hypernyms per term in the three categories of the SNOMED-database. As detailed in Table 23, the highest figures of co-occurrence appear in the area of disorders where there is 81% partial matching between the hypernyms of the ontology and the co-occurrence terms. One of the explanations for this higher match is that, in the categories if disorders, the ontology provides a higher proportion of hypernyms per term, and, therefore, is more likely to hit a correct hypernym. The difference in the term-hypernym ratio is rooted in the nature of the analyzed entities, which is also an interesting fact in itself.

As already stated in 7.1.3, the assessment of the results of the first order co-occurrence set of experiments included the implementation of an automatic evaluation procedure. The advantage of such automation is, on the one hand, that it provides an objective measure of performance and, on the other hand, that it allows for a cost effective large scale evaluation, which is more reliable than the conclusions that can be drawn from a few trials. However, it also implies several disadvantages. For instance, often (particularly in WordNet) the hypernym proposed by the algorithm was not the same as the one found in the ontology but was still a valid hypernym for the term in question. A typical case is that of synonyms. Thus, for the algorithm, a given term can be a *disease* while the ontology states that it is a *sickness*. The second stage of the experiments, the experiments on second order co-occurrence, shows some degree of amelioration of the results in comparison to the first order co-occurrence experiment conducted on the same sample. The number of correctly retrieved hypernyms is augmented and the average ranking of these is also reduced, which means that a better confidence can be achieved with respect to the hypernym candidates.

With respect to the evaluation of the results obtained in each stage of the process, the comparison of Stages 1 and 2 demonstrates that there is an improvement in the percentage of the collected hypernyms as well as in the average ranking where those hypernyms were found.

With respect to the third stage of the experiments, with the final generation of taxonomies for each of the seed terms, the figures of

precision shown in section 7.1.3 are much better than those of other authors (see last paragraphs of previous section), in spite of the noise and silence introduced in the first stages of the experiments. This is explained by the design of the algorithm: it is robust because it reinforces the certainty of the hypernymy assignments through the number of trials triggered by the set of seed terms. These results would not have been possible with only one or with very few seed terms. There are, though, important differences between domains. Most of the problems are found in the hormones domain, and the reason in most cases is the interference of the terms related to diseases. Hormones are often related to the cause or to the treatment of diseases, and it is likely that circumstances may have affected the results in that particular field. There are, though, other errors not attributable to the interference of disorders terminology. In some cases, this is because of the polysemy of some terms. The term *somantin*, for example, which is a hormone, happens to be also a song written by musician Trey Anastasio, and therefore reference to it abound in the web. Even when the information yielded is in fact correct, it is not relevant in this case and the trial is therefore considered a failure. As already commented, cases of polysemy such as these would be less frequent in a carefully selected corpus than in one directly downloaded from the web.

The figures of recall are, on the other hand, considerably low, which means that many of the SNOMED hypernyms are missing in the automatically generated taxonomies. This might suggest that one should not expect to find, for a given term, the whole chain of hypernym terms co-occurring frequently with the seed term in documents downloaded from the web. For any given term, there must be some natural level of description, which does not necessarily include hypernyms for that term if they are too far in the hypernymy chain. In other words, there are hypernyms that are more relevant than others to the description of the concept denoted by a given term (and therefore, co-occur more frequently with it).

7.1.5 Future Work for this Experiment

One of the possibilities of future work for this experiment would be to apply this technique of taxonomy extraction to the automatic compilation of bilingual terminology. Taxonomic structures offer valuable information for the decision of which is the correct translation of a term from a list of translation candidates. No such attempt has been made in this thesis, although it includes an experiment on bilingual terminology extraction (Section 7.4). In this line of research, it would also be possible to conduct an experiment to extend the technique to the automatic extraction of synonyms. A technique for the extraction of synonyms would select as synonym candidates units that have a high overlap in their respective

graphs. Experiments in this line of research, as well as the extraction of dialectical differences in the denomination of a concept in different variants of the same language are attractive subjects of study.

It would be also interesting to combine the knowledge obtained with this algorithm with other sources of information using some degree of linguistic processing. This linguistic processing could be done with the lexico-syntactic patterns approach as the one reported in Chapter 5, combining the quantitative strategy sketched in this thesis with a language-specific strategy based on lexical patterns. This would take the research in the direction of a hybrid strategy. The hybrid strategy would work as follows: with the quantitative strategy it is possible to select a list of hypernym candidates; then, with the symbolic strategy, the weight of a given term-hypernym candidate pair from the obtained list would be reinforced if the term and the hypernym candidate were found in the corpus occurring in any of the predefined lexico-syntactic patterns.

7.2 Second Experiment: Distinction between Sense and Reference

The present section includes a set of experiments which gather empirical evidence in support of the second hypothesis, according to which it is possible to distinguish referential expressions on a distributional basis. The experiment consists, thus, in the automatic detection of referential expressions from a given sample of text. It is possible, however, to replicate the experiment without providing the sample of text. In this case, instead of extracting units from a text, what is done is to provide a unit or a set of units and to obtain, for each unit, a value representing the probability of it being a referential unit. As it has been explained in Chapter 6, the algorithm that executes this analysis uses the web as a corpus to obtain information about the units subject to analysis. The reason for proceeding in this way is that there is a chance that the analyzed unit will have different senses in different contexts. Therefore, providing a sample of text where the unit occurs (with a specific sense) benefits the quality of the result, since the contexts of occurrence contain information that is useful to disambiguate polysemous terms.

The experiment presented here is to distinguish among sense and reference and addresses two different kinds of tests. The first one is from a synchronic point of view and the second from a diachronic one. One of the ways to undertake the experiment in the synchronic perspective is to inspect the profile of the co-occurrence of a given expression. That is, to

observe if in a very general corpus (such as the web) there is a determined set of words that tend to appear at short distances from the target expression forming dense graphs. From the diachronic perspective, in contrast, the study shows how the distribution of the units evolves over time. Referential units tend to be highlighted by a distinctive pattern: they have a life span, in contrast to the rest of the units of the vocabulary which are in permanent use with independence of the context.

7.2.1 Differences from the Synchronic Point of View

7.2.1.1 Experimental Set-up

The idea of the experiment is to rank expressions that occur in a given text according to a score that indicates their probability of having a referential function in that particular context. Each of these expressions, henceforth, each referential unit candidate, is defined as a set $T = \{t_1, t_2, \ldots, t_n\}$. Each $t \in T$ is scored on the basis of several geometric properties of the graph that is generated by taking t as input for the process outlined in Chapter 6. Such properties of the graph are measured by the following coefficients: 1) Density coefficient; 2) Frequency ratio coefficient; 3) Overlap coefficient; 4) Frequency of t in the analyzed text, 5) Weighted Frequency and 6) Indirect Overlap.

Ranking of referential expression candidates

$T \leftarrow$ {list of units to classify}
$d \leftarrow 100$ // initialize number of documents to download
$G \leftarrow \varnothing$ // initialize matrix to store graph
$M \leftarrow \varnothing$ // initialize matrix to store scores of candidates
for each $t \in T$ do
  $G(t) \leftarrow$Co-occurrenceGraph(t, d)
  M(Cf-1, t) $\leftarrow$Density($G(t)$)
  M(Cf-2, t) $\leftarrow$FrequencyRatio($G(t)$)
  M(Cf-3, t) $\leftarrow$Overlap($G(t)$)
  M(Cf-4, t) $\leftarrow$Frequency($G(t)$)
  M(Cf-5, t) $\leftarrow$WeightedFrequency($G(t)$)
  M(Cf-6, t) $\leftarrow$IndirectOverlap($G(t)$)
endfor
FinalRank(M)

Figure 37: Pseudo-code for the extraction of referential units

In order to rank expressions of a given input list according to their probability of having a referential value, it is necessary to combine these six coefficients into a final weight. The decision upon a discrete vs. continuous gradation is not trivial because in the first case one can classify a given unit while, in the other case, one is confronted with a scenario where there are some units which are clearly referential, others which are not and then a boundary region where it is not clear whether a unit belongs to one class or another. The decision in favor of a continuous gradation is thus not merely methodological. It is motivated by the fact that in this case the continuous gradation is assumed to be an intrinsic property of the object of study.

The experimental set-up now offers the details of each of the coefficients and the manner in which they are combined. The sequential steps undertaken in this experiment are shown in Figure 37. In the experiments presented in this section, the value of *d*, which is the number of documents representing the sample downloaded from the web for each analyzed unit, is set to 100. This does not guarantee, of course, that the search engine will return that number of documents (there may be fewer). In what follows, details of the coefficients mentioned in Figure 37 are given.

Cf-1) Density coefficient: The density coefficient (Formula 15) of a graph is the logarithm of the number of secondary connections over the number of contexts where *t* occurs.

$$\text{Cf-1}(t) = \log\left(\frac{A(t)}{C(t)}\right) \qquad (15)$$

$A(t)$ = number of secondary connections in the graph of a given candidate *t*.
$C(t)$ = number of contexts where the candidate *t* occurs

Graphs produced using referential expressions as query show a tendency to select a determined lexical cohort and are more dense. Graphs produced with non-referential expressions, in contrast, show rather poorly interconnected graphs. Graphs from non-referential expressions do not show these dense graphs because the documents where they occur are not related with respect to the content. For instance, a word like *requisite* will not generate a dense graph because it appears in such a variety of contexts that there is no selected set of vocabulary units that are statistically related to it. A proper noun like *Boudou*, on the contrary, will generate a dense graph because that is currently the name of Minister of the Economy of a country (and, thus, it is a "cultural unit", and entity of shared knowledge). As a consequence, the graph will be populated

with expressions related to this person and his affairs (such as *Argentina, Debt, Economy, IMF*, etc).

It is important to recall the difference between primary and secondary connections stated in Section 6.1: A primary connection is the one that holds between the candidate term and any other node of the graph generated for it, while the secondary connection is the one that holds between two nodes of the same graph, neither of which is the candidate term. Thus, for a given number of contexts of *t* the Frequency ratio coefficient of *t* (henceforth Cf-1(*t*) to abbreviate) will increase as the number of secondary connections increase.

Cf-2) Frequency ratio coefficient: Under the assumption that graphs of referential expressions are dense and have strong connections, one can also examine the strength of these connections. The Frequency ratio coefficient for a graph is the ratio between the sum of the relative frequencies of the units occurring in the downloaded sample (the corpus from which the graph is drawn) and the relative frequency of the candidate *t* in that sample. *Fr*, in Formula 16, represents the subset of weighted vocabulary of the downloaded collection, ordered by decreasing frequency, such that Fr_1 is the most frequent unit in the downloaded collection (excluding *t*). It is to be expected that, in the case of non-referential expressions, there will be a greater difference between the frequency of the query term and the sum of the frequencies of the *n* most frequent terms in the sample, because non-referential expressions tend to be independent of the content. Using again the above example, there will be a bigger difference in frequency, in the case of a non referential expression such as *requisite,* between such unit and the rest of the most frequent units in the contexts of *requisite* than in the case of the name *Boudou*, where there will be other frequent units as well, such as, again, *Argentina, Economy, IMF,* etc. In a sense, this and the previous coefficient are measuring the same property but from different perspectives. The value of *n*, in the case of this experiment, is limited to 1 (i.e., the most frequent co-occurrent unit) for reasons of efficiency[35].

[35] Setting this parameter to 1 accelerates the process but at the risk of error because it could be the case that there is only one important node in the graph of *t* and in that case the coefficient would wrongly result in a high score. This should be subjected to future experiment and optimization.

$$\text{Cf-2}(t) = \frac{\sum_{(j=0)}^{n} Fr_j}{Fr_t}$$

Fr_j = frequency of unit j in the downloaded sample.
Fr_i = frequency of candidate t in the downloaded sample.
$j \neq t$

(16)

Cf-3) Lexical Overlap: The measure of lexical overlap is intended to correct problems derived from the potential ambiguity of *t*. Reference can change depending on the context. For instance, the term *string* can have different referents in Computer Science and in String Theory. Thus, the purpose of this coefficient is to ensure that the graph of *t* represents the terminology related to the same sense of *t* as it is used in the document under analysis. As shown in Formula 17, the lexical overlap intervenes thus as a disambiguation factor for terms because it measures the overlap of the most informative lexicon in the analyzed document and in the downloaded sample.

$$\text{Cf-3}(D, G(t)) = \frac{|D \cap G(}{|D|}$$

D = weighted vocabulary from the analyzed document.
$G(t)$ = weighted vocabulary of the downloaded collection.

(17)

Cf-4) Frequency in the Analyzed Document: this coefficient registers the frequency of *t* in the analyzed document. Frequency alone is not, of course, a sufficient source of information to discriminate referential units. However, as it has been empirically demonstrated, its contribution to the final ranking improves results.

Cf-5) Weighted Frequency: this coefficient complements the previous one by penalizing the units that have a low informational value, according to the same language model that was described in Section 6.3., and used in the first of the experiments of this chapter. As explained, the rationale of this coefficient is to reward elements which are frequent in the analyzed document and at the same time infrequent in the reference corpus. Those units which are frequent in the analyzed document but are also very common words of the vocabulary of a language receive zero weight (according to a threshold k which in this case is 1000).

$$\text{Cf-5} = \begin{cases} \log\left(\frac{1+\text{fd}(t)}{1+\text{Rc}(t)}\right) & (\text{Rc}(t) < k) \\ 0 & \text{otherwise} \end{cases}$$

fd*(t)* = frequency of t_i in the analyzed document.
Rc*(t)* = frequency of t_i in the language model.
k = frequency threshold (1000)

(18)

In the case of multiword expressions, this estimation is done by averaging the weight of the individual components of the expression, as explained in Section 6.3. Note, however, that if the first or the last component of a multiword expression has a frequency in the model higher than k, then the unit as a whole is considered uninformative and receives zero frequency.

Cf-6) Indirect Overlap: reflects the number of times the candidate t has appeared in the graphs generated for other candidates in T. This coefficient is, in a way, close to the Overlap coefficient, however posed from another perspective, and it is meant to correct the Overlap coefficient. As stated above, the phenomenon that the Overlap coefficient is intended to capture is the lexical overlap between the graph generated for a given unit and the vocabulary of the analyzed document. However, there are cases where such an overlap does not exist, even when there is an evident semantic relation of the concept denoted by the candidate and the content of the document. For instance, in a text about terminology one will probably find vocabulary units such as *word* and *term*. In such a case, they are referential expressions because in the context of that document such vocabulary units refer to the concept of *word* and *term*. Thus, while there would be other referential expressions that would show a high overlap, like a proper noun such as the name of the terminologist *Eugen Wüster*, in the case of the co-occurrence graphs of the units *words* and *terms*, there will be a poor or non-existent overlap with the vocabulary of the analyzed document, for the simple reason that these units are extremely general and polysemous. The Indirect Overlap (Formula 19), in the case of such units (*word* and *term*) will be high because they will be present in many of the graphs generated for the rest of the candidates in T.

$$\text{Cf-6}(t) = \sum_{j=1}^{|T|} k \begin{cases} k=1 & (t_j \in G(t)) \\ k=0 & \text{otherwise} \end{cases}$$

T = referential unit candidates
G(t) = weighted vocabulary of the downloaded collection for t.

(19)

Final Ranking: The final ranking is given by the mean of the rankings of each coefficient. When two competing units have the same value, they are both positioned in the same rank. The mean is calculated using the positions that a given unit has in the ranking proposed by each coefficient. In this way, the first position in the rank receives the value 1 and the last position the value 0.

Few additional factors are used to give more importance to some coefficients over others. The first three, Density, Frequency Ratio and Overlap, are multiplied by two because they are considered more important. In addition, a coefficient is multiplied by two if it surpasses an upper value (m) or if it does not surpass a lower value (n). The motivation for this procedure is to increase the weight of a coefficient to reward or penalize a unit if it is too low or too high in the ranking. This is intended to ensure that a coefficient corresponds with a high certainty about the score of a candidate. This judgment is considered more relevant than the case of another coefficient that rates the candidate in a boundary region with a mediocre scoring. Obviously, the fact that a coefficient is multiplied by two implies that the denominator to calculate the mean is increased by 1, in order to maintain the norm.

The final rank is, thus, given by the mean of the rankings of each individual coefficient. For a unit t to be ranked as a referential expression candidate, it is defined as a set F_t (20) as values of the rankings provided by each of the six coefficients. Thus, the final rank of t or FR(t) is simply the mean of F_t.

$$\mathrm{FR}(t) = \bar{F}_t \qquad F_t = \{\mathrm{Cf}_1(t), \ldots \mathrm{Cf}_6(t)\} \qquad (20)$$

7.2.1.2 Results

This section shows the result for this experiment after submitting for classification an arbitrary selection of 20 units from a newspaper article[36]. The resulting classification of the units is shown as a ranking in Table 27. Shadowed cells are the units that were considered referential by manual examination of the text before the experiment.

[36]The newspaper article reference is: "Sería inminente el lanzamiento del nuevo canje de la deuda". Martín Kanenguiser - Diario La Nación. http_//www.lanacion.com.ar/nota.asp?nota_id=1183021 [accessed October 2009]

In the case of this set of units, the distinction offers no difficulties because those which were selected as referential expressions by the native speaker are also proper nouns. The columns of the table includes the values obtained for each coefficient described in 7.2.1., expressed as positions in their respective rankings, as explained in the paragraph “Final Ranking”. Table 27 also includes the number of hits that each expression found in the search engine (restricted to Spanish results) however that information does not contribute to the final weight. The last two columns are the percentages of precision and recall taking a native speaker's criterion as reference. In the matrix, values marked with the sign (+) and in bold font face are those which surpassed the upper value threshold. In a similar way, values marked with the sign (-) and in italics are those which are below the lower threshold and are penalized accordingly.

rank	candidate	hits	coefficients						mean
			1	2	3	4	5	6	
1	club de parís	41260	(+) **0.90**	0.80	(+) **0.85**	0.75	0.80	(+) **0.85**	0.84
2	boudou	154363	0.80	0.65	(+) **0.85**	(+) **0.90**	(+) **0.95**	(+) **0.85**	0.83
3	fondo monetario internacional	570881	(+) **0.95**	(+) **0.95**	0.80	0.70	0.75	(-) *0.27*	0.77
4	redrado	32091	0.75	(+) **0.85**	(+) **0.90**	0.80	(+) **0.85**	(-) *0.27*	0.75
5	martín kanenguiser	1192	(+) **0.85**	(+) **0.90**	0.75	0.70	0.75	(-) *0.27*	0.73
6	dominique strauss kahn	588611	0.70	0.75	(+) **0.95**	0.70	0.75	(-) *0.27*	0.70
7	argentina	73836859	(-) *0.10*	0.40	(+) **0.85**	(+) **0.95**	(+) **0.95**	(+) **0.95**	0.67
8	fmi	3104935	(-) *0.10*	0.50	0.70	(+) **0.85**	(+) **0.90**	(+) **0.90**	0.62
9	washington	108347238	0.55	0.60	0.75	0.70	0.75	(-) *0.27*	0.58
10	narices	374013	0.60	0.30	0.60	0.70	0.75	(-) *0.27*	0.50
11	aclaró	790100	0.40	0.55	0.50	0.70	0.75	(-) *0.27*	0.49
12	críticas	6095885	0.45	(-) *0.25*	0.70	0.70	0.75	(-) *0.27*	0.46
13	términos	16155999	(-) *0.10*	0.70	0.60	0.70	0.75	(-) *0.27*	0.44
14	cuestión	4513887	(-) *0.10*	0.45	0.65	0.75	0.80	(-) *0.27*	0.42
15	negociación	1909133	0.50	(-) *0.05*	0.60	0.75	0.75	(-) *0.27*	0.40
16	requisito	1347087	(-) *0.10*	0.35	0.60	0.70	0.75	(-) *0.27*	0.38
17	moody	4907712	(-) *0.10*	(-) *0.20*	0.75	0.70	0.75	(-) *0.27*	0.37
18	instalación	4088996	0.35	(-) *0.10*	0.55	0.70	0.75	(-) *0.27*	0.37
19	grupo de periodistas	76160	0.65	(-) *0.00*	0.65	0.70	(-) *0.23*	(-) *0.27*	0.36
20	mucha plata	27055	(-) *0.10*	(-) *0.15*	0.70	0.70	0.75	(-) *0.27*	0.34

Table 27: Ranking of a sample of units of a text according to their referential value

The ranking presented in Table 27 shows a high coincidence with the human judgment. The coefficients derived from co-occurrence graphs accurately assign the highest punctuations to those expressions that have a specific reference. The only case that escaped the prediction has been the unit *Moody,* which refers to Moody's Corporation. This candidate has a low score in both of the coefficients that measure vocabulary overlap as well in the one that measures Frequency ratio. In the case of the overlap, examination of the data revealed that it was motivated by the fact that most documents downloaded with the query expression *Moody* are in English. As it will be commented more extensively in the conclusions of this experiment, the search engine's language selector is not reliable enough. This problem could be ameliorated by an automatic language selection system or by expanding the query to retrieve only documents that are in the same language as the analyzed text. With respect to the low score on the Frequency ratio coefficient, this is more difficult to explain, and manual examination of the data is not helpful. By definition, a low Frequency ratio score indicates that the element in question has weak ties with the vocabulary that it co-occurs with. One could consider, as a tentative explanation, that being an investment services corporation, it is related to a great variety of actants. However, this does not explain why another more general unit such as *Washington* scores better in the same scale. What this data suggests is that *Washington* has stronger ties with other units than in the case of *Moody's*, which is, at least, counter-intuitive.

7.2.1.3 Evaluation

The evaluation of this experiment was conducted using a more difficult case than the one shown in Table 27, where most of the referential expressions are easily recognizable as being proper nouns. In this second experiment, a new level of difficulty is added because among the units to classify there are proper nouns but also terms, and the terms do not have the prototypical structure of terminology of domains like medicine, which show characteristic morphological patterns and Greek and Latin formants. This time, the analyzed text is from the domain of linguistics[37], and some of those terms coincide in appearance with words that in other contexts are not referential, as will be seen in more detail below.

The procedure used to evaluate was, as in the previous result, to manually classify a set of 28 arbitrarily selected single and multiword expressions from the analyzed document, selecting those that are considered

[37]The text that was used is: "Términos y palabras en los diccionarios", from M. T. Cabré, (2007). In: Cuartero Otal, J.; Emsel, M. (ed.). Vernetzungen: Bedeutung in Wort, Satz und Text. Festschrift für Gerd Wotjak zum 65. Geburtstag. Frankfurt am Main: Peter Lang, pp. 71 - 84.

referential. Afterwards, the same sample of units was submitted to classification by the tested algorithm, along with the analyzed document. The output is the ranking of the 28 units, shown in Table 28, and it demonstrates that there is a strong similarity to the human criterion, given that all except 3 of the units manually tagged as referential are in the upper half of the ranking.

Rank	Candidate	Human	Baseline	Algorithm
1	eugen wüster	1	1	0.71
2	teresa cabré	1	1	0.63
3	lexicografía	1	1	0.63
4	términos	1	1	0.63
5	terminografía	1	1	0.60
6	cahiers de linguistique sociale	1	1	0.50
7	palabras	1	1	0.50
8	valor especializado	1	0	0.48
9	teoría comunicativa de la terminología	1	1	0.47
10	cambio en el proceso general	0	0	0.46
11	campo de conocimiento	1	1	0.45
12	sentido preciso	0	0	0.44
13	françois gaudin	1	1	0.44
14	disposición de las informaciones	0	0	0.43
15	esquema conceptual	1	0	0.42
16	perfiles de usuario	0	0	0.34
17	producción de nuevo conocimiento	0	0	0.34
18	concepción restrictiva	0	1	0.33
19	método semasiológico	1	1	0.32
20	dos listas de unidades	0	0	0.32
21	cuestión esencial	0	0	0.31
22	materia de conocimiento	0	0	0.31
23	necesidad de disponer	0	1	0.30
24	noticias sobre los avances técnicos	0	0	0.29
25	centro del campo	0	1	0.28
26	factores de separación	0	1	0.27
27	trabajo terminográfico	1	1	0.26

Rank	Candidate	Human	Baseline	Algorithm
28	herramientas de trabajo	0	0	0.21

Table 28: Ranking of units by the tested algorithm in comparison to binary classification made by human and baseline algorithm

A comparison with other authors' work is not possible because there is no precedent of an experiment of this nature. However, in order to provide some assessment of the quality of these results, a comparison with a baseline algorithm is provided. This baseline algorithm is a very simple rule-based system for the extraction of referring expressions. It works based on two main assumptions: 1) that every unit which always occurs in the document with initial upper case letters is a proper noun and, therefore, is referential and 2) that every noun phrase that occurs at least once immediately preceded by a determiner (definite articles, as in Spanish are *el*, *la*, *los*, *las*) is considered referential (Alcina, 1994). Based on these two assumptions, the baseline can provide the binary classification shown in the *Baseline* column of Table 28 and, in comparison with the human criterion, it is remarkably good despite its simplicity. However, a third assumption is added to the baseline in order to transform the binary classification of the baseline into a ranking of units. This is necessary to achieve a better comparison with the tested algorithms in terms of precision and recall. In the ranking of units by the baseline, units with value 1 precede units with value 0 and, within each group, the units are ordered in decreasing frequency, assuming that frequency in the analyzed document is correlated with referential function. In this way, it is possible to compare both rankings from the tested and baseline algorithm with the manual classification of the units, and to then plot a graph of precision and recall (Figure 38).

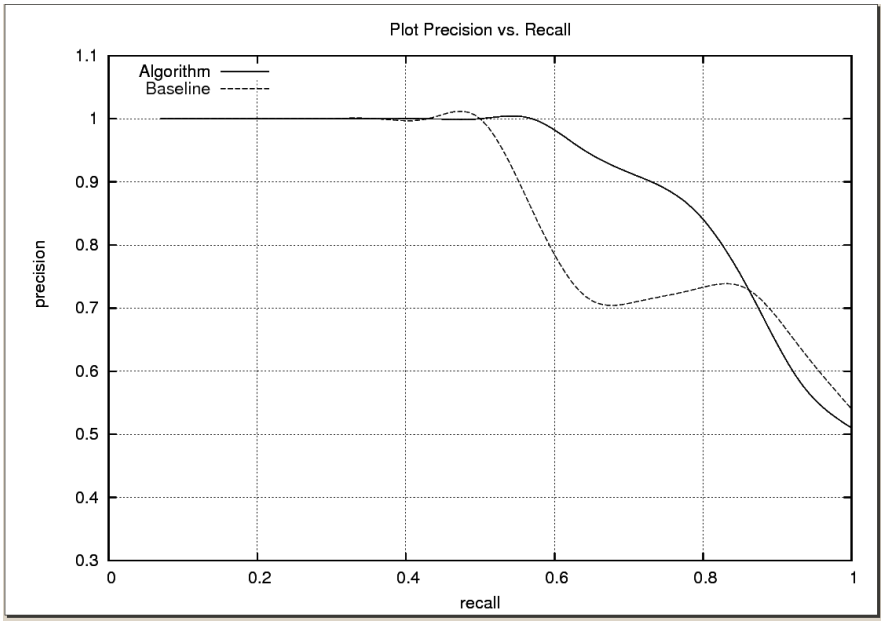

Figure 38: Comparison of Precision and Recall of the tested algorithm against the baseline algorithm

The comparison of the two algorithms reveals that both are quite efficient in the classification of the units in the first positions of the rank. However, the tested algorithm demonstrates better performance in the lower positions. While the baseline shows a sharp decrease in precision at around 50% recall, the tested algorithm is still at over 80% precision after 80% recall, which is an appreciable difference.

7.2.1.4 Discussion of the results of this Experiment

The results of this experiment have shown that there exists a measurable statistical difference between the co-occurrence graphs generated from referential expressions in comparison to those graphs generated from units which do not have referential function, with the graphs from referential expression being more dense, being highly interconnected, and with an appreciable lexical overlap with the vocabulary found in the specific context of occurrence of such expressions. The fact that such a difference can be measured allows for the possibility of developing a model which will have the power to predict -with reasonable accuracy, as seen in the previous section- when a given expression in a given document is referential. This is done, of course, without taking into account any

information about the domain or the language of the analyzed text, and under the assumption that the analyzed referential expressions pertain to a world of shared-knowledge and that, as such, are represented in some corpus (like the web).

With respect to error analysis, it is possible to observe that most of the errors in the classification of units are caused by the same factor: many times the referential units are independent of the language of the analyzed text and this affects the Overlap Coefficient. In the case of the results presented in Table 27, section 7.2.1.2, the name of Moody's Corporation, is not selected as referential because most of the words that appear in its graph are in English, and, since the analyzed document is written in Spanish, there is almost no lexical overlap and it receives, therefore, a poor scoring. The same happens with other units, like the name of François Gaudin in the experiment evaluated in section 7.2.1.3. This particular unit obtains high values by the Frequency ratio and Density coefficients, meaning that the graph it generates is dense, having many nodes and being well interconnected. However, and for the same reason, it shows poor overlap since most nodes of its graph are in English or French. In principle, this problem should be corrected by a proper selection of the language of the web. The problem here is that the language selector of the search engine that was used (Yahoo) is not reliable enough, even though the "language=strict" parameter of the Yahoo interface was active during the search. That means that even when one asks for documents in Spanish, many documents in English are returned. Previous research (Nazar et al., 2008) has demonstrated that to automatically recognize the language of a text is a relatively easy task. Thus to include such recognition and to exclude documents from the downloaded collection that are not from the same language of the analyzed text would be one of the ways to improve the quality of the results.

Improvements of this nature would be included in a future stage dedicated to optimize the results. At this point, however, the idea is only to offer empirical evidence in favor of a hypothesis. Even in these circumstances, there are enough reasons to believe that this proposal is also interesting from a practical point of view in fields such as terminology extraction, because of the relative simplicity of the strategy and the robustness and versatility of a statistical, language and domain independent approach.

The present experiment has similarities with terminology extraction. The field of automatic extraction of terminology was already introduced in section 5.2. Some would possibly also relate this kind of experiment to the field of *named entity recognition* (Boufaden et al., 2004, among others). In the latter case, there are plenty of strategies for recognizing named entities that use surface features such as the presence of initial upper case

letters in the case of proper nouns or by detecting "trigger words", as in the work of Gaizauskass & Willks (1998), already commented in Section 5.2.1. The experiments presented in this section, however, offer a quite different approach. The methodology proposed here takes advantage of structural properties of linguistic expressions that designate entities in discourse, units which are not necessarily written with upper case letters. Despite the differences in methodology, however, it is claimed that the methodology and perspective explained in this section can be used, with some adjustments, in such areas as terminology extraction and named entity recognition.

If this strategy is applied as a term extractor, some shallow strategies to distinguish and eliminate proper nouns may be implemented, as well as some degree of parsing for the correct segmentation of multiword expressions. In the case of terminology extraction, often the user will not be interested in proper nouns, and it will be the other way around when a user is interested in named entities. Thus, in both scenarios, the idea would be to determine which subset of the units classified as referential expressions are terminological units and which are named entities. In that case, information such as the already mentioned morpho-syntactic patterns as well as the probability of a unit appearing with an initial upper case letter[38] should be taken into account.

7.2.2 Differences from the Diachronic Point of View

This section presents a different attempt to extract referential units from text this time from a diachronic point of view. Even when the experimental design does not use co-occurrence graphs, it is included here because it allows us to study the same phenomenon from a complementary perspective.

7.2.2.1 Experimental Set-up

If, as it is hypothesized in Chapter 2, referential units can be identified by a function of their contexts, then it is to be expected that the units which have a homogeneous frequency distribution across the time line are the ones that are independent of the context and are thus not referential. Referential expressions, on the contrary, will describe the opposite behavior: their frequency will be concentrated in certain time intervals.

It is not possible, though, to make such a claim in a categorical manner. There are referential units -*Jesus Christ*, for instance- that have a continuous presence over time. However, it can be stated that it is a

[38]This is, naturally, a language dependent feature.

general tendency of referential units to show greater variance over time because they are more restricted to the topic of discourse. The topic, in turn, depends on the agenda setting of the media for every particular period of time. Two heterogeneity measures are used in this experiment. One is *tf.idf* (Spark Jones, 1972) and the other, the same dispersion coefficient that was introduced in Section 6.3.

The corpus used for this experiment consists of the archives of El PAÍS[39] newspaper in the time interval of years [1976-2007]. Given the characteristics of this corpus, instead of being partitioned into documents, its parts represent years, including all the text edited in a year. For illustration, consider Figures 39 and 40. The first one describes the relative frequency of two units of the core vocabulary of the Spanish language across the time line from 1976 to 2007. These units are *después* (after) and *entonces* (then). The curves of these two units are practically horizontal because their use is constant (because it does not depend on the context). They both refer to relative positions in time which are of great generality. In the case of Figure 40, in contrast, *Alzheimer*, in reference to the disease, shows a discontinuous use which is characteristic of referential expressions.

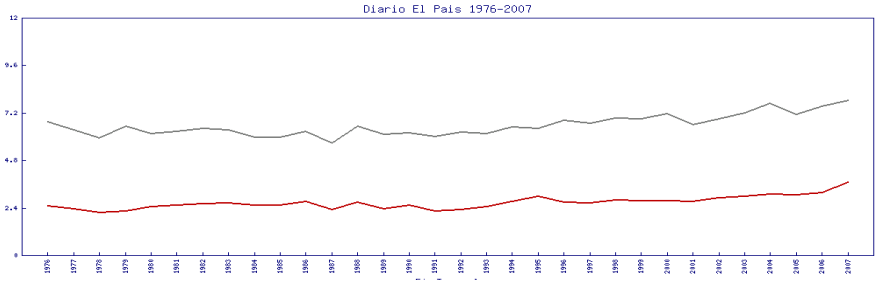

Figure 39: Frequency distribution of "después" (after) and "entonces" (then)

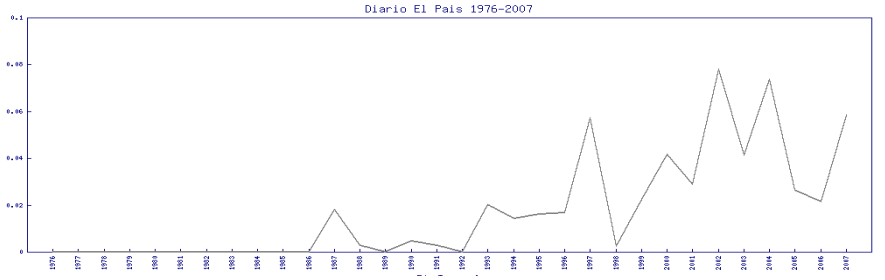

Figure 40: Frequency distribution "Alzheimer"

Measuring the heterogeneity of the frequency distribution, it is possible to estimate the probability that a unit is referential. The first measurement was made, as it was already mentioned, with *tf.idf*, (Formula 21) which is

[39] http://www.elpais.com [downloaded in December 2007].

an estimate of the probability for a term to appear *k* times in a document. It is therefore a measurement of the amount of information that the term has.

$$w(i,j)=(1+\log(tf_{i,j}))\log\frac{n}{df}$$

$tf_{i,j}$ = frequency of the term *i* in a document *j*.
df_i = number of documents where the term *i* occurs. (21)

The second measure of dispersion (Formula 22) supersedes tf.idf when applied to the task of identifying referential terms. The referential coefficient for a unit is equal to the quantity of subparts of a corpus under a frequency threshold indicated as parameter times the highest relative frequency in any subpart of the corpus.

$$D(t) = \mathrm{Mf}(t)\ .\ \mathrm{Cr}(t)$$

t = term under analysis.
Mf(*t*) = maximum frequency of t in a partition.
Cr(*t*) = number of partitions with a frequency of $t \leq k$.(with *k* being an arbitrary parameter) (22)

7.2.2.2 Results

This section shows the result of this experiment after submitting for classification a random sample of units taken from the set of units with the same frequency in the year 2007, the last year of the collection. These units appear listed both in columns 1 and 3 of Table 29. The first list in the ranking is the result from the weighting of *tf.idf*, while the second is the ranking output by Formula 22. The units that are considered referential are shadowed in the table. The right hand side rank is better because most of the shadowed units are near the top. In the list, there is a unit *fondos de alto,* which should be considered referential because it always occurs in this corpus as forming part of the larger expression *fondos de alto riesgo* (hedge funds), which is a very specific concept (a specialized term) that began circulating in the press in the last ten years. The reason why the last component is missing is that these *n*-grams were randomly selected. This means that a sequence like *fondos de alto riesgo* has the same probability of being selected as *fondos de alto* and in this case it was the latter that was selected.

Unit	IDF
lochte	0.0035196307
cercanía de las elecciones	0.0035264085
fondos de alto	0.0036700418
steve jobs	0.0038079582
algo diferente	0.0046240239
torre eiffel	0.0049145284
riviera maya	0.0055951460
puedo asegurar	0.0057730436
esos hechos	0.0058992065
dominaba	0.0059404898
gamarra	0.0060189185
doris lessing	0.0061264360
contestaron	0.0066184511
empresas europeas	0.0071734555
helga	0.0072380365
norman foster	0.0074954551
desborde	0.0078116418
invitando	0.0079066132
alambre	0.0087341321
hasta el último minuto	0.0091513478
mitigar	0.0101654941
himnos	0.0109961743
fugados	0.0117770547
contraposición	0.0122750024
titular de interior	0.0159323703
ultranza	0.0233978589
air france	0.0234926888
celebrando	0.0235067848
originado	0.0244623311
varios miles	0.0316631844

a) Terms sorted by *tf-idf*

Unit	Dispersion
riviera maya	0.76680
lochte	0.58530
fondos de alto	0.56579
steve jobs	0.50726
doris lessing	0.50547
cercanía de las elecciones	0.44873
gamarra	0.43586
helga	0.43350
titular de interior	0.39221
norman foster	0.39214
air france	0.29944
torre eiffel	0.27314
algo diferente	0.23412
alambre	0.16926
empresas europeas	0.16310
hasta el último minuto	0.16105
desborde	0.15882
esos hechos	0.15608
invitando	0.15355
fugados	0.14619
dominaba	0.04664
himnos	0.04446
mitigar	0.04320
puedo asegurar	0.02233
contraposición	0.00000
originado	0.00000
contestaron	0.00000
varios miles	0.00000
ultranza	0.00000
celebrando	0.00000

b) Terms sorted by $D(t)$ value.

Table 29: Comparison between IDF and this experiment's measure of dispersion.

7.2.2.3 Discussion of the results of this Experiment

In the upper part of column 3 of Table 29 there are the units that are more clearly referential, as technical terms or proper nouns, such as *Steve Jobs*, *Doris Lessing*, *Norman Foster*, etc., while in the lower part of column 3 there are the units which are clearly non-referential, including predicates and units that form part of the core vocabulary of the language, such as *celebrando* (celebrating), *puedo asegurar* (I can assure), *varios miles* (several thousands), etc. In the middle part of the table there are some units that are much more difficult to categorize as one class or the other. It

is the case of units such as *titular del interior* (Minister of Internal Affairs) o *empresas europeas* (European companies). A distinction based on intensional and extensional types of definition would not be useful in this case, because units such as these can function in both ways. The expression *titular del interior* can admit an extensional definition as a proper noun -or as a set of proper nouns- as well as it can admit an intensional definition as the type of charge or function performed by the Minister of Internal Affairs. To solve problems such as these one should analyze instance by instance instead of doing this general estimation.

It should be stressed, though, that even in the case of the analysis of a particular document, presented in the previous experiment as the synchronic perspective, the diachronic axis is still of great importance. The same document can have units of different referential status depending on the time in history when it is produced (or interpreted). The same unit, *climate change* in the same document is not referential in 1976 but it is in 2007. In addition, there are subjective factors that of course cannot enter into the analysis. The work of the reader in recognizing the referential units of a text depends on his or her knowledge of the world, as well as the clues given by the text. In the same use, sociological variables, such as age or social extraction and anthropological variables relative to the cultural conditions play a key role in the interpretation of a text.

7.3 Third Experiment: Analysis of Polysemous Terms

This section presents the results of a series of experiments undertaken on the disambiguation of expressions using graphs of lexical co-occurrence, based on the assumption that these are useful to characterize the contexts of occurrence of a given unit. The present experiments cast the problem of disambiguation as a problem of document clustering. That is, given a set of documents where a linguistic expression is used with multiple senses, an application of the algorithm based on Chapter 6 undertakes the double task of identifying the uses that an expression shows in a sample of contexts of occurrence and to cluster such contexts according to the different senses they show.

In order to accomplish this task, the system develops the co-occurrence graph described in Chapter 6 and then proceeds to travel through the generated graph identifying different clusters -or "attractors"- which are regions of the graph with a concentration of highly inter-connected nodes. Each node is linked to the contexts of occurrence of the terms in the documents. As a consequence, clustering a group of nodes from the graph

is the same as clustering a subset of documents from the collection, according to Yarowsky's (1992) one-sense-per-document assumption.

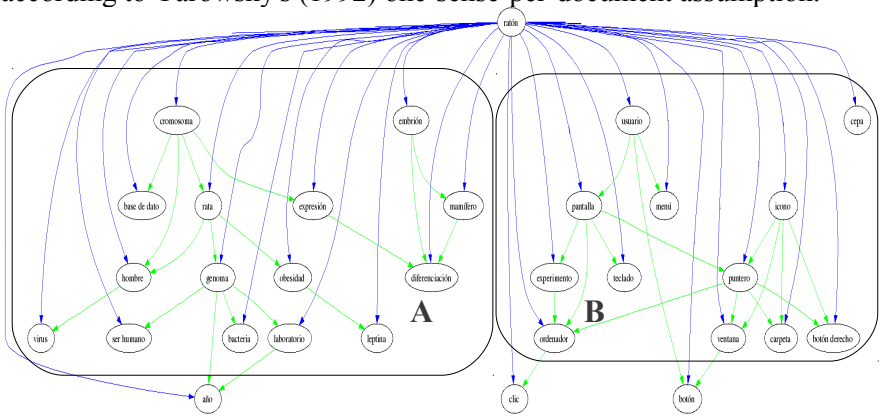

Figure 41: Two clusters of nodes in a graph generated by the polysemous Spanish term "ratón" (mouse).

For illustration, consider the graph depicted in Figure 41. This graph has been generated from IULA's Technical Corpus with the Spanish word *ratón* (mouse). This unit has two clearly distinguished uses in two of the domains of this corpus: Genomics and Computer Science. In the first domain (marked as cluster A in Figure 41), it is the name of the animal which is often subjected to experimentation in Genomics. In its use in the Computer Science domain, in contrast, it is the computer peripheral (marked as cluster B).

The experiments that will be presented in this section used the Web as a corpus. The language of the experiments is English and the units to be disambiguated are homonym acronyms. Acronyms have certain properties that make them especially suitable for an experiment of this nature. In first place, acronyms are convenient because it is easy to collect them in a systematic and straightforward way such as consulting the corresponding category in the Wikipedia or in other specific resources (see Giraldo Ortiz, 2008). Secondly, acronyms are appropriate for this experiment because they are used to designate specific entities, thus homonymy, in their case, means reference to different entities. A final and important feature of acronyms is that they are prone to be ambiguous out of context. Acronyms are an important resource of economy of language when an entity is several times referred to in a text, because they reduce the size of the linguistic sign. The size of a unit is inversely proportional to its frequency of occurrence and to its potential of polysemy, therefore very often the same acronym can refer to different entities.

Finally, and as a reference for comparison with other authors' work, the same experiment is replicated in section 7.3.4 using the same corpus that Véronis (2004) used in his experiment on disambiguation already introduced in section 5.1.2.2.

7.3.1 Experimental Set-up

Given a sample of homonym acronyms, the experimental procedure consists of obtaining contexts of occurrence of these acronyms from the web and to cluster these contexts in order to determine which are the different uses of each acronym in the sample. The list of acronyms is an arbitrary selection obtained from Wikipedia. This resource is convenient because it includes a specific category for acronyms with a large number of units, with the corresponding enumeration of uses in the case of the homonym acronyms. The selected units are listed in Table 30.

1. AASC	11. KPS	21. RLA
2. APCS	12.KSP	22. SEU
3. ASG	13.LEP	23. TDD
4. BVM	14. NAFTA	24. TLA
5. CKD	15. NCO	25. WTF
6. DDO	16. NPN	
7. ETN	17. OLA	
8. FYI	18. OOV	
9. IED	19. PCR	
10. JUB	20. PTW	

Table 30: Acronyms selected for the experiment.

The criterion for the selection of the units in Table 30 is that they have at least 3 different senses and no more than 15 senses registered in Wikipedia. Most of the three letter acronyms meet this constrain, except in the case of those which begin with the letter *A*, where there is more homonymy. This is why in that case some of the selected units had four letters (because, as already mentioned, longer units tend to be less polysemous).

The experiment consists thus in the application of the same algorithm described in Chapter 6, which loops through the list of units shown in Table 30 and searches each of them in Google, downloading the first 100 documents. This set of documents is submitted to a graph-based clustering process which automatically divides the set of documents into clusters. The minimum number of members per cluster is 3. As stated in Chapter 6, this program ignores the HTML code, the Javascript code, the hyperlinks and advertisement banners. It also ignores web sites that are not made of

text, like web sites composed of bitmap images of Flash movies. Web sites made with frames are also discarded[40].

Given a sample of acronyms T, a distinct trial for each acronym $t \in T$ is performed. A graph of the characteristics of Chapter 6 is generated and, within each graph, the steps listed in Figure 42 are performed.

Word Sense Disambiguation Algorithm

For each node $j \in G(t_i)$ do:
(starting from the node that has more secondary connections)
1: Follow each of the secondary relations with other nodes.
2: Create a new cluster and assign all documents containing the term that labels node j into a cluster (the same document can be assigned to more than one cluster).
3: Compare the newly created cluster with each of the clusters already created, starting in chronological order.
3.1 -If the comparison shows an overlap greater than a threshold k, collapse the new cluster with the existent one.
3.2 -Else if the new cluster shows the same degree of overlap to more than one cluster, destroy the cluster and return its documents to the document stack.
3.3 -Else if the overlap is not greater than k, create a new cluster with these documents.
4: Once there are no more documents that can be assigned to clusters, perform a pairwise comparison of all clusters. If two of them have an overlap of documents greater than k, then collapse them.
endfor

Figure 42: Pseudo-code for the disambiguation of polysemous terminology

The name of each cluster is automatically selected by the algorithm by choosing the most dispersed node in a cluster. The dispersion of a node is defined as the number of documents that contain the term that labels the node. It may or may not be the same as the expanded version of the acronyms. It should be stressed that, contrary to Experiment 1, in this chapter the result of the trial of each acronym in the sample is independent of the results of the other trials. This is because no information from the previous outcomes is learned for the next trial.

[40]A more careful process for the extraction of text from this type of documents is feasible. It is currently possible to extract text from bitmaps and Flash movies, and of course from frame-based websites as well. However, since this optimization is not essential to the experiment, it is left for future work in the case of a real application.

7.3.2 Results

Although the experiment was undertaken with the sample of 25 acronyms, this section shows the result of only one of them. The results of the rest of the experiments are presented in Appendix A, but they do not include as many details as the one presented here. The results of each experiment include information such as examples of the clusters that the program found, examples of clusters that the program rejects for not having the minimum number of members (three) and also some examples of documents that show different expanded forms of the terms the acronyms stand for.

==========
EXPERIMENT 1. Expression**:** AASC
==========

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
Asian American Studies Center	9	3	0	66	100
American Association of State Climatologists	4	1	0	75	100
Arizona Association of Student Councils	3	0	0	100	100

Clusters overlooked:
Asian American Student Association
Academic Assembly Steering Committee

Examples of clusters rejected for not having enough (3) members:

Clusters of two documents:
American Association of Service Coordinators
American Association of State Climatologists
Architectural Advisory Service Centre - Metal Finishing
Adaptive Avoid Second-collision backoff algorithm for multihop wireless sensor networks
Applied Aerospace Structures Corp
American Association of Service Coordinators

Isolated documents:
African-American Steering Committee
Active After-school Communities
Appalachian Agency for Senior Citizens
Active after school communities
Alamerda Applied Science Corporation
Advanced Agri-Solutions Co-Op
American Association of Safety Councils
Austrian Association for Statistical Computing
Alliance of AIDS Services Carolina
Astronomy and Astrophysics Survey Committee

Asian American Student Coalition
Association for the Anthropological Study of Consciousness
Arkansas Association of Student Councils
Alumnae Association of Smith College
African-Americans and South Carolina
Australian Archaeological Survey Consultants Pty Ltd
Ashfield Amateur Swimming Club Inc
AASC - Autodesk Animator codec
Aluminum Alloy Stranded Conductor
Assam Administrative Staff College
Atlanta Aesthetic Surgery Center
Active After-school Communities
Alopecia Areata Support Community
African American Studies Center
African American Student Center
African-American Student Coalition
Auditing and Assurance Standards Council
Applied Aerospace Structures Corp
Augusta Arsenal Soccer Club
American Association of Sporting Clubs

Table 31: Example of the output for the trial for "AASC"

Table 31 shows an example of the result for the acronym *AASC*. Notice that there are three main clusters of documents for these acronym: one for *Asian American Studies Center*, with 9 documents, another for *American Association of State Climatologists,* with 4 documents and finally another for *Arizona Association of Student Councils,* with three documents. After that, the proportion of documents per cluster follows a Zipfean distribution, meaning that there are few clusters that have many documents and many clusters that have few documents, in particular two documents and isolated documents. Having this distribution means that there are a few prominent uses and then a great variety of other different uses. This distribution depends on the number of uses of an expression[41] and is the reason why the algorithm only keeps a cluster if it has more than two members.

Overall figures for the quality of the results will be given in the next section, but first let us examine, in this example, what a good or a bad result is. In the result, there are different aspects that must be examined. First, the algorithm has found 3 out of 5 prominent senses of the acronym in the downloaded collection. There were two groups of documents, of three documents each, one about the *Asian American Student Association* and the other about the *Academic Assembly Steering Committee*, which the algorithm was not able to detect. However it succeeded in the

[41]Notice that referential expressions like acronyms have much more different uses than non-referential expressions. This is clearly demonstrated by inspecting the number of different uses that acronyms have in the experiments shown in Section 7.3.3. in comparison to the replication of the experiment with non-referential expressions that will be shown in Section 7.3.5.

detection of the three prominent clusters, those which comprised the largest numbers of documents. Another aspect that can be examined in this result is the internal coherence of each cluster. For instance, the algorithm assigned 9 documents to the first cluster, the one labeled as the *Asian American Studies Center*. Only 6 of those documents were relevant to the cluster, thus it means that for that particular cluster the algorithm performed with 66% precision. The recall, however, is in this case of 100%, because the 6 documents that were correctly assigned to the cluster were all the relevant documents that existed in the downloaded collection.

7.3.3 Evaluation

The evaluation consisted in determining whether the clusters found by the algorithm were internally consistent -in the above mentioned sense- and in determining how many relevant documents were left outside of the corresponding cluster. Evaluation is expressed, thus, in precision and recall from two different aspects: with respect to the documents and with respect to the clusters.

As explained in 7.3.2, precision with respect to documents is the proportion of documents that were correctly assigned to a cluster, while recall is the proportion of relevant documents that were effectively classified. In this first measure, the mean of precision that the algorithm achieves is 89% and recall is 88%. That is, 89% of the times that the algorithm assigned a document to a cluster it did it correctly, and that 88% of the total number of documents were classified.

Precision with respect to the clusters is measured by the number of clusters correctly identified by the algorithm, while recall would be the proportion of clusters identified in relation to the total number of clusters that should be detected. The total amount of spurious clusters generated is 11%, and for causes that are well known (see discussion in section 7.3.5). Precision with respect to clusters is, thus, estimated in 89%.

Acronym	Detected Uses	Undetected Uses
AASC	3	2
APCS	5	1
ASG	2	0
BVM	4	0
CKD	2	0
DDO	3	0
ETN	1	1
FYI	2	0
IED	4	0
JUB	6	1
KPS	7	0
KSP	5	1
LEP	4	1
NAFTA	1	0
NCO	5	0
NPN	2	2
OLA	0	0
OOV	5	0
PCR	3	0
PTW	6	0
RLA	5	0
SEU	3	0
TDD	2	0
TLA	8	1
WTF	5	1
Total	93	11

Table 32: Evaluation of the results

Table 32 shows the number of detected clusters per trial against the number of senses found manually for the respective collection of documents that could have formed a cluster but were undetected by the algorithm. There, each row is an experiment, involving each of the ambiguous acronyms. The first column shows the different acronyms of the sample, the second column shows the number of uses per acronym that the algorithm was able to detect, and the third column shows the number of uses per acronym that the algorithm overlooked. The first case, for *AASC*, was already examined in Table 31. As it was shown there, in this case the algorithm detected 3 uses but overlooked 2. However, the total number of clusters found by the algorithm after the 25 experiment is

93, leaving only 11 groups of documents undetected, thus a very good recall with respect to clusters (89%).

7.3.4 Replication of the Experiment with the HyperLex Dataset

As already stated in section 7.3, the algorithm was applied to another dataset in order to obtain a reference for evaluation. This was the dataset used to test the HyperLex disambiguation algorithm (Véronis, 2004), kindly provided by the author for this replication. The kind of problem that HyperLex confronts is slightly different than the experiment that has just been shown in 7.3.3. Instead of disambiguating acronyms, Véronis reports an experiment disambiguating the uses of polysemous words in French. Despite these differences, the comparison with Véronis' results is a useful measure of performance. Recall the details of his experiment, already commented in Chapter 5. He presents a set of ten arbitrarily selected French polysemous words: *barrage*, *detention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* and *vol.* The first of these words is presented in Table 33. A larger table with the rest of the examples is shown in Appendix B. The problem consists in determining which are the different uses that each word shows in a given corpus. The evaluation, in the case of Véronis, is done manually by a native speaker who tags each instance of those words in the corpus. The corpus was downloaded from the web in the year 2002 (Véronis, personal communication). Table 33 also shows, in the first column, the different uses that the native speaker found, while the second column shows some examples of contexts for each use, according to the speaker's own criterion. The idea for replicating this experiment is thus to compare the performance of both algorithms on the same dataset taking the same native speaker as reference.

Uses of word *barrage*	
Use	**Example**
routier	Barrages non autorisés . Il y a un mois , le ministre d'Etat , ministre de l'Intérieur , de la Sécurité et de la Décentralisation , Emile Boga Doudou , se réjouissait de la baisse de la criminalité en Côte d'Ivoire . Il expliquait cette baisse par la levée des barrages et par l'instauration des patrouilles mobiles . Un mois plus tard , le constat est toujours alarmant. Catastrophique , même . Non seulement les barrages illégaux sont toujours en vigueur mais les forces de l'ordre redoublent de férocité . S'estimant mal payées, elles revendiquent un droit de passage auprès des automobilistes. 500 F CFA (5 FF) par conducteur. Les chauffeurs de taxi et les transporteurs routiers sont les premiers à se plaindre des agissements des policiers .
frontie re	La défense des frontières par les barrages terrestres est une des préoccupations majeures du commandement depuis que le Maroc et la Tunisie sont devenus indépendants, c'est-à-dire depuis 1956 et jusqu'au jour de l'indépendance

Uses of word *barrage*	
	algérienne. Les barrages sont une création continue ; ils évoluent en fonction du développement des armées de libération nationale (ALN) se constituant à l'extérieur des frontières. Si les passages à travers les barrages se tarissent, les harcèlements se développent et la crainte d'un passage en force d'une ALN de Tunisie bien armée est entretenue jusqu' en 1962. La construction, l'entretien et la défense de ces barrages sont une mission interarmées puisque y participent marins (1) et aviateurs aux côtés de toutes les armes de l'armée de Terre, comme l'illustrent les études et témoignages ici réunis.
match	1-VI Barrages, 1ère journée Profitant de la victoire aux tirs au but, qui ne rapporte que deux points, de Saint-Brieuc face à Segré, le FC Mantois prend la tête du mini-championnat après son succès 2-1 face à la réserve de Sedan. Il ne fallait pas arriver en retard au stade Aimé-Bergeal... ni partir avant la fin du match! Après 120 secondes de jeu, les Ardennais débloquent la marque sur une tête lobée de Harreau. Piqués au vif, les Mantois se ruent à l'assaut des buts sedanais et égalisent logiquement sur un coup de tête de Rameau (23e). La rencontre est haletante avec de nombreuses occasions de part et d'autres ; la plus belle est pour les Mantois avec un tir de Preira sur la barre (70e). La dernière poussée des locaux est la bonne : lancé en profondeur par Challouf, Ictoi inscrit le but de la victoire à l'entame des arrêts de jeu! Saint-Brieuc est passé tout près de la victoire dans le temps réglementaire : alors que le score est de 1-1, les tentatives de Morin (79e) et Desriac (80e) trouvent les montants segréens!! Grâce à deux parades de leur gardien Druguet, les Griffons remportent la séance de tirs au but, 4-2. 1-VI Mantes 2, Sedan Rés. 1 Rameau (23e), Ictoi (90e+1) pour Mantes ; Harreau (2e) pour Sedan Rés. 800 spec. Saint-Brieuc 1, Segré 1 Saint-Brieuc vainqueur aux tirs au but, 4-2. Allainmat (50e) pour Saint-Brieuc ; Vallais (p 31e) pour Segré. 500 spec.
eau	Les barrages mobiles Les barrages mobiles sont des ouvrages qui sont construits dans la partie aval (basse) du cours des rivières ou la pente est faible. Ils sont constitués de bouchures (grandes vannes) et d'une infrastructure en béton. L'avantage des barrages mobiles : On peut règler l'ouverture des vannes et ainsi réguler l'arrivée d'eau selon le niveau de crue de la rivière . D'autre part , on les associe à des écluses permettant le franchissement, pour la navigation, de la chute crée par le barrage . Pour favoriser une production optimale , il faut donc non seulement étudier duquel on tire différentes édifications mais il faut aussi étudier les hauteurs de chute pour éviter les risques tout en produisant au maximum. Pour montrer ces différentes hauteurs de chutes nous présenterons le tableau ci-après énumérant plusieurs hauteurs selon le type de construction ou de matériaux avec lesquels ils sont édifiés.

Table 33: Uses of the polysemous French word “barrage”, annotated by a native speaker

The dataset was submitted for execution to the same algorithm described in Section 7.3.1, analyzing each unit and determining which are the uses that the unit exhibits in the corpus, as well as to select examples of contexts of occurrence that correctly reflect that usage. The way to evaluate the results, in this case, is to calculate the proportion of uses that the algorithm was able to isolate in comparison to those that were isolated by the native speaker in the same sample. Another parameter that must be evaluated is the internal consistency of the examples selected for that use. That is, whether the contexts selected for each cluster are indeed correct. Notice that this second parameter is another way of looking at the

problem of tagging each polysemous word in context with the appropriate use, which is the kind of problem that Véronis is interested in. Thus, there are two kinds of problems, although they are inherently related. The first one is, given a polysemous word in some language (and some sample of contexts to take the data from), to discover which are the different uses that this word has. Once this has been done, the second problem can be solved: to tag each word in context with the correct use (i.e., to disambiguate). While this thesis is focused on the first problem, Véronis, instead, is more interested in the second. Therefore, in order to compare the behavior of both algorithms, the one in this thesis must be applied as a disambiguation algorithm.

As an example of the result of this experiment, Table 34 shows the clusters made by the algorithm after analyzing the web pages where the word *barrages* occurs. Each cluster has a name, which is automatically assigned by the algorithm, the list of the document identifiers, and the tag which was assigned by a native speaker according to the sense in which the word is used in each context. Notice the remarkable internal coherence that is present in the clustering, showing a high degree of agreement with the human criterion. The only disagreement occurs in position 18 of Cluster 2, that a document tagged with the sense of *routier* by a native speaker, is mistakenly placed in a cluster where all documents have received the tag *eau*. Recall is also high in this case. Only 10 documents were left out without being assigned to a cluster.

Cluster 1 **Name:** *barrages_routiers*			**Cluster 2** **Name:** *bassins_versants*		
#	Doc ID	Hand tag	#	Doc ID	Hand tag
1	1451	routier	1	1748	eau
2	2832	routier	2	1749	eau
3	2834	routier	3	1750	eau
4	3007	routier	4	2625	eau
5	3010	routier	5	2859	eau
6	3046	routier	6	3031	eau
7	3066	routier	7	3338	eau
8	3069	routier	8	3627	eau
9	3168	routier	9	4005	eau
10	3265	routier	10	4269	eau
11	3408	routier	11	4278	eau
12	3995	routier	12	4463	eau
13	426	routier	13	4478	eau

14	4707	routier
15	5461	routier
16	547	routier
17	5500	routier
18	5663	routier
19	6476	routier
20	6495	routier
21	6560	routier
22	6601	routier
23	6713	routier

14	4595	eau
15	5209	eau
16	536	eau
17	5460	eau
18	5528	routier
19	5554	eau
20	5688	eau
21	5692	eau
22	5798	eau
23	6088	eau

Cluster 3
Name: *frontières_algéromarocaine*

#	Doc ID	Hand tag
1	2009	frontiere
2	2225	frontiere
3	3073	frontiere
4	3600	frontiere
5	3892	frontiere
6	3895	frontiere
7	3983	frontiere
8	4158	frontiere
9	4180	frontiere
10	4252	frontiere
11	4304	frontiere
12	4510	frontiere
13	4737	frontiere
14	5393	frontiere
15	5717	frontiere
16	5924	frontiere
17	5955	frontiere
18	806	frontiere
19	5554	frontiere
20	5688	frontiere
21	5692	frontiere
22	5798	frontiere

Cluster 4
Name: *gardien_druguet*

#	Doc ID	Hand tag
1	1473	match
2	1554	match
3	1555	match
4	1957	match
5	1958	match
6	3922	match
7	4140	match
8	5211	match
9	5227	match
10	5480	match

Cluster 5
Name: *dépourvu_aviation*

#	Doc ID	Hand tag
1	4593	frontiere
2	5015	frontiere
3	4291	frontiere
4	5883	frontiere

23	6088	frontiere

Cluster 6
Name: *barrage_électrifié*

#	Doc ID	Hand tag
1	1462	frontiere
2	2017	frontiere
3	4291	frontiere
4	4576	frontiere

Cluster 7
Name: *barrages_servent*

#	Doc ID	Hand tag
1	287	eau
2	5072	eau
3	759	eau
4	4576	eau

Table 34: Clusters made by the algorithm with “barrage”

With respect to the first problem, which is to discover the possible senses of words in accordance with the native speaker's criterion, the evaluation for the ten words is shown in Table 35, below. The procedure for evaluation is similar to that of the previous experiment. For instance, in the case of *barrage*, the native speaker detected four different uses of the word in the collection, and all of these uses were reflected in the clusters generated by the algorithm. There are, however, seven clusters. This is because there is a more finely graded partition of senses than the native speaker identified. While for the speaker there is only one cluster for *barrage* in the sense of *frontiere* (frontier), for the algorithm there are three more specialized clusters. For instance, in Cluster 6, on Table 34, the word barrage refers to electrified devices for the defense of frontiers. The proportion of clusters per senses is also included in Table 35, where, for each word the number of resulting clusters can be seen, the clusters that were detected, and the total number of senses (the ones that the native speaker found). Based on the results of this experiment, it can be estimated that the chances of correctly finding a given sense for a given word in a given collection of contexts of occurrence is around 88%, enough to be considered useful in real applications, such as the discovery of semantic neologisms (see Chapter 9).

Term	Clusters	Detected Senses	Total Senses	%
barrage	7	4	4	100
detention	6	4	4	100
formation	1	1	1	100
lancement	5	2	2	100
organe	12	7	8	87
passage	16	6	9	66
restauration	13	6	7	85
solution	6	3	3	100
station	10	6	8	75
vol	3	3	4	75
Average				**88**

Table 35: Discovery of senses

	Low recall run		High recall run	
Term	**% P**	**% R**	**% P**	**% R**
barrage	99	40	99	90
detention	94	74	73	100
formation	100	12	100	100
lancement	98	66	90	98
organe	99	66	97	94
passage	100	42	89	80
restauration	95	65	89	91
solution	100	64	100	88
station	98	47	84	82
vol	100	51	91	78
Total	**98**	**52**	**91**	**90**

Table 36: Internal consistency of the clusters

Another aspect of the evaluation is the internal consistency of the clusters that have been formed. This is evaluated taking as reference the most frequent tag assigned by the speaker. Therefore, 100% precision would be the case in which all the members of a cluster have been assigned the same tag by the native speaker. Table 36 shows the mean of the precision of the clusters per word. The results in this table are divided into two runs. In the first case, there is a run with low recall, where only the most reliable results are given, therefore, providing better precision to the detriment of recall. In the second run, recall is improved to the detriment of precision by forcing the algorithm to take more risky decisions. With

respect to internal consistency of the clusters, having around a 90%-90% trade-off between precision and recall can be considered good enough to be considered as an application in real case scenarios.

With respect to the second problem, which is to tag every instance of the polysemous word in text with the appropriate sense, Table 37 shows the comparison between the results achieved by this algorithm and Hyperlex, as well as a baseline also provided by Véronis. This baseline is obtained by assigning the most frequent word use to each instance in the collection, which is a common practice in the field of Word Sense Disambiguation. The second column shows the precision achieved by Hyperlex. Véronis does not show figures of recall for every single word. Instead, a general figure of recall is calculated by a reliability coefficient, which allows for a recall of 82% in the whole collection. For this reason, the third column shows the precision achieved by this algorithm also at 82% recall, in order to make the figures comparable. This was not possible in two cases, more precisely in *passage*, where there was only 80% recall, and *vol*, which resulted in 78%. Although different, the figures are still comparable to the expected 82% recall.

Term	**Baseline**	**Hyperlex**	**This Experiment**
barrage	77	100	99
detention	87	100	85
formation	100	100	100
lancement	99	100	97
organe	40	88	98
passage	52	88	88*
restauration	44	100	90
solution	84	98	100
station	84	100	83
vol	62	100	88*
Total	**73**	**97**	**93**

Table 37: Precision in tagging each context at 82% recall
(* cases where 82% recall is not reached)

HypeRlex performs very well, with 97% precision in total. The precision of this algorithm is 4% under this mark, but there is still 20% improvement over the baseline. In three cases, however, the baseline performed better (*detention*, *lancement* and *station*) and in two cases this algorithm performed better than Hyperlex (*organe* and *solution*). Surely, more than ten trials would be necessary for a reliable estimation, though this is left for future work.

7.3.5 Discussion of the results of this Experiment

This experiment has demonstrated that it is possible to cluster a collection of documents or contexts of occurrence of a polysemous term, on the basis of the different senses of such term. It is important to recall some of the characteristics of this algorithm, which were introduced in Chapter 6. In first place there is no hard-coded information, either from ontologies or from the language under study. It is an unsupervised learning approach, and it needs no training phase previous to the task. Therefore, the system does not have prior or external knowledge about which or how many of the senses the linguistic sign under analysis has. Another characteristic is that it is much faster than clustering algorithms that use distance matrices which entail quadratic complexity (such as Aldenderfer & Blashfield, 1984) because their execution time grows exponentially with the size of the input data, while the execution time of the algorithm presented here grows linearly with the input data size.

There remain, however, a number of errors, fundamentally because the language model does not include units like *advanced search*, *email address*, *web page*, *web site*, *flash player*, *contact us* or certain symbols or sequences of characters such as →, which are typically found in web pages. In a real case scenario, these errors could be corrected by creating a stopword list, either manually or automatically, meaning that the precision documented in this research may increase. In the context of this research, however, it would not be legitimate to compile such a list manually. It would be possible, though, to compile it in an automatic way using statistics because these are units that have a very low informational content, since their occurrence in a given document is not inherently related to the content of the document (see section 7.3.6). In any case, the purpose of this experiment was only to provide empirical evidence in support of the claims presented in Chapter 2.

In the case of the experiment with acronyms, there is a natural reason for the low number of documents assigned to clusters: web documents are characterized by their diversity, and the distribution of documents per clusters follows a Zipfean distribution. Approximately 25% of the clusters have only one document, 15% have only two, and the series end with few clusters of more than three documents (depending, of course, on the number of documents that make up the sample). This distribution is very regular across most samples. Whether the algorithm can be forced to have a higher recall, as it was done with the replication of Véronis' experiment, depends on the corpus. This is an interesting fact in itself: in this sense, acronyms do not behave in the same way as polysemous words do. While the number of possible uses of an acronym follows a Zipfean distribution, this does not happen with common polysemous words because they are used less often.

In relation to the replication of Véronis' experiment, the conclusions are that even though the algorithm presented in this thesis does not supersede his results, the performances of both are at least comparable, and the fact that the one presented here is language independent is enough reason to believe that it is worth pursuing this line of research. As already mentioned, there is a difference in the number of clusters that this thesis' algorithm generated in comparison to the criterion of the native speaker. For instance, one of the uses in French for the word *restauration* is fast food. But while for the native speaker this was one whole cluster, for the algorithm there were different kinds of fast food (i.e., one cluster for pizza and another cluster for hamburgers, and so on). Something similar occurs with *barrage*, where there are more clusters than senses, though the clustering still reflect the division of senses. This is an interesting phenomenon in its own right, thus it is not be considered an error of the algorithm.

With the necessary adjustments, it seems clear that there is a potential for application of this prototype to the problem of information retrieval. In the context of this thesis, it has been considered better to present the results without polishing them for two reasons: firstly because in this way it can be demonstrated that the method is effective even under non-ideal conditions. Secondly, because later it will be possible to separate which part of this efficiency is due to the basic idea and which part to the result of subsequent adjustments.

7.3.6 Future Work for this Experiment

Lines of future research go in two directions. The first one is to optimize several aspects of the algorithm and to expand some of its limitations. The second line would be to extend the research and its areas of application.

With respect to the first line, it has been demonstrated that there is a need to elaborate a more complete and updated model of the language of study. It would be important to recognize these units by automatic means because this set will always be renewed with new jargon appearing on the Internet. There is a possibility of an automatic correction of the errors that could be made on the basis of the frequency that noisy units have in the whole set of experiments. The rationale behind this strategy is that a unit that continues appearing as a relevant node independent of the query term that is generating the graph is an indicator of a small amount of information. For instance, a unit like *contact us* has a high probability of occurrence in a corpus downloaded from the web, independent of what the query expression was. A measure similar to this one, called Statistical

Noise Reduction, has already been reported in previous research (Nazar et al., 2008) and is used in the next experiment, section 7.4 of this Chapter.

Another possibility for the algorithm could be to incorporate additional resources such as following the first level of hyperlinks of the web pages that are analyzed in order to obtain more information about them, which in turn would be of help for the proper classification of those web pages. In this same line, information from the URL could also be used, for example to relate documents that come from the same server. This resource was not used in the thesis' experiment because it would be an exogenous source of information and therefore not valuable from the scientific point of view. From a practical point of view, however, it would be worth trying. In any case it would be a complex task, since sharing the same server is by no means a guarantee that the documents are related from the semantic point of view.

With respect to the second line, the application of the algorithm to real life cases, several example of application are described in Chapter 9. However, one line of research that has already started, beyond the most obvious case that would be the disambiguation of words or terms, is the application to the detection of semantic neologisms (Nazar & Vidal, 2010).

7.4 Fourth Experiment: Acquisition of Bilingual Terminology

This experiment exploits a variant of the algorithm presented in Chapter 6. It is aimed at the extraction of bilingual terminology from unrelated corpora (i.e., the web) and it is meant to be both an empirical verification of both hypotheses of the thesis and an example of a practical application. The main idea in this experiment is to develop a software that, given a query term in a source language, finds the appropriate translation into the target language. It will be done by observing which is the term, in the target language, that has the most similar graph to the one generated from the term in the source language, provided that one can identify the equivalent nodes of both graphs.

Related literature on the basis of paradigmatic statistics was introduced in Section 5.1.2.3. There, it was shown that co-occurrence graphs that are generated from equivalent terms have significantly similar profiles of co-occurrence (Harris, 1954; Grefenstette, 1994; Schütze & Pedersen, 1997; Curran, 2003; etc). Starting from this premise, the task of finding equivalent terms in different languages can be cast as the task of finding a proper similarity measure for the comparison of graphs of terms of different languages. The terms whose graphs are sufficiently similar

should correspond to equivalent terms. A similar intuition was followed by Fung (1998) and Rapp (1999), among others, who also apply the distributional hypothesis to extract bilingual dictionaries. They calculate different similarity coefficients between co-occurrence vectors to find equivalent words in different languages. The present experiment builds on this idea, and applies it to the extraction of specialized terminology with a different methodology.

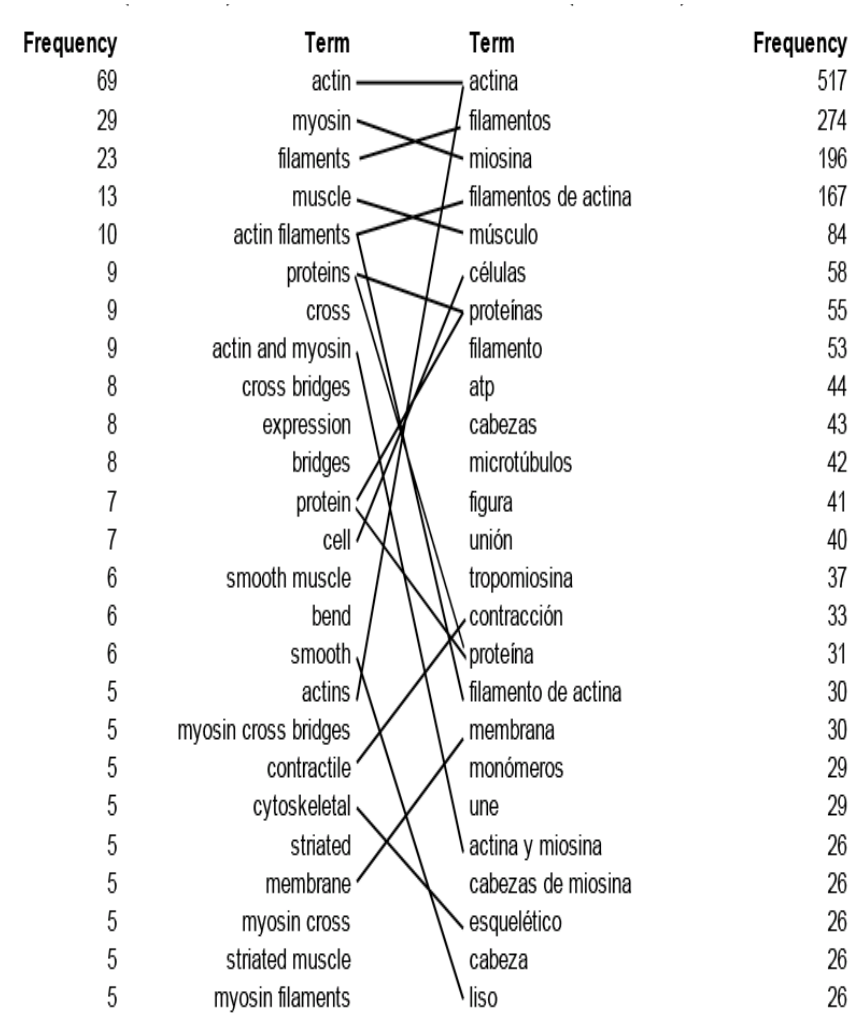

Figure 43: Most frequent *n*-grams in 65 contexts of the English term "actin" (left) and 418 contexts of the Spanish term "actina" (right)

As an illustration of the methodology proposed in this Section, consider again Figure 3 in Chapter 2 (reproduced here in Figure 43), which shows an example which was built out of concordances extracted from the CT-IULA corpus in random order. Recall that this figure represents the units

that occur more frequently in a context window of one sentence with the term *actin*, in the case of English, and *actina*, which is the Spanish equivalent. Figure 43 represents the co-occurrents as vectors in one dimension and shows the coincidence of the two groups of co-occurrents, a phenomenon which will allow us to undertake the experiment of bilingual lexicon extraction.

One of the major challenges of this experiment is that the profiles of the co-occurrence of both the query term and the evaluated candidate are not directly comparable because the units which make up their graphs are in different languages. This is why an initial bilingual dictionary is needed, and its relevance is critical because the quality of this resource will directly affect the quality of the results. Obviously, the idea is not to simply look for the translation of the input term in the dictionary. The idea would be, on the contrary, to try to find the translation of terms that are not contained in the initial bilingual dictionary. The English-Spanish vocabulary that was used in this experiment was the Wiktionary[42], a basic vocabulary of around 10000 entries of Spanish-English equivalents.

Probably the greatest difficulty, leaving aside the potentially insufficient coverage of the nomenclature of the bilingual dictionaries, is that these dictionaries may return more than one translation. To reduce this risk, the methodology avoids using terms that propose more than one translation for a given term. Of course this is no guarantee, because even with one meaning the translation returned from a dictionary may imply an unfortunate bias for the experiment. For example, if the English term *component* is only translated as *ingrediente* (ingredient), this may be appropriate in some contexts, but in this case the term *componente* would have been better not only because *ingredient* is useless here but also because it will produce noise by relating to other words that are semantically unrelated to an initial query term.

As a final comment for this overview, consider also that working with non-parallel bilingual corpora presents challenging difficulties especially because, frequently, documents have fragments in different languages. This is particularly the case with scientific literature, because the abstracts are usually in English, as are the references to other studies. This situation prevents us from giving a uniform treatment to the vocabulary of the corpus, in that it requires a previous identification of the language of each fragment of text in each document.

7.4.1 Experimental Set-up

This section presents the algorithm (Figure 44) in the form of a list of instructions and a description of the coefficients that are used. With the

[42] The Wiktionary (http://www.wiktionary.org/) [accessed June 2010].

exception of step 3 (a function that checks for the existence of a term in a database), steps 1 to 4 perform as explained in the chapter on methodology (Chapter 6). This means that the output of the algorithm that draws the graphs is the input for this experiment.

Bilingual Terminology Extraction

$t \leftarrow$ term to be translated
LM $\leftarrow$ Language model // as described in section 6.3
L // bilingual Lexicon
$n \leftarrow 300$ // number of documents to download
$k \leftarrow 4$ // minimum number of documents to output a result
$e \leftarrow 10$ // minimum number of secondary connections to accept a node
$m \leftarrow 3$ // minimum frequency of a term in the downloaded collection

Do:

1- Download n documents in the source language using t as query.
2- Draw a co-occurrence graph $G(t)$
3- create a vector V_t with t as first value.
4- Include in V_t any component of t that can be translated with L.
5- Sort nodes in $G(t)$ in decreasing number of secondary connections.
for each node $i \in G(t) \wedge |V_t| < e$ do
5.1- translate i into i' using L unless L provides multiple translations.
5.2- if translated, include i' in vector V_t .
endfor
6- D $\leftarrow$ Download n documents in the target language using V_t as query.
7- return 0 if $|V_t| = 0$.
8- if D < k delete the last element from V_t and go to Step 6
} else {
$C \leftarrow$sort candidates // sort the vocabulary of the downloaded collection in decreasing frequency order (vocabulary defined as n-grams of up to 5 grams)
$CM \leftarrow$initialize candidate matrix // stores the values of each candidate according to coefficients.
for each $c \in C \wedge c > m$ do
CM(Cf-1, c) $\leftarrow$Weight based on LM(t, c)
CM(Cf-2, c) $\leftarrow$Length Ratio(t, c)
CM(Cf-3, c) $\leftarrow$Expected Frequency Ratio(t, c)
CM(Cf-4, c) $\leftarrow$Measure of Expected Dispersion(c)
CM(Cf-5, c) $\leftarrow$Measure of Observed Dispersion(c)
CM(Cf-6, c) $\leftarrow$Orthographic Similarity Measure(t, c)
CM(Cf-7, c) $\leftarrow$Statistical Noise Reduction(C, c)
CM(Cf-8, c) $\leftarrow$Context Language Detection(c)
endfor
EliminationRules(CM)
return FinalRank(CM)
}

Figure 44: Pseudo-code for the extraction of bilingual terminology

Step number 7 is critical because, once a collection of translation candidates for a given input term is gathered, the sorting of these candidates begins. Each candidate is placed in a matrix *CM* where it receives values for a number of coefficients that will be described in

detail below. The absolute frequency of a candidate in the downloaded collection is not considered a coefficient on its own, but it does contribute to the final ranking of the candidates. The eight coefficients[43] that are in use[44] are shown in Table 38.

Coefficients
Cf-1: Weight based on a language model
Cf-2: Length Ratio
Cf-3: Expected Frequency Ratio
Cf-4: Expected Dispersion
Cf-5: Document Frequency
Cf-6: Orthographic Similarity Measure
Cf-7: Statistical Noise Reduction
Cf-8: The Language of the Context

Table 38: Names of the involved coefficients

The set of "Elimination Rules", shown in Figure 45, has the purpose of eliminating candidates that are considered false according to the values that the candidates obtained with the coefficients.

[43]Note that the use of these coefficients does not necessarily imply that they are linguistic generalizations. Many of these coefficients are useful in the scope of this experiment on specialized terminology. However, this does not mean that there is a general correlate within all languages concerning the length, the expected frequency or the orthographic resemblance of equivalent words in different languages. This comment, however, does also not imply the contrary. It is, in any case, an interesting topic for a another study.

[44]The order in which the coefficients are applied is irrelevant from the theoretical point of view but relevant for reasons of computational efficiency.

Elimination Rules

v_j ← candidate to evaluate
d ← number of downloaded documents
q ← 1000 // maximum frequency in language model
p ← -3 // minimum weight
d ← 3 // minimum document frequency
r ← 0.2 // minimum length ratio
s ← 0.3 // minimum expected frequency ratio

A candidate is eliminated if it meets any of the following conditions:

1- it is a multiword unit that begins or ends with an element with a frequency greater than q in the target language model (Cf-1).
2- it has as components units of a length of only one character.
3- it is more frequent in the source language model than in a target language model (Cf-1).
4- it has a weight less than p in the language model described in Section 6.3.
5- it has a length ratio (Cf-2) less than r.
6- its document frequency (Cf-5) < d.
7- it has an expected frequency ratio (Cf-3) less than s.
8- it includes a word that has a frequency greater than q in the source language model.
9- its contexts of occurrence are mostly in the source language (70%)

Figure 45: Rules for the elimination of false candidates

Once all the elimination rules have been applied and the maximum number of false candidates has been eliminated, the Ranking Coefficient is applied, which is the normalized sum of the values of all the coefficients for the remaining candidates. A detailed description of all the coefficients follows:

Cf-1: Weight based on a language model

This coefficient is based on a language model as used in previous experiments (see section 6.3) which is basically the frequency that a given unit has in a corpus of general language. It informs the relative frequency of a unit i in a source language model –Fs(i)– and its frequency in the target language model – Ft(i). It is used by the Elimination Rule 1 to eliminate candidates that are frequent but uninformative, such as *would* or the sequence *has been* in the case of English. These are mostly grammatical units (empty words). It is also used by the Elimination Rule 3, which eliminates a translation candidate v for the term t if Fs(v) > Ft(v), because this would indicate that v is a word of the source language.

Cf-2: Length Ratio

At least in the context of technical terminology, it is often observed that pairs of equivalent terms in different languages are more similar in length than pairs of terms that are not equivalent. Building on this assumption, and without claiming any linguistic generalization (see note 43), let length(*i*) be the length in characters of a term *i.* Formula 23 defines the coefficient Cf-2(*i*,*j*) for the length ratio of two text strings *i* and *j*.

$$\text{Cf-2}(ij) = \frac{\min(\text{length}(i), \text{length}(j))}{\max(\text{length}(i), \text{length}(j))} \tag{23}$$

The idea is not to find an exact match in length (equivalents rarely have exactly the same length). Thus, this coefficient is used only if its value is higher than a lower threshold (0.2) and smaller than an upper threshold (0.7). If the value is equal or smaller than the lower threshold, the resulting value of this coefficient will be considered 0. If the value is equal or higher than the upper threshold, it will be considered 1. Any other value in the interval between the two thresholds is left without transformation

Cf-3: Expected Frequency Ratio

Using the values Fs(*i*) and Ft(*i*) defined in Cf-1, which represent the frequency of a term *i* in source and target language models, this coefficient 3 now defines a ratio of expected frequency between the two. The frequency in the language model of a term *i* and its translation candidate *j* gives us the ratio of expected frequencies, and is expressed in Formula 24.

$$\text{Cf-3}(ij) = \frac{\min(\text{Ft}(i), \text{Fs}(j))}{\max(\text{Ft}(i), \text{Fs}(j))} \tag{24}$$

This coefficient is useful, for instance, to infer that a unit such as *este trabajo* (this work) is not a good candidate for the translation of *alkyl group*, because *este trabajo* is much more frequent in the Spanish (target) language model than *alkyl group* is in the English (source) language model. In contrast, the correct candidate, which is *grupo alquilo*, has the same zero frequency in the language model as *alkyl group*. It should be obvious, however, that the fact that a candidate that has the same zero expected frequency as the term to translate is not enough reason to select the candidate as the translation, but it is enough reason to consider others as bad candidates

when their respective expected frequencies are not comparable, as in the above example.

In the case of multiword candidates, it is difficult to estimate an expected frequency for them since there is an increased chance of not having them in the language model. One possibility, then, is to calculate the overall expected frequency on the basis of the expected frequency of each of the components of the multiword term, though it is obviously not an ideal solution since it entails the risk of a wrong estimation. In this case, the calculation of the ratio ignores the components that are too frequent in the language model -the so-called empty words-, because multiword terms may include them and this would distort the ratio (e.g. the elements *of* and *the* in the term *Cancer of the esophagus*).

Cf-4: Expected Dispersion

Another way in which the language model can help us in measuring the information of a given lexical unit in a corpus is by relating the frequency of the unit to its document frequency (or sample frequency) within the language model. This is, again, the application of the model of language explained in Chapter 6.2. It is to be expected that an uninformative unit will not only be frequent in that corpus but also more or less homogeneously dispersed within such corpus. It is called a measure of Expected Dispersion (Formula 25), because it is based on the dispersion observed in the language model. As already explained in Chapter 6, Mf(t) is the maximum frequency of unit t in one of the partitions of the language model and Cr(t) the number of partitions with a frequency of t less than or equal to a parameter k.

$$\text{Cf-4}(t) = \text{Mf}(t) \,.\, \text{Cr}(t) \tag{25}$$

Cf-5: Document Frequency

This coefficient simply counts the number of documents in the downloaded collection where a candidate occurs. The correct translation of a term is not only expected to be the most frequent among the downloaded collection, but also to have a high document frequency. If the downloaded collection for the current experiment has more than five documents, this coefficient is also used by elimination rule 6, which eliminates all units that have a document frequency less than 3. This implies a significant reduction of the vocabulary size and computational cost.

Cf-6: Orthographic Similarity Measure (Dice)
This is an application of the vector similarity measure coefficient for the detection of an orthographic resemblance between two units. First, both units are transformed into vectors *X* and *Y*, which have sequences of two characters as components. Then, a Dice similarity coefficient is computed (Formula 26), as is explained in Chapter 6.

$$\text{Cf-6}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{26}$$

The coefficient is applied in two different ways. The first one is a pseudo-lemmatizer in order to facilitate frequency counts of the candidates. Candidates are subjected to a pairwise comparison using this coefficient, and if two of them are too similar (using a threshold of similarity of 0.85) then they are agglutinated as if they were the same word, summing their respective individual frequencies.

The second way in which this coefficient is used is to detect an orthographic resemblance this time not between the different candidates among themselves but between each of the candidates and the input term. This is because, as is usually observed especially in scientific or technical domains, term equivalents in different languages are cognates. It is perhaps important to remark that the idea is not to select a candidate as the correct translation only on the basis of an orthographic similarity with the input term, since the correct translation is not necessarily going to be a cognate. As it has been stated in the explanation of other coefficients, the idea in this case is to provide another small contribution to what will be the final ranking of a candidate.

Cf-7: Statistical Noise Reduction
When one is translating a list of terms and those terms happen to be of the same domain (such as diseases, in this case), an interesting phenomenon occurs, and this is the persistence of a set of lexical units that represent noisy vocabulary. But this is a noise of a specific nature, a domain-specific type of noise. Within the domain of diseases, for instance, there is a set of units which persistently appears as equivalent candidates with independence of the query input term, such as *pacientes* (patients); *síntomas* (symptoms) and *frecuente* (frequent), among others. Thus, even when the set of units changes depending on the domain and language, this noise

can be statistically reduced by using a distributional criterion. These units have an exaggerated dispersion among the sets, thus their weight can be reduced accordingly to make up for their high frequency. This means that it is possible to generate a sort of "dynamic stoplist" on run time with units that have been proposed as a candidate *n* number of times. A threshold *g* is the proportion of the total number of terms to translate rather than an absolute value. In this experiment, $g = 0.25$. To put it formally, let us denote as *S* the number of trials in this experiment (the number of terms in the sample to be translated). As expressed in Formula 27, a candidate *t* will be penalized with a high value depending on the number of times that it has appeared as translation candidate for other trials.

$$\text{Cf-7}(t) = \sum_{j=1}^{|S|} \begin{cases} 1 & \text{if } t \in S_j \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

Cf-8: Context Language Detection

This final weighting coefficient is reliable and simple to implement. It is placed at the end of the list because it is computationally costly, therefore it is applied only to the final list of candidates, which is supposed to have fewer units than at the previous stages of the process. Out of a random sample of 20 contexts of occurrence of a translation candidate, this coefficient counts the number of contexts which are in the target language. The greater this number is, the higher the probability that it is a correct translation, unless the term is enclosed by parenthesis or marked with some typographical convention such as italics[45], because that would indicate that the sign is extraneous to the language of the text. This coefficient is used by elimination rule 9, which eliminates a candidate if the ratio of source language contexts vs. target language contexts where the candidate occurs is below a certain threshold (0.7). Formula 28 defines Cf-8, where Sctx(*t*) is the number of source contexts from the sample where the query term *t* occurs and Tctx(*t*) the number of target contexts in the same sample.

$$\text{Cf-8}(t) = \frac{\text{Tctx}(t)}{\text{Sctx}(t) + 1} \tag{28}$$

[45]Of course this is impossible to do when working with plain text which has no format.

Final Ranking Coefficient

The candidates that remain after the application of the elimination rules are submitted to final ranking. In order to produce a single final ranking, the values that each candidate obtains from each of the coefficients must be normalized. In this way, all coefficients will range between 0 and 1, and therefore the values will be comparable. The normalization of the values is expressed in Formula 29. The values of each coefficient are divided by the maximum value of the same coefficient. In the normalized coefficient the maximum value will be 1.

$$\mathrm{Norm}(\mathrm{Cf}_{i,t}) = \frac{\mathrm{Cf}_{i,t}}{max(\mathrm{Cf}_i)} \tag{29}$$

Thus, given Ft = {Norm(Cf1,t), ... Norm(Cf_6,t) }, the final rank of *t* or FR(*t*) (Formula 30) is defined as the mean of the values of F_t.

$$\mathrm{FR}(t) = \bar{F}_t \tag{30}$$

7.4.2 Results

This section shows a detailed example of the results for one of the input terms which in this case is *manic disorder*. The next section will evaluate the results for a random sample of 100 terms. According to explanations given in Section 7.4.1, a procedure similar to the one presented in Chapter 6 has been followed with this term *t*, which means that a number *d* documents has been downloaded from the web in the source language, a co-occurrence graph *G*(*t*) has been developed and, from the selection and subsequent translation to the target language of the most prominent nodes in such graph *G*(*t*), a new query to the search engine has been made, this time in the target language. Again, the same number of documents -this time in the target language- is downloaded and converted from their different document formats to plain text. At this point, the vocabulary of the downloaded collection (totaling a number of 157 translation candidates) is sorted in a matrix and organized in *n*-grams.

Rank	E. Rule	Bigram	Freq	Cf-1 (Ft)	Cf-1 (Fs)	Cf-2	Cf-3	Cf-4	...
1)		trastorno bipolar	116	8|0	0|0	1	0.42	4	...
2)		bipolar ii	41	0|232	0|189	1	0.08	15	...
3)	3	~~bipolar depression~~	28	0|0	0|47	1	0	10	...
4)	1	~~major depression~~	63	12|0	1039|47	1	0.79	13	...
5)	3	~~manic-depressive~~	42	0|0|0	0|55	0.583	0	11	...
6)	3	~~manic depression~~	102	0|0	0|47	1	0	17	...
7)	3	~~mood swings~~	51	0|0	56|17	1	0	19	...
8)	3	~~more information~~	23	0|0	244|584	1	0	12	...
9)	1	~~they may~~	29	0|0	1639|2183	0.571	0	9	...
10)	2	~~j psychiatry~~	22	127|0	179|0	1	0.15	4	...
...	...	...	...	...	...	...	...	...	...

Table 39: Example of a listing of bigrams with values of some of the coefficients

Table 39 shows, for example, a list of bigrams obtained from this downloaded collection of documents in the target language along with the values of the coefficients that are relevant at this point to apply the corresponding elimination rules. *N*-grams are organized in different tables, so there will be another table similar to Table 39 for listing single words, another for trigrams, and so on until the desired number of *n* is reached, which in the case of this experiment is limited to 5. Taking a closer look at Table 39, the values produced by some of the coefficients as well as the application of some of the elimination rules can be seen. The candidates that are eliminated are in crossed out and the rows of the table are shadowed. The translation candidate *major depression*, for instance, in row number 4 of the table, is eliminated by rule number 1 (the number of the elimination rule which applied is indicated in the second column of the table) because, as prescribed by this rule, the first component of the bigram (*major*) has a frequency in the source language model, indicated as Cf-1(Fs), of 1039 occurrences. This is taken as indication that the candidate in question does not belong to the target language but to the source language and is, therefore, eliminated. Similarly, other elimination rules apply according to the values produced by the corresponding coefficients. Candidate number 4, for instance, is eliminated by rule 3, because it has components that are more frequent in the source language model than in the target language model.

Final Ranking						
Rank	**Form**	**contexts**	**target cont**	**source cont**	**ratio**	**Cf-8**
1	~~depressive~~	233		21	0	0
2	~~psychiatry~~	131		21	0	0
3	~~bipolar~~	1630		21	0	0
4	trastorno bipolar	115	21		21	0.83
5	~~bipolar ii~~	41		21	0	0
6	~~manic~~	472		21	0	0
7	~~lithium~~	229		21	0	0
8	~~mania~~	390		21	0	0
9	~~mental~~	197	4	17	0.22	0
10	trastorno	166	21		21	0.72

Table 40: Final Ranking of candidates for "manic disorder"

After this first round of eliminations, the remaining candidates are subjected to the final elimination procedure based on coefficient 8, which is the detection of the language of the contexts of occurrence of the candidate. This last rule is separated from the rest for reasons of computational efficiency. Table 40 shows subsequent stage of the experiment with *manic disorder*. This table contains, thus, the best rated *n*-grams which are now listed in the same table regardless of their *n* because the number of terms has now been considerably reduced. In this table, different columns can be seen indicating the number of contexts of occurrence found for each term (column "contexts"). From these contexts, a random sample of 21 contexts is taken and, for each context of occurrence, the language of the context is automatically detected. The "target cont" column indicates the number of contexts that are found to be in the target language while the "source cont" column indicates the number of contexts that are in the source language. A candidate is eliminated if the ratio (indicated in the next column) is below the threshold of 0.7, as already explained. What was originally the first row, for instance, with the candidate *depressive*, is eliminated because all of its contexts of occurrence in the random sample are in the source language (21 contexts) instead of being in the target language, which was to be expected for the correct translation. As already explained, the final ranking of the remaining candidates (FR) is the normalized sum of the values given by the individual coefficients. In this way, the process finishes with the correct candidate, *trastorno bipolar*, as the best rated candidate.

7.4.3 Evaluation

As a preliminary and small scale evaluation, this subsection shows an experiment translating, from English to Spanish, 100 randomly selected terms from the Karolinska (2008) database of names of diseases. Results were then checked using different bilingual dictionaries and databases such as the Mosby dictionary; the MedlinePlus database; the database provided by "The Health on the Net Foundation"[46] and Wikipedia. Table 41 shows some examples from the experiment of 100 terms in English where the translation to Spanish was considered successful.

English Term	Spanish Equivalent	Rank
african lymphoma	linfoma de burkitt	12
alzheimer type senile dementia	demencia senil tipo alzheimer	6
breast neoplasms	cáncer de mama	1
candidiasis	candidiasis	1
chronic obstructive lung disease	enfermedad pulmonar obstructiva crónica	1
fibrous histiocytoma	histiocitoma fibroso	2
herniated disk	hernia discal	2
hughes syndrome	síndrome de hughes	3
keratitis	queratitis	1
keratoconus	queratocono	1
krabbe disease	enfermedad de krabbe	1
langerhans-cell granulomatosis	histiocitosis de células de langerhans	4
motion sickness	mareo por movimiento	9
multiple personality disorder	desordern de personalidad múltiple	12
neuroendocrine tumors	tumores neuroendocrinos	2
pre-eclampsia	preeclampsia	1
quadriparesis	cuadriparesia	1
retinal detachment	desprendimiento de retina	3
talipes equinovarus	pie zambo	3
warts	verrugas	1
wilms tumor	tumor de wilms	3

Table 41: Examples of correct alignments

The evaluation of the overall performance of the algorithm in translating terms needs to take into consideration the current availability of

[46]http://www.hon.ch/ [accessed June 2010].

information for those terms in the web. Obviously, the algorithm will perform worse with terms that do not return many documents in an internet search engine. In a random sample of 100 names of diseases, it is to be expected that for around the half of them there will be insufficient information on the web because these diseases are very rare. There is an evident correlation between frequency of occurrence of a term in the web and the chances of finding a correct translation. This is unfortunate because the distribution of web pages per name of disease also obeys a Zipfean distribution. A positive side of the situation may be that the web is constantly growing and new information about diseases is also being added at a growing pace (including new rare diseases). In any case, when there is not enough information about a given disease, then a human translator will also have problems finding a translation. Because of this, it is difficult to determine what proportion of the failed trials are due to the experimental design when the algorithm relies on the assumption of the availability of information. One way to evaluate to evaluate the algorithm while taking this consideration into account is to divide the sample of 100 terms into subsets according to the number of hits that each term has in a search engine. In this way, Table 42 shows evaluation figures with the sample divided into three subsets. The first subset is made up of the half of the sample that obtained the largest number of hits in the search engine, a second subset with the same criterion using the top 75% of the sample in terms of hits, and finally, the whole sample.

Correct in rank	**50 most frequent**		**75 most frequent**		**Total (100)**
	cases	%	cases	%	cases
1	15	30	19	25	19
2	4	8	4	5	5
3	7	14	8	16	12
4	3	6	5	8	5
≤10	36	72	46	61	53
≤15	41	82	54	72	62

Table 42: Number of correct cases per rank

Precision is expressed according to the different positions in the ranking of candidates. Because of this, Table 42 shows the different rankings in rows. It must be read as follows: the first column indicates the position in the ranking of candidates where the correct candidate was found. The correct candidate was found in 30% (15 times) of the trials using the first subset of the most frequent terms, 25% (19 times) in the second subset and again 19% (19 times) in the total set. Taking into account the finding of a correct translation within the first 15 position in the ranking of candidates (a number considered reasonable taking into account the

cognitive effort of a potential user browsing the lists), the estimation of the precision of this algorithm would be 82% within the sample of the 50 most frequent terms, 72% within the 75 most frequent terms, and 62% within the whole sample.

In order to provide some point of reference for the evaluation of these results, the same sample of terms was submitted to three different baselines, whose figures of performance are provided in Table 43. The performance is measured as the proportion of terms of the sample that were correctly translated into the target language. The first of the baselines is the commercial translation software Lucy[47], which can be configured to translate technical terminology from English to Spanish. This program correctly translated 46% of the sample. The second baseline is a very simple term-translation script based on Wikipedia. It accepts a term in a source language, searches for a page with that title in Wikipedia, checks if there is a translation to the target language and returns the title of the translated page. The result of this simple script on the same sample is 48%. Google's translation service[48] produced much better results, translating correctly 88% of the sample, leaving the evaluated algorithm in second position with its 62% precision rate.

System	Precision on the 50 most frequent terms (%)	Precision on the 75 most frequent terms (%)	Precision on all terms (100) (%)
Lucy Software	48	50	46
Wikipedia	74	61	49
This Experiment	82	72	62
Google Translator	94	93	88

Table 43: Performance in comparison to other systems

Table 43 also shows how these numbers differ according to the number of hits in the search engine taking into account the division of the sample. As it was to be expected, the precision figures of the script based on Wikipedia as well as the Google translator improve. Lucy, in contrast, is not affected by the frequency of the terms it translates, a characteristic of the rule-based translation services.

It should be obvious, in any case, that the figures of comparison are merely for orientation. The idea has never been to demonstrate that the

[47]http://www.lucysoftware.de/cms/front_content.php?idcat=45&changelang=4 [accessed June 2010].
[48]http://www.google.es/language_tools?hl=en [accessed June 2010].

system based in this thesis can perform better than other systems, but the selected baselines provide an idea of where these results stand. Google's superiority is not at all surprising, considering the size of the corpus from which they trained the translator. While Lucy is a system based on dictionaries, Google is based on corpus, and it provides translations which do exist in its index. In comparison with other authors results, the performance of this thesis' algorithm does compare favorably. In a similar experiment, translating 2102 medical terms from English to Spanish using a language independent method, Langlais et al. (2008) reported 19.6% precision considering only the top translation candidate and 30.6% taking into account up the 25th position in the rank.

7.4.4 Discussion of the results of this Experiment

This experiment has shown that, on the basis of the hypotheses presented in the thesis, it is possible to find, for a given input term in one language, candidates of term equivalent in another language. It is obvious that several weaknesses of this experiment can be ameliorated (as it will be explained in 7.4.5) if this methodology is to be applied in a real practical situation. The figures of precision are surely not high enough for a final user, but the results may save time and effort in the case of a lexicographic project or may be even useful for a translator or a terminologist when dealing with terminology that is not yet registered in dictionaries.

A fair comparison with the results of other authors is difficult because, apparently, no attempt has been made at the extraction of specialized bilingual terminology by using the web instead of a parallel corpus. The attempts made so far with the extraction of bilingual lexicon from non-parallel corpora, as those commented at the beginning of this experiment, have focused mainly on general vocabulary and most of the times with single word entries, as in the case of Fung (1997) or Rapp (1999). Experiments with specialized terminology are much more challenging for many reasons, mainly the poor availability of raw material and also the problems of segmentation due to the fact that a great number of the entries are multiword terms. Langlais et al. (2008) work with terminology of the same domain (medicine) and language-pair (English-Spanish, among others), but these experiments are not comparable because they obtain the translation candidates from ontologies instead of attempting to extract them from a large corpus like the web, which is a more realistic scenario.

As a final conclusion on this experiment, it is also important to remark the fact that the subject of research is also interesting from a purely theoretical perspective, because it uncovers macroscopic and structural regularities of language that are visible at this moment of history because

of the massive amount of data from the web. The web is a chaotic corpus, the result of a great number of individual linguistic choices made by different authors in different languages and disciplines. Within this unorganized social behavior there are measurable regularities, and one of these emergent properties, as it has been shown in this experiment, is the statistical association of equivalent terms in different languages.

7.4.5 Future Work for this Experiment

There are many possibilities to improve this method of finding candidates for bilingual terminology equivalences described in these sections. One such possibility, which is conceptually simple but computationally costly, is to eliminate false candidates by iterating the process in the opposite direction. That is to say, to repeat the process with each of the equivalent candidates as input to find their translation in what was originally the source language. It is to be expected that the correct translation will have the original term among the equivalent candidates, this time in the original source language.

Another interesting idea would be to try using hybrid methods. These can be the simple ones, like part of speech (POS) tagging or others of greater complexity such as syntactic analysis. These alternative sources of information can be exploited in order to avoid proposing a candidate which is morphologically or syntactically incompatible with the query term, at least in the case of specialized terminology where, as stated before, multiword terms are very frequent. With a set of grammar rules, it is possible to parse most of the noun phrases in Spanish. A few rules can parse phrases and give us, as a result, trees that can be expressed in brackets, such as **SN[H[N5-FS] JQ—FS]**. The keys are SN for noun phrase, SP for prepositional phrase, *H* for head, *N5-FS* for common noun, feminine singular, *JQ-FS* for qualifying adjective, feminine singular, etc. Table 44 shows some examples of terms chunked with a hand made grammar of fifteen rules.

1) **amencia nevoide**
N5-FS JQ--FS
SN[H[N5-FS] JQ--FS]

2) **factor de alivio**
N5-MS P N5-MS
SN[H[N5-MS] SP[P N5-MS]]

3) **reflejo de defensa**
N5-MS P N5-6S
SN[H[N5-MS] SP[P N5-6S]]

4) **vitamina k2**
N5-FS N5-MS
SN[H[N5-FS] N5-MS]

5) **corazón de atleta**
N5-MS P N5-6S
SN[H[N5-MS] SP[P N5-6S]]

Table 44: Examples of morphological and syntactic analysis.

It is even possible to generate the rules automatically with the aid of corpus statistics and co-occurrence graphs, in a completely unsupervised manner. In this case, the parser would chunk sequences of POS tags that are known to co-occur frequently, such as Noun + Adj or Determinant + Noun, etc. With this procedure it would be possible to find out which are the typical syntactic patterns of the multiword terms in a given language.

Yet another possibility for the automatic acquisition of syntactic knowledge of both source and target language would be with supervised learning, using a bilingual dictionary as a training set. This is, it is possible to use the dictionary to apply a tagger both to the English and the Spanish entries and see which is the aligned pair of syntactic trees that co-occurs most frequently in both languages. For instance, an English term such as *histoid neoplasm,* with its syntactic pattern Adj + Noun, is translated as *neoplasia histoide*, which has the pattern Noun + Adj. Since most of the times an English term with pattern Adj + Noun will have as Spanish translation a term with the pattern Noun + Adj, a computer system can statistically learn to associate those syntactic patterns as equivalents. This would give us the syntactic equivalence between both languages, which means that one could know in advance what the syntactic tree of an ideal translation candidate should look like. Therefore, for any query term, which has a given syntactic pattern, the algorithm will prefer a candidate which has the corresponding target language syntactic pattern.

To rephrase the previous paragraph, syntactic knowledge would be beneficial not only because it would help to eliminate false candidates, but also because it would help to relate a candidate with a query term, or to

make even more refined inferential assertions. For instance, given a query term such as *hemophilia A*, several instances of candidates such as *patients with hemophilia A* and *pacientes con hemofilia A* may be encountered. At this point, the algorithm can only recognize the first candidate as an English expression in order to eliminate it. However, it is missing the information that could be obtained from the syntactic equivalence of these two phrases, inferring the equivalence between *hemophilia A* and *hemofilia A,* which would be later reinforced with other coefficients such as the one based on orthographic similarity. Furthermore, provided that *enfermedad* and *disease* are equivalents, then it is to be expected that *enfermedad de Refsum* and *Refsum disease* are equivalents too, which would be another level of reasoning beyond frequency analysis.

Chapter 8: General Conclusions

This thesis has presented a statistical approach to the analysis of concepts by means of the study of the distributional behavior of terms in large samples of discourse. More specifically, on the correlation that exists between the statistical association of terms and their relation in the conceptual plane. Statistical techniques, such as graphs of term co-occurrence and vector space models, among others, can help us to uncover implicit conceptual information lying beneath the distribution of terms in corpus. This thesis shows that, using these techniques, it is possible to develop an algorithm that performs a certain degree of conceptual analysis taking only raw text as input. The model has been applied to solve a variety of practical problems and, in doing so, no recourse to human criterion or sources of pre-built knowledge from domains or languages has been taken. This analysis has been done purely on the basis of the mathematical properties of the graphs, that is, not taking into account any information on the specific language that is analyzed or ontological information on the topic of the analyzed discourse. The thesis presented an empirical demonstration of its hypotheses, but it is also explanatory, as it has tried to find a linguistic interpretation of the observed phenomena, which is that the properties of the co-occurrence graphs are possible because authors of argumentative texts have a tendency to name some of the basic properties of the concepts that they introduce in discourse. Apart from being empirical and explanatory, the thesis also offers a predictive model because it can be extrapolated to other languages and domains.

The thesis has been structured around two hypotheses to explain the observable statistical behavior of term co-occurrence making conceptual structures emerge naturally from discourse. The first hypothesis implied that analytical statements are also prominent with respect to synthetical statements. In absolute numbers, there are more synthetical statements in discourse, however they are much more diverse. Analytical statements tend to be repeated. Independent of their particular wording, and this could explain why humans remember the properties and conceptual features of concepts. The second hypothesis is that terms that refer to specific concepts are necessarily "conceptually related" to other terms in a Saussurean solidarity with the rest of the system. The rest of the units of the vocabulary, which have a predicative instead of a referential function, engage in relations with such a variety of contexts that no strong relations with specific concepts grow.

The background necessary for the thesis is rather complex. The reflection of the "concept of concept" has a long standing tradition and a certain degree of historical account has been necessary to place the discussion of concept analysis in full perspective. The source of ideas on this matter has

ranged from ancient philosophy to modern day semantics, including a minimal contribution from psychological and psycholinguistic advances. With respect to the studies that deal more specifically with the automatic analysis of concepts, the literature reviewed in the thesis was presented in two main blocks. The first one is devoted to the strategies of term and conceptual relations extraction from corpus using "symbolic rules", that is, systems based on explicit coding. The prototypical example of this strategy is the extraction with lexico-syntactic patterns (*X is a type of Y*) and even the employment of deeper language and domain specific knowledge. The second block includes the set of strategies that involve the statistical and "knowledge poor" approach with authors that use graph and vector space models to solve a variety of problems of automatic concept analysis.

As empirical evidence, the thesis has shown examples of application in a variety of operations that a human would perform using intellectual skills. The methodology and experimental evaluation of the model presented in the thesis consisted basically in the test of a single algorithm of term co-occurrence analysis on four basic operations of conceptual analysis: 1) The ability to recognize when the text refers to an entity of shared knowledge. Terms that refer to specific concepts or entities in texts can be highlighted thanks to the geometrical properties of the co-occurrence graphs that they generate. 2) The ability to classify these entities in taxonomic structures. Terms that refer to entities can be clustered in taxonomic structures thanks to the asymmetric relation that derives from the vocabulary overlap of their graphs. 3) The ability to distinguish when a term refers to more than one concept and to determine which is the concept that is referred to in every specific context. Terms that refer to different concepts can be disambiguated in contexts thanks to the different "hubs" of nodes of their graph. 4) The ability to recognize when different terms refer to the same concept. Different terms that designate the same concept (as it is the case of equivalents in different languages) can be associated thanks to the symmetric relation that is detected by the vocabulary overlap of their graphs.

As the conclusion, this thesis proposes a research in a theoretical plane, because it intends to study the mechanisms of language and discourse which encodes conceptual structuring. This structuring emerges naturally as the result of the interaction of a linguistic community. The claims that the thesis presents are sustained by a considerable amount of empirical evidence, conforming, as said above, a descriptive, explicative and predictive theorization. Thanks to its predictive power, the thesis also promises a variety of possibilities of application in fields such as terminology extraction, taxonomy extraction, term disambiguation and bilingual terminology extraction, among others. In the face of these results, it is certainly striking to compare the general proposal of this

thesis with statements of Eco (2004), who apparently is not acquainted with the methods of computational linguistics and does not seem to be ready to agree with the hypothesis presented in this thesis, even when this thesis is much in debt to his ideas. He points out that Internet does not provide the kind of mediations that previous media offered. These mediations function as a filter, and readers need them. Otherwise, they cannot distinguish relevant information from nonsense. He thinks one should know when Julius Caesar died, but the day his wife died is unimportant. Since in the Internet this valuation is waived, the Internet is a chaotic medium from which no knowledge can be extracted and the risk is that everyone will have an individual encyclopedia. His comment shows an understanding of the nature of language that is very different to the one exposed in this thesis. This thesis is intended to show order where Eco sees only chaos. And this is possible because, as it is hypothesized in this study, inside the most chaotic collection of documents, analytical propositions are salient.

The mechanism underlying the equilibrium between analytical and synthetical statements is undoubtedly rhetorical: topics in general are introduced in discourse. When a new actant is presented in discourse, it is usually "padded with redundancy" (Bougnoux, 1993). This is nothing more than the exordium of ancient rhetoric, where the speaker, normally at the beginning of the exposition, tries to be accommodating to the audience, showing respect for the topics that are already established in the community and that are supposed to be shared among participants of the interaction (Vignoux, 1986). By this same regulatory principle, every time that a speaker introduces a new term or topic, this term is defined according to the meaning it has in the community that is being addressed. These principles make up a semantic system that is in permanent fluctuation and evolution. With the aid of corpus statistics, it is now possible to make a temporal cut in the different moments of this system. Metaphorically, it would be like taking a snapshot of the concept denoted by the term.

On the one hand, as it was already stated in the introduction, this work entails an interesting line of research from an engineering point of view. On the other hand, it is also a purely theoretical contribution, since this thesis deals with structural properties of language and the behavior of its vocabulary, which transcends the study of the terminology of a specific field or language.

Chapter 9: Prospective Outlooks

Many potential applications remain unexplored in this thesis, but they are now open as futures lines of research. This chapter is divided in two main points. The first one describes what it would a possibility for the extension of the proposal with a symbolic approach, using full morphological, syntactic and semantic parsing. The second point delves into what would be different lines of practical application, some of which were already evaluated. Evaluating those ideas in real life cases is now the most immediate prospective.

9.1 Extensions of the Approach

The first and most important pending work at the moment is a deeper semantic analysis of the relation between the terms that are found in co-occurrence. Extraction by quantitative means of certain relations between terms, such as hypernymy or synonymy, have already been evaluated in the research without involving explicit knowledge about a particular language. However, there is still much to be done with the extraction of conceptual relations between terms by adopting a language-specific perspective.

This thesis has shown how it is possible to separate those units that refer to concepts from the units that function as predicates, and by doing so, to represent concepts and relations between concepts. As an example, consider a small graph in which there are two nodes, one for *keratitis* and the other for *contact lens,* with a connection between them labeled with the predicate *associated with,* which is the predicate that has been found most frequently between these two nodes.

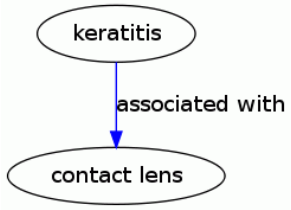

Figure 46: Labeling of the conceptual relations between nodes.

An important limitation at this moment is that the algorithm has no internal representation of what it means to be *associated with*, which in this case has no consequences since this is a symmetric type of predicate. However, considering an asymmetric predicate, such as the relation of cause, the situation is quite different.

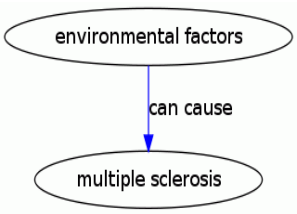

Figure 47: Asymmetrical relation.

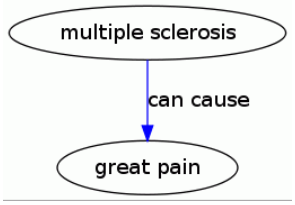

Figure 48: Another case of asymmetrical relation.

Consider, for illustration, Figures 47 and 48. At this point, the algorithms presented in this research are not able to tell the difference between the roles that the argument *Multiple sclerosis* has in relation to the predicate *cause* in these two examples. In one, it is the cause, and in the other, the consequence. An internal representation that could state this difference explicitly could be applied to an automatic question answering system, although this would be a different line of research. Thus, the next step of this thesis would be a combination of statistics and language-dependent strategies. This is because, knowing which are the key concepts in the meaning of a term, it would be possible then to extract and parse the contexts where the terms that designate these key concepts are engaged and then select the most frequent relation. The reason for this is, again, that analytical statements about a given term are different on the surface but similar with respect to the information they transmit.

Finally, as was already stated in Chapter 2, another possibility of extending this model is to implement a system based on feedback. This would be a system that learns from its own errors. In this case, the figure of a user arise (individual or collective), because the user would be the one who would *teach* the system, to avoid using the word *train*, which has already acquired a specific technical sense in the computational linguistics community.

9.2 Possibilities for Application

Information Retrieval: An implementation based on this thesis is intended to overcome a basic assumption in most studies in a field such as Information Retrieval (IR): that the response of a search engine is a list of documents ranked by degree of relevance to the user's query (Salton & McGill, 1983; Baeza-Yates & Ribeiro-Neto, 1999; Manning et al. 2008). A list of documents means that the user is confronted with a one-dimensional object and is therefore expected to scan this list in the search of a relevant document. Users may feel familiar with this procedure not only because of the popularity of web search engines but also because

language, as a serial code, is one-dimensional too. This paradigm in the fieldwork of IR is well established and has a great community of users, although other alternatives and other uses are still possible, such as a structure of terms in two or three dimensions.

While text is linear, this cannot be said about its macro-structure. It is possible to re-arrange the textual collection in order to make it look like a topography that represents an organization of the topics covered by the documents. Thus, the idea is to decompose the linearly encoded text into more dimensions, the space of a conceptual structure. In cartographic maps, the different distances between the spots defined by analogy those that existed in the territory. In the case of a concept map, spots are signaled by significant terms found in a document collection. This thesis has shown how conceptually related terms emerge naturally from a search query and how groups of documents are formed according to their content, all things that might be highly beneficial for information retrieval systems. An example of application of this thesis for information retrieval could be a search engine which, instead of returning a list of results for a query, offers a graph of arcs and nodes with the terminology related to the concept or concepts that are referred to by that query. With the terms linking their contexts of occurrence, a user could quickly organize the relevant documents of a collection without the need to go through the list of URL references and snippets.

Terminology Extraction: Besides the main application of the thesis to information retrieval, there are other interesting possibilities. One example would be to apply these algorithm to the task of terminology extraction. A task that is related to this particular problem has been shown in the second experiment of Chapter 7. In that case, the extraction is done without resort to any kind of knowledge source, such as specialized dictionaries or lexical databases, but using a simple web search engine. The identification is to be carried only by means of the extraction of the lexical cohort of each of the lexical units of a document under analysis.

This method may, however, encounter problems in the case of less obvious examples, particularly with highly polysemous terms. Consider, for instance, the case of the term *object* in a document on semiotics. The probabilities of getting a document on semiotics in a web search using just the word *object* as the query are, at the very least, negligible. It should be possible, however, to combine the query *object* with some other term from the document whose technical status is more certain. To continue with same example, the term *representamen*, that may appear near *object* in the document on semiotics, would increase the chances of retrieving documents on the same subject. This is, however, a claim that needs to be grounded in further empirical evidence and is, indeed, an area of active research (Nazar et al., 2007).

Term Translation: Translators are often confronted with scientific terminology. The translator can use terminological resources to find a term in a language and its equivalent in a different language. However, new domains are permanently appearing and evolving and there is continuous coinage of terminology. Dictionaries cannot keep pace with this rhythm, so automatic and accurate procedures will be welcome. This thesis has shown how it is possible to apply co-occurrence graphs in bilingual acquisition from monolingual corpora. Other experiments along this line have already been reported in Nazar et al. (2007; 2008) and Nazar (2008b).

Dialectology: Dialectological studies, with the automatic characterization of the lexical and syntactical differences in variants of the same language, are another active line of research. This line of application is intrinsically related to the models of extraction of equivalent terms in different languages, because the same can be done now with different variants of the same language. Spanish is, for instance, a language that is geographically widespread and this has given place to great variation that is expressed in the lexicon, apart from phonology and syntax. There are many variations of Spanish even at the register of the media, especially with respect to vocabulary. The same referent receives different names in one country and another. Thus, if a Spaniard wants to know how to say *aguacate* (avocado) in the Spanish variant of Argentina or Chile, he or she could use this method to know that the correct form is *palta*. The case of biological species is particularly straightforward since the most strongly associated node in the network is usually the scientific name, *persea americana* in the case of the avocado. A second search with this last term will yield *palta* as a significant collocate. In the case of this example, perhaps it would be easier to look the word up in an encyclopedia, but the example is only intended to show how the same method could be used for less obvious examples, still not listed in dictionaries or encyclopedias. The web currently offers enough data about the distinct varieties of Spanish (as well as other languages) to begin such an exploration. This is the subject of ongoing research.

Opinion Mining and other Marketing Strategies: Another application that could be considered is in the field of marketing and advertisement. It could offer a quick way to know what people are thinking about a product -or a politician-, what opinions they are giving in blogs, forums and web pages, and how this concept is evolving in time (for example, before and after advertisement campaigns).

Companies also need to be careful when selecting the names of their products, a practice called *naming* or *branding*. This is especially important when companies want to introduce a product in a foreign

culture. Apart from the possibility of using a name which is foreign to their phonological system and making it therefore difficult to pronounce, one of the biggest risks is the possibility of using a name that has negative connotations in such culture.

In the same scope of application but from another perspective, there is the possibility of the exploitation of ambiguity for creative purposes. Ambiguity is one of the most common strategies used in advertisement. It would also be useful to extract playful word associations with this method. In this way, it would be possible to find senses of words that one may not be able to find by intuition, and to use these senses in the creative process of an advertisement campaign. The methodology here would be, given as input a list of terms related to the concept of the campaign, the product or brand, to obtain as output, from the whole vocabulary of a language, units that are related to the concept in question and can also be ambiguous in certain contexts.

From a somewhat similar perspective, different studies of ideology could be undertaken with the methods described in this research. A sociologist, for instance, or discourse analyst, could find this method useful to study how people have formed opinions on certain topics. Say, for instance, how people in Europe have changed their concept of an *immigrant* in the last ten years.

Semantic-Neologism Extraction: Diachronic studies, such as the automatic extraction of semantic neology, is another active line of research that will give continuity to this thesis. The detection of semantic neologisms, is a challenge to general neology, because in this case the neologism has the same form of a unit that is already collected in databases. However, the distributional perspective offers an alternative for the solution to this problem. Having an extensive and chronologically ordered corpus of a language, it would be technically feasible to advert abrupt differences in the frequencies of certain units of the lexicon in combination with a shift in their contexts. Promising results have been reported in this line of research (Nazar & Vidal, 2010).

Automatic Text Summarization: Another interesting line of research that the distinction between analytical and synthetical statements offers, is its application to automatic summarization. What is important to keep in a summary is not precisely the analytical statements, which transmit a knowledge that is already established. What is wanted in a summary are the synthetical statements, because they are the ones that bring new information.

References

AGUADO DE CEA, G.; ÁLVAREZ DE MON, I. & PAREJA LORA, A. (2002). "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag". Revista Iberoamericana de Inteligencia Artificial, no. 17, pp. 37–49.

AITCHINSON, J. (1987). "Words in the Mind, an Introduction to the Mental Lexicon". Oxford Basil Blackwell.

AIZAWA, A & KAGEURA, K. (2001). "A graph-based approach to the automatic generation of multilingual keyword clusters", in Bourigault, D., Jacquemin, C. and L'Homme, M-C. (eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins. pp. 1–28, 2001.

ALCINA, M. A. (1994). "Algoritmo para la resolución del grado de especificidad de las expresiones referenciales", in Procesamiento del Lenguaje Natural, Num. 14, pp. 79–90.

ALDENDERFER, M. S. & BLASHFIELD, R. K. (1984). "Cluster analysis". Beverly Hills, Sage.

ALSHAWI, H. (1989). "Computational lexicography for natural language processing", in *Analysing the dictionary definitions*, pp. 153–169. Longman Publishing Group, White Plains, NY, USA.

AMSLER, R. (1981). "A taxonomy for English nouns and verbs", in Proceedings of the 19th annual meeting on Association for Computational Linguistics, pp. 133–138, Morristown, NJ, USA.

ANANIADOU, S. (1994). "A Methodology for Automatic Term Recognition". *Coling 1994, 15th International Conference on Computational Linguistics*, Kyoto, Japan, 1994.

ANDERSON, J. R. (1983). "The architecture of cognition". Harvard. University Press.

ARISTOTLE (c. 350 b.c.). "Categories".

ARNHEIM, R. (1954/1979). "Arte y percepción visual", Madrid, Alianza.

ARTSTEIN, R & POESIO, M. (2008). "Inter-Coder Agreement for Computational Linguistics". Computational Linguistics. December 2008, vol. 34, no. 4, pp. 555–596.

ATTAR, R. & FRAENKEL, A.S. (1977). "Local feedback in full text retrieval systems". Journal of ACM, vol. 20, no. 3, pp. 397–417, July.

AUGER, A. & BARRIÈRE, C. (Eds.) (2008). "Pattern-based Approaches to Semantic Relation Extraction Special issue of Terminology", vol. 14, no. 1, (2008). Defense Research and Development Canada (Valcartier) / National Research Council of Canada.

BAEZA-YATES, R. & RIBEIRO-NETO, B. (1999). "Modern Information Retrieval", ACM Press, Essex.

BAEZA-YATES, R. & CASTILLO C. (2001). "Analysis of Link-Based Ranking for the Web". Technical report, University of Chile.

BARONI, M. & LENCI, A. (2008). "Concepts and properties in word spaces". In Italian Journal of Linguistics . Rivista di linguistica, vol. 20, no. 1, 2008, pp. 55–88.
BARRIÈRE, C. & POPOWICH, F. (1996). "Building a noun taxonomy from a children's dictionary", in Proceedings of Euralex'96, pp. 65–70.
BATESON G. (1972). "Steps To An Ecology of Mind". Balantine: New York.
BEKINSCHTEIN, T.; SHALOM, D.; FORCATO C.; HERRERA, M.; COLEMAN, M.; MANES, F. & SIGMAN, M. (2009). "Classical conditioning in the vegetative and minimally conscious state". Nature Neuroscience, vol. 12, pp. 1343–1349.
BELINCHON, M.; IGOA, J.M. & RIVIERE, A. (1992). "Psicología del Lenguaje. Investigación y teoría". Madrid. Trotta.
BENVENISTE, É. (1966). "Problèmes de linguistique générale", Paris: Gallimard.
BERTELS, A.; SPEELMAN D. & GEERAERTS, D. (2007). "Les corpus spécialisés au service de la sémantique quantitative : la polysémie du français technique", in G. Williams (ed.), *Revue électronique Texte et Corpus*. Actes des 4èmes Journées de Linguistique de Corpus 2005 à Lorient.
BIRRELL, A.; MCJONES, P; LANG, R. & GOATLEY, H (1995). "Pstotext - extract ASCII text from a PostScript or PDF file. README file". Online Document. http://pages.cs.wisc.edu/~ghost/doc/pstotext.htm [accessed June 2010].
BOGURAEV, B. (1991). "Building a lexicon: The contribution of computers." International Journal of Lexicography, vol. 4, no. 3, pp. 227–260.
BÖHM, K.; MAICHER, L.; WITSCHEL, H. & CARRADORI, A. (2004). "Moving Topic Maps to Mainstream - Integration of Topic Map Generation in the Users' Working Environment", in J.UCS - Journal of Universal Computer Science (Springer), vol. 10, Special Issue I-Know 2004, pp. 241–251.
BORGES, J.L. (1944/1962). "Ficciones", New York, Grove Press.
BOSQUE, I. (1999). "El nombre común"; in Bosque, I.; Demonte, V. (eds.) *Gramática descriptiva de la lengua española*, Cap. 1. Madrid. Espasa-Calpe.
BOUFADEN, N.; LAPALME, G & BENGIO, Y. (2004). "Extended Semantic Tagging for Entity Extraction". Beyond Named Entity Recognition Semantic labeling for NLP tasks Workshop held Jointly with LREC 2004, Lisbon, Portugal, May 2004, pp. 49–54.
BOUGNOUX, D. (1993). "Sciences de l'information et de la communication". Paris: Larousse.

BRILL, E. (1992). “A simple rule-based part-of-speech tagger”, in Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing, Trento, IT, 1992, pp. 152–155.
BRISCOE, T. (2001). “From dictionary to corpus to self-organizing dictionary: Learning valency associations in the face of variation and change”, in Proceedings of Corpus Linguistics 2001, Lancaster University, pp. 79–89.
BROWN, P.F.; COCKE, J.; DELLA PIETRA, S.A.; DELLA PIETRA, V.J.; JELINEK, F.; LAFFERTY, J.D.; MERCER, R.L. & ROOSIN, P. (1990). “A Statistical Approach to Machine Translation”. Computational Linguistics, vol. 16, pp. 79–85.
BUITELAAR, P.; CIMIANO, P. & MAGNINI, B. (2005). “Ontology Learning from Text: An Overview”, in Buitelaar, Cimiano & Magnini (Eds.), *Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press, pp. 3–12.
BUNGE, M. (1960). “La ciencia, su método y su filosofía”, Buenos Aires, Eudeba.
BUNGE, M. (1974). “Semantics I: Sense and Reference”. Treatise on Basic Philosophy, v.I. Boston. Reidel.
BYBEE, J.L. & HOPPER P.J. (eds.). (1997). “Frequency and the Emergence of Linguistic Structure”. Amsterdam, Philadelphia: John Benjamins.
CABRÉ, T. (1992). “La Terminología: Teoría, metodología, aplicaciones”. Barcelona, Antártida / Empúries.
CABRÉ, T. (1999). “La Terminología: Representación y Comunicación”. Barcelona, Institut Universitari de Lingüística Aplicada.
CABRÉ, T. & ESTOPÁ, R. (2005). “Unidades de conocimiento especializado, caracterización y topología”, in Cabré, M. T.; Bach, C. (eds.). *Coneixement, llenguatge i discurs especialitzat.* pp. 69–94. Barcelona. Institut Universitari de Lingüística Aplicada.
CABRÉ, M. T.; Estopà, R. & Vivaldi, J. (2001). “Automatic Term Detection: A review of Current Systems”. In Bourigault, D., Jacquemin, C. and L'Homme, M-C. (eds.) Recent Advances in Computational Terminology. Amsterdam: John Benjamins, pp. 1–28, 2001.
CABRÉ, M. T.; BACH, C. & VIVALDI, J. (2006). “10 anys del Corpus de l'IULA”. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.
CAIRNS, W. (2007). “About the Size of It: The Common Sense Approach to Measuring Things”, London, Macmillan.
CALZOLARI, N. (1977). “An empirical approach to circularity in dictionary definitions”. Cahiers de Lexicologie, vol. 31, no. 2, pp. 118–128.
CARABALLO, S. (1999). “Automatic construction of a hypernym-labeled noun hierarchy from text”, in Proceedings of the 37th

Annual Meeting of the Association for Computational Linguistics, pp. 120–126.

CHANG, J.; KER, S. & CHEN, M. (1998). "Taxonomy and lexical semantics - from the perspective of machine readable dictionaries", in AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, London, UK. Springer-Verlag, pp. 199–212.

CHARNIAK, E. (1993). "Statistical language learning", Cambridge (Mass.) MIT Press cop.

CHEN, H.; SCHATZ, B.; NG, T.; MARTINEZ, J.; KIRCHHOFF, A. & LIN, C. (1996). "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project". IEEE Transactions on Pattern Analysis and Machine Intelligence. August 1996, vol. 18, no. 8, pp. 771–782.

CHODOROW, M.; BYRD, R. & HEIDORN, G. (1985). "Extracting semantic hierarchies from a large on-line dictionary", in Proceedings of the 23rd annual meeting on Association for Computational Linguistics, July 08-12, 1985, Chicago, Illinois, pp. 299–304.

CHOMSKY, N. (1957). "Syntactic Structures". The Hague, Mouton.

CHOMSKY, N. (1959). "A Review of B. F. Skinner's Verbal Behavior" in Language, vol. 35, no. 1, pp. 26–58.

CHOMSKY, N. (1965). "Aspects of the Theory of Syntax". Cambridge, MIT Press.

CHRIST, O.; SCHULZE, B.; HOFMANN, A. & KÖNIG, E. (1999). "The IMS Corpus Workbench: Corpus Query Processor (CQP) - User's Manual". Institute for Natural Language Processing, University of Stuttgart.

CHURCH, K. & HANKS, P. (1991). "Word Association Norms, Mutual Information and Lexicography", Computational Linguistics, vol 16, no. 1, pp. 22–29.

CIGARRÁN, J.; GONZALO, J.; PEÑAS, A. & VERDEJO, F. (2004). "Browsing search results via Formal Concept Analysis: Automatic selection of Attributes", in Proceedings of the Second International Conference on Formal Concept Analysis, ICFCA (ICFCA 2004). Lecture Notes in Computer Science. Springer-Verlag, pp. 74–87.

CIGARRÁN, J.; PEÑAS, A; GONZALO, J. & VERDEJO, F. (2005). "Automatic selection of noun phrases as document descriptors in an FCA-based Information Retrieval system", in Proceedings of the International Conference on Formal Concept Analysis (ICFCA 2005). Lecture Notes in Computer Science. Springer-Verlag, vol. 3403, pp. 49–63.

COLLINS, A. & LOFTUS, E. (1975). "A spreading-activation theory of semantic processing", Psychological Review. Nov, vol. 82, no. 6, pp. 407–428.
COSERIU, E. (1967). "Teoría del lenguaje y lingüística general", Madrid, Gredos.
CRUSE, D. A. (1986). "Lexical semantics". Cambridge University Press.
CUENCA, Mª J. & Hilferty, J. (1999). "Introducción a la lingüística cognitiva", Barcelona, Ariel.
CUENCA, J.M. (2000). "La lingüística cognitiva com a paradigma emergent", in M. T. Cabré, C. Gelpí (ed.). *Cicle de conferències i seminaris 97-98: lèxic, corpus i diccionaris.* Barcelona, IULA.
CURRAN, J. (2004). "From Distributional to Semantic Similarity". PhD thesis, University of Edinburgh.
DAILLE, B. (1994). "Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques". Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.
DALE, R. (1992). "Generating Referring Expressions". MIT Press. London.
DAVIDSON, L. (1997). "Knowledge extraction technology for terminology". Master's thesis, School of Translation and Interpretation, University of Ottawa.
DE YZAGUIRRE, Ll.; RIBAS, M.; VIVALDI, J. & CABRÉ, M.T. (2000). "Alineación automática de traducciones: descripción y usos en los ámbitos de la profesión, de la docencia y de la investigación traductológica", in Proceedings of IV Encuentros Alcalaínos de Traducción, Universitat d'Alcalà d'Henares, 17-18 de febrer de 2000.
DODGE, M. (2005). "An atlas of cyberspaces". Online Document. http://www.cybergeography.org/atlas/topology.html [accessed June 2010]
DOLAN, W.; VANDERWENDE, L. & RICHARDSON, S. (1993). "Automatically deriving structured knowledge bases from on-line dictionaries", in Proceedings of the First Conference of the Pacific Association for Computational Linguistics (Vancouver, Canada), pp. 5–14.
DROUIN, P. (2003). "Term extraction using non-technical corpora as a point of leverage". Terminology, vol. 9, no. 1, pp. 99-117.
ECO, U. (1979/1981). "Lector in fabula la cooperación interpretativa en el texto narrativo", Barcelona, Lumen.
ECO, U. (1968/1989). "La Estructura ausente: introducción a la semiótica". Barcelona, Lumen.
ECO, U. (2004). "Bitte, sagen Sie uns die Wahrheit, Umberto Eco über den Islam, das Internet, seinen neuen Roman und die Rolle des Intellektuellen".
EVERT, S. (2004). "The Statistics of Word co-occurrences"; PhD Thesis; IMS; University of Stuttgart.

FELDMAN, J. (2006). “From Molecules to Methafors”. Cambridge. MIT Press.

FELIU, J. (2004). “Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica”. PhD Thesis; IULA; Universitat Pompeu Fabra.

FELIU, J.; VIVALDI, J & CABRÉ, T. (2006). “SKELETON: Specialised knowledge retrieval on the basis of terms and conceptual relations”, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Paris: European Language Resources Association. pp. 2377–2382.

FELLBAUM, C. (1998). “WordNet: An Electronic Lexical Database”. MIT Press.

FIRTH, J. R. (1957/1968). “Selected papers of J.R. Firth, 1952-59”; edited by F.R. Palmer, London, Longmans.

FOX, E.; NUTTER, J.T.; AHLSWEDE, T.; EVENS, M. & MARKOWITZ, J. (1988). “Building a Large Thesaurus for Information Retrieval”, in Proceedings of the ACL Conference on Applied Natural Language Processing, February, 1988, pp. 101–108.

FREGE, G. (1884/1953). “The Foundations of Arithmetic”, Oxford Basil Blackwell.

FREGE, G. (1892/1993). “On sense and reference”, in A.W. Moore (ed.) Meaning and Reference. Oxford: Oxford University Press.

FREUD, S. (1913/1986). “Totem y Tabú”, Obras Completas, Tomo 13, Bs. Aires, Amorrortu.

FUNG. P. (1995). “Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus”, in Proceedings of the Third Workshop on Very Large Corpora, pp. 173–183.

FUNG, P. (1998). “A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora”, in Proceedings. of the Third Conference of the Association for Machine Translation in the Americas, pp. 1–16.

FUNG, P. & MCKEOWN, K. (1997). “Finding Terminology Translations From Non-Parallel Corpora”, in Proceedings of the 5th Annual Workshop on Very Large Corpora, Hong Kong, Aug. 1997, pp. 192–202.

GAINES, B.R. & SHAW, M.L.G. (1995). “WebMap: Concept mapping on the Web”, in Proceedings of the Fourth International World Wide Web Conference, December 11-14, 1995, Boston, Massachusetts, USA. World Wide Web Journal 1, December 1995.

GAIZAUSKAS, R. & WILKS, Y. (1988). “Information Extraction: Beyond Document Retrieval”, Computational Linguistics and Chinese Language Processing vol. 3, no. 2, August 1998, pp. 17–60.

GALE, W. & CHURCH, K. (1991). "Identifying word correspondences in parallel texts", in Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 152–157.
GANTER, B. & WILLE, R. (1999). "Formal Concept Analysis. Mathematical, Foundations". Berlin, Springer.
GÄRDENFORS, P. & WILLIAMS, M. (2001). "Reasoning about Categories in Conceptual Spaces", in Proceedings of the 17th international joint conference on Artificial intelligence vol. 1 Seattle, WA, USA, pp. 385–392.
GIRALDO ORTIZ, J. J. (2008). "Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente". PhD Thesis, IULA; Universitat Pompeu Fabra.
GIRJU, R. (2002). "Text Mining for semantic relations", PhD Thesis, University of Texas.
GLADKIJ, A. V. & MELCUK, Y. (1972). "Introducción a la lingüística matemática". Barcelona. Planeta.
GODBY, C.; MILLER, E. & REIGHART, R. (1999). "Automatically Generated Topic Maps of World Wide Web Resources". Annual Review of OCLC Research.
GRANITZER, M.; AUGUSTIN, A.; KIENREICH, W. & SABOL, V. (2009). "Taxonomy extraction from german encyclopedic texts", in Proceedings of the Malaysian Joint Conference on Artificial Intelligence 2009, Kuala Lumpur, Malaysia.
GREFENSTETTE, G. (1994). "Explorations in Automatic Thesaurus Discovery", Kluwer Academic Publishers, Norwell, MA.
GREIMAS, A. J. (1966). "Sémantique structurale", Paris, Larousse.
GRIES, S.T. & DIVJAK, D. S. (forthcoming). "Behavioral profiles: a corpus-based approach towards cognitive semantic analysis", in Evans, Vyvyan and Stephanie S. Pourcel (eds.), *New directions in cognitive linguistics*. Amsterdam, Philadelphia, John Benjamins.
GROSSBERG, S. (1986). "The adaptive self-organization of serial order in behavior: Speech, language, and motor control", in E.C. Schwab and H.C. Nusbaum (Eds.) *Pattern Recognition by Humans and Machines*, vol. 1: Speech Perception, pp. 187–294. New York: Academic Press.
GRUBER, T. (2008). "Ontology". To appear in the Encyclopedia of Database Systems, Ling Liu & M. Tamer Özsu (Eds.), Springer-Verlag.
GUIHONG, C.; JIANFENG G. & JIAN-YUN N. (2007). "A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web Pages", in the Proceedings of MT Summit XI.
HA, L., MITKOV, R. & CORPAS PASTOR, G. (2008). "Mutual terminology extraction using a statistical framework". *Procesamiento del lenguaje natural*, no. 41 (sept. 2008), pp. 107–112.

HALLIDAY, M.A.K. (1978). “Language as Social Semiotic: The Social Interpretation of Language and Meaning”. London, Arnold.
HALLIDAY, M.A.K. (1966/2008) “Lexis as a Linguistic Level”, in P. Hanks (ed.) *Lexicology: Critical Concepts in Linguistics*, vol. IV: Syntagmatics, pp. 3–15.
HALLIDAY, M. A. K. (1994). “An Introduction to Functional Grammar”. 2nd ed. London, Arnold.
HANKS, P. (forthcoming). “Analyzing the Lexicon: Norms and Exploitations”, MIT Press.
HARRIS, Z. (1954/1985). “Distributional Structure”, in J.J. Katz (ed.) *The phylosophy of linguistics.* New York: Oxford University Press, pp. 26–47.
HARRIS, Z. (1968). “Mathematical Structures of Language”. New York. Wiley-Interscience.
HEAPS, H. S. (1978) “Information Retrieval: Computational and Theoretical Aspects”. New York, Academic Press.
HEARST, M. (1992). “Automatic Acquisition of Hyponyms from Large Text Corpora”, in Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, pp. 539–545.
HEARST, M. (1999). “User Interfaces and Visualization”, in Baeza-Yates and Ribeiro-Neto (eds.) Modern Information Retrieval, ACM Press, Essex.
HEGEL, G. W. F. (1830/1975). “Logic. Part One of the Encyclopaedia of the Philosophical Sciences”, trans. William Wallace, Oxford, Clarendon Press.
HEGER, K. (1974). “Teoría semántica: hacia una semántica moderna”. Madrid, Herausgeber.
HEIDDEGER, M. (1927/1962). “Being and Time”, trans. J. Macquarrie & E. Robinson, Oxford, Blackwell.
HERACLITUS (ca. 500BC) “The Fragments of Heraclitus”, trans. by G.T.W. Patrick (1889).
HERDAN, G. (1964). “Quantitative Linguistics”. Washington. Butterworths.
HJELMSLEV, L. (1943/1953). “Prolegomena to a Theory of Language”. Baltimore, Indiana University Publications in Anthropology and Linguistics.
HOEY, M. (1991). “Patterns of Lexis in Text”. Oxford, Oxford University Press.
HOEY, M. (2004). “The textual priming of lexis”, in Guy Aston, Silvia Bernardini & Dominic Stewart (eds.), *Corpora and Language Learners*, (Amsterdam: John Benjamins), pp. 21–41.
HOEY, M. (2005). “Lexical Priming: A New Theory of Words and Language”. London: Routledge.
HOPFIELD, J. (1982). “Neural networks and physical systems with emergent collective computational abilities”, in Proceedings of the

National Academy of Sciencies, USA, vol. 79, April 1982. Biophysics, pp. 2554–2558.
HUME, D. (1777/1975). "An Enquiry concerning Human Understanding". Oxford. Clarendon Press.
IBEKWE-SANJUAN, F. & SANJUAN, E. (2004). "Mapping the structure of research topics through term variant clustering: the TermWatch system"; 7ème Journées internationales d'Analyse statistique des données textuelles (JADT' 2004). Louvain-la-Neuve, Belgium, 10-12 March 2004, pp. 589–600.
IDE, N. & VÉRONIS, J. (1993a). "Extracting knowledge bases from machine-readable dictionaries: have we wasted our time?", in KB&KS'93 Workshop, Tokyo, pp. 257–266.
IDE, N. & VÉRONIS, J. (1993b). "Refining taxonomies extracted from machine readable dictionaries". Research in Humanities Computing 2. Oxford University Press.
IDE, N. & VÉRONIS, J. (1995). "Knowledge extraction from machine-readable dictionaries: An evaluation". Machine Translation and the Lexicon, vol. 898, Berlin. Springer Verlag, pp. 19–34.
JACKENDOFF, R. (1990). "Semantic structures", The MIT Press, Cambridge, Massachusetts.
JACKENDOFF, R. (2002). "Foundations of Language: Brain, Meaning, Grammar, Evolution". Oxford/New York: Oxford University Press.
JACKENDOFF, R. (2007). "Linguistics in cognitive science: the state of the art". The Linguistic Review.
JACQUEMIN, C. (1997). "Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale", Université de Nantes, Nantes.
JACQUEMIN, C. (2001). "Spotting and Discovering terms through natural language processing". Cambridge (Mass.). MIT Press.
JAKOBSON, R. (1960). "Closing statement: Linguistics and poetics", in Sebeok, T. A. (ed.) *Style in language*. Cambridge, Mass.: MIT. pp. 350–377.
JANSSEN, M. (2002). "SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua". PhD Thesis, Utrecht University.
JUILLAND, A. & CHANG-RODRÍGUEZ, E. (1964). "Frequency Dictionary of Spanish Words". The Hague, Mouton.
KAGEURA, K. & UMINO, B. (1996). "Methods of Automatic Term Recognition". Terminology, vol. 3, no. 2, pp. 259–290.
KAGEURA, K.; TSUJI, K. & AIZAWA, A. (2000). "Automatic thesaurus generation through multiple filtering", in Proceedings of the 18th conference on Computational linguistics, July 31-August 04, 2000, Saarbrücken, pp. 397–403.
KANT, I. (1787/1929). "Critique of Pure Reason"; trans. N. Kemp Smith, McMillian.

KAROLINSKA INSTITUTET (2005). “Alphabetical List of Specific Diseases/Disorders”. Online Document. http://www.mic.stacken.kth.se/Diseases/Alphalist.html [accessed June 2010].
KATZ, J. J., (1985) “The philosophy of linguistics”. Oxford: Oxford University Press.
KILGARRIFF, A. (1997). “I don’t believe in word senses”. Computers and the Humanities vol. 31 no. 2, pp. 91–113.
KILGARRIFF, A.; RYCHLY, P; SMRZ, P. & TUGWELL, D. (2004). “The Sketch Engine”, in Proceedings of EURALEX 2004, Lorient, France. pp. 105–116.
KLEIN, M. (1932) “The Psychoanalysis of Children”. London. Hogarth Press.
KÖHLER, C.; KORTHAUS, A. & SCHADER, M.(2004). “(Semi-) Automatic Topic Map Generation From A Conventional Document Index”. International Conference on Knowledge Sharing and Collaborative Engineering, St Thomas.
KOMMERS, P. & LANZING, J.(1998). “Mapas conceptuales para el diseño de sistemas de hipermedia: navegación por la Web y autoevaluación”, in Carmen Vizcarro and José A. León (eds.) *Nuevas tecnologías para el aprendizaje*. Madrid, Pirámide, pp. 102–127.
KORZYBSKI, A. (1951). “The Role of Language in the Perceptual Processes”, in Robert R. Blake and Glenn V. Ramsey (eds.) *Perception: an approach to personality*. New York, Ronald Press Company.
KRONFELD, A. (1990). “Referrence and Computation: An Essay in Applied Philosophy of Language”. Cambridge University Press. Cambridge.
L'HOMME, M.C. (2005). “Sur la notion de 'terme'”, Meta, vol. 50, no. 4, pp. 1112–1132.
LACAN, J. (1956/1981). “Le séminaire, Livre III: Les psychoses”. Paris. Seuil.
LAKOFF, G. & JOHNSON, M. (1980) “Metaphors We Live By”. University of Chicago Press.
LANCIA F. (2007). “Word Co-occurrence and Similarity in Meaning”, in Salvatore S. and Valsiner J. (eds.), *Mind as Infinite Dimensionality*, Roma, Ed. Carlo Amore.
LANDAUER, T. K. & DUMAIS, S. T. (1997). “A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge”. Psychological Review, vol. 1, no. 4, pp. 211–140.
LANDAUER, T. K.; FOLTZ, P. W.; & LAHAM, D. (1998). “Introduction to Latent Semantic Analysis”. Discourse Processes, vol. 25, pp. 259–284.

LANGACKER, R. (2000). “A symbolic view of grammar”, in M. T. Cabré and C. Gelpí (eds.). *Cicle de conferències i seminaris 97-98: lèxic, corpus i diccionaris*. Barcelona, IULA.

LANGLAIS, P.; YVON, F. & ZWEIGENBAUM, P. (2008). “Analogical translation of medical words in different languages”, in Proceedings of the 6th International Conférence GoTAL 2008, Advances in Natural Language Processing, Gothenburg, Sweden, Aug. 2008, pp. 284–295.

LARA, L.F. (1997). “Teoría del diccionario monolingüe”. México D.F. El Colegio de México.

LEECH, G. (1981). “Semantics”. Harmondsworth, Penguin Books.

LEIBNIZ, G. (1765/1980). “Nuevo tratado sobre el entendimiento humano”, vol. I. Buenos Aires. Aguilar.

LENAT, D. (2006). “Problems of Scale in Building, Maintaining and Using Very Large Formal Ontologies”, in Proceedings of the Fourth International Conference (FOIS 2006) p. 3.

LENCI, A. (2008). “Distributional semantics in linguistic and cognitive research”. Italian Journal of Linguistics. Rivista di linguistica, vol. 20, no. 1, 2008, pp. 1–32.

LÉVI-STRAUSS, C. (1947/1993). “Las estructuras elementales del parentesco”, vol. I, Planeta, Barcelona.

LEWANDOWSKA-TOMASZCZYK, B. (2007). “Polysemy, prototypes and radial categories”, in Dirk Geeraerts and Hubert Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics*,. Oxford: Oxford University Press, pp. 139–169.

LLORÉNS, J. & ASTUDILLO, H. (2002). “Automatic generation of hierarchical taxonomies from free text using linguistic algorithms”, in Proceedings of OOIS Workshops 2002, pp. 74–83.

LIMA, M. (2005). “Visualcomplexity”. Online Document. http://www.visualcomplexity.com/vc/ [accessed June 2010].

LORENTE, M. (2005). “Ontology for economics and Information Retrieval”. Yearbook hipertext.net no. 3. Online Document. http://www.hipertext.net/web/pag259.htm [accessed June 2010].

LORITZ, D. (1999). “How the brain evolved language”. Oxford. Oxford University Press.

LUGLI, A.; ZLOBEC, I.; SINGER, G.; KOPP LUGLI, A.; TERRACCIANO, L. & GENTA, R. (2007). “Napoleon Bonaparte's gastric cancer: a clinicopathologic approach to staging, pathogenesis, and etiology”. Nature Clinical Practice Gastroenterology & Hepatology (2007) vol. 4, pp. 52–57.

LYONS, J. (1963/1989). “Semántica”. Barcelona: Editorial Teide.

MAGNUSSON, C. & VANHARANTA, H. (2003). “Visualizing Sequences of Texts Using Collocational Networks”. In P. Perner and A . Rosenfeld(Eds). MLDM 2003, LNAI 2734, Springer-Verlag, Berlin, Heidelberg, pp. 276–283.

MAICHER, L. (2004). “Subject Identification in Topic Maps in Theory and Practice”, in Tolksdorf, R., Eckstein, R. (eds.) *Berliner XML Tage 2004*. XML-Clearinghouse, Berlin, pp. 301–307 .

MALINOWSKI, B. (1923). “The problem of meaning in primitive languages”, in C. Ogden and I. Richards *The meaning of meaning*. London: Routledge and Kegan Paul, pp. 146–152.

MANDELBROT, B. (1957). “Linguistique Statistique Macroscopique”, in L. Apostel, B. Mandelbrot and A. Morf (eds.), *Logique, Language et Theorie de l'Information.* Paris, Presses Universitaires de France.

MANDELBROT, B. (1961). “On the theory of word frequencies and Markovian models of discourse”, in Proceedings of the Symposia on Applied Mathematics, vol. 12, American Mathematical Society, pp. 190–219.

MANDELBROT, B. (1983). “Los objetos fractales”. Barcelona. Tusquets.

MANNING, C. & SCHÜTZE, H. (1999). “Foundations of Statistical Natural Language Processing”, MIT Press. Cambridge, MA.

MANNING, C.; RAGHAVAN P. & SCHÜTZE, H. (2008). “Introduction to Information Retrieval”. Cambridge University Press.

MARKOV, A. (1913). “Essai d'une recherche statistique sur le texte du roman 'Eugene Oneguine',” Bulletin d'Academie Imperiale des Sciences de Saint-Petersbourg, vol. VII.

MAYNARD, D. & ANANIADOU, S. (2000b). “TRUCKS: A Model for Automatic Multi-Word Term Recognition”. Journal of Natural Language Processing, vol. 8, no. 1, pp. 101–125.

McCLELLAND, J.L. & RUMELHART, D.E. (1985). “Distributed Memory and the represenations of General and Specific Information”. Journal of Experimental Psychology: General, vol. 114, no. 2, pp. 159–188.

McCLELLAND, J. L. & BYBEE, J. (2007). “Gradience of Gradience: A reply to Jackendoff”. The Linguistic Review, 24, pp. 437–455.

MEL’ČUK, I.A. (1996). “Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon”, in Wanner, L. (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: Benjamins Academic Publishers, pp. 37–102.

MERLEAU-PONTY, M. (1945). “Phénoménologie de la perception”. Paris. Gallimard.

MILLER, G.A. (1995). “Virtual meaning”. In *Gothenburg Papers in Theoretical Linguistics* vol. 75, pp. 3–61.

MITCHELL, T. M.; SHINKAREVA, S. V.; CARLSON, A.; CHANG, K.M.; MALAVE, V. L.; MASON, R. A. & JUST, M. A. (2008) “Predicting Human Brain Activity Associated with the Meanings of Nouns”. *Science*, 320, May 30. pp. 1191–1195.

MORIN, E. & JACQUEMIN, C. (1999). “Projecting corpus-based semantic links on a thesaurus”, in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

MORTON, J. (1969). “Interaction of information in word recognition”. Psychological Review, vol. 76, pp. 165–178.

MOSBY (2001). “Diccionario Mosby de Medicina, Enfermería y Ciencias de la salud”. Quinta edición. Harcourt: Madrid. Versión en lengua española de la 5.ª edición de la obra original en inglés: Mosby's Medical, Nursing, and Allied Health Dictionary Copyright ©MCMXCVIII by Mosby Year Book, Inc.

MOSTELLER, F & WALLACE D. (1984). “Applied Bayesian and classical inference the case of the Federalist papers”. New York, Springer.

MULLER, C. (1973). “Estadística Lingüística”. Madrid, Gredos.

NAGATA, M., SAITO, T. & SUZUKI, K. (2001). “Using the Web as a Bilingual Dictionary”, in Proceedings of the workshop on Data-driven methods in machine translation, July 07-07, 2001, Toulouse, France, pp. 1–8.

NAZAR, R. (2005). “Aproximación cuantitativa al mapeo conceptual”. PhD Thesis Project. IULA, UPF.

NAZAR, R. (2008). “Bilingual Terminology Acquisition from Unrelated Corpora”, in Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008). Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra; Documenta Universitaria, pp. 1023–1029.

NAZAR, R.(2009). “Diferencias cuantitativas entre referencia y sentido”, in Applied Linguistics Now: Understanding Language and Mind = La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Actas del XXVI Congreso de la Asociación Española de Lingüística Aplicada. Almería: Universidad de Almería.

NAZAR, R.; VIVALDI, J. & WANNER, L. (2007). “Towards Quantitative Concept Analysis”, Procesamiento del Lenguaje Natural, no. 39, pp. 139–146.

NAZAR, R.; WANNER, L & VIVALDI, J. (2008). “Two Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora” in Proceedings of the twelfth conference of the European Association for Machine Translation. Hamburg: HITEC. pp. 138–147.

NAZAR, R.; VIVALDI, J. & CABRÉ, MT. (2008). “A Suite to Compile and Analyze an LSP Corpus”, in Proceedings of LREC 2008 (The 6th edition of the Language Resources and Evaluation Conference) Marrakech (Morocco, 28-30 May 2008), pp. 1164–1169.

NAZAR, R. & VIDAL, V. (2010). “Aproximación cuantitativa a la neología”, in M. Teresa Cabré i Castellví, Ona Domènech Bagaria, Rosa Estopà Bagot, Judit Freixa Aymerich, Mercè Lorente Casafont (Ed.) *Actes del I Congrés Internacional de Neologia de les llengües romaniques*. (CINEO 2008, Barcelona, 7-10 May, 2008). Documenta Universitaria, pp. 861–864.

NOONBURG, D. (2004). “Pdftotext - Portable Document Format (PDF) to text converter (version 3.00) README File”. Online Document. http://www.foolabs.com/xpdf/README [accessed June 2010].
NOVAK, J. & CAÑAS, A. J. (2006). “The Theory Underlying Concept Maps and How To Construct Them”. Technical Report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition.
OGDEN, C. & RICHARDS, I. (1984) “El Significado del significado: una investigación acerca de la influencia del lenguaje sobre el pensamiento y de la ciencia simbólica”. Barcelona, Paidós.
OLIVEIRA, H.; SANTOS, D. & GOMES, P. (2009). “Relations extracted from a portuguese dictionary: results and first evaluation”, in Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009), Aveiro, Portugal, pp. 541–552.
PANTEL, P. & LIN, D. (2001). “A Statistical Corpus-Based Term Extractor”, in Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence. London, UK, Springer-Verlag, pp. 36–46.
PARK, J. & HUNTING, S. (2003). “XML Topic Maps: creating and using topic maps for the Web”, Boston, Addison-Wesley cop.
PATRY, A. & LANGLAIS, P. (2005). “Corpus-Based Terminology Extraction”, in Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Danemark, pp. 313–321.
PEDRAZA, R. (2007). “Generación semiautomática de ontologías”. Online Document. http://www.iula.upf.edu/materials/070223pedraza.pdf [accessed June 2010].
PEIRCE, C.S. (1867). “On a New List of Categories”, Proceedings of the American Academy of Arts and Sciences 7 (1868), pp. 287–298.
PEIRCE, C.S. (1878). “How to Make Our Ideas Clear”. Popular Science Monthly 12, pp. 286–302.
PIAGET, J. (1963). “The origins of intellingence in children”. New York, Norton.
PHILLIPS, M. (1985). “Aspects of Text Structure: An Investigation of the Lexical Organization of Text”. North-Holland, Amsterdam
PLATO (ca. 380 B.C). “Meno”. Trans. by Benjamin Jowett. The Internet Classics Archive, Massachusetts Institute of Technology. Online Document. http://classics.mit.edu/Plato/meno.html [accessed June 2010].
PLAUT, D. (1995). “Semantic and Associative Priming in a Distributed Attractor Network”, in Proceedings of the 17th Annual Conference of the Cognitive Science Societym, Hillsdale, NJ, Lawrence Erlbaum Associates, pp. 37–42.
POPPING, R.(2000). “Computer assisted Text Analysis”, London, Sage.

PORTNER, P. & PARTEE, B. (2002). "Formal Semantics. The Essential Readings". Oxford, Blackwell.
POTRICH, A. & PIANTA, E. (2008). "L-ISA: Learning Domain Specific Isa-Relations from the Web", in Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA)}, Marrakech, Morocco, may, 28-30, 2008.
PRISS, U. (2006). "Formal Concept Analysis in Information Science". Annual Review of Information Science and Technology, vol. 40. Online Document. http://www.upriss.org.uk/papers/arist.pdf [accessed June 2010].
PURVES D.; AUGUSTINE G.; FITPATRICK D.; KATZ L.; LAMANTIA A. & MCNAMARA J. (2001). "Invitación a la Neurociencia". Buenos Aires, Ed. Panamericana.
QUASTHOFF, U.; RICHTER, M. & BIEMANN, C. (2006). "Corpus portal for search in monolingual corpora", in Proceedings of the LREC 2006, Genoa, Italy.
QUILLIAN, M. (1968). "Semantic Memory", in M. Minsky (ed.), Semantic Information Processing, MIT Press; reprinted in Collins & Smith (eds.), Readings in Cognitive Science, section 2.1, pp. 227–270.
QUINE, W. (1951). "Two Dogmas of Empiricism". *The Philosophical Review* vol. 60, pp. 20–43.
RAPP, R. (1999). "Automatic Identification of Word Translations from Unrelated English and German Corpora", in Proceedings of 37th Annual Meeting of the Association for Computational Linguistics. pp. 5190–526.
RATH, H.(1999). "Technical Issues on Topic Maps", STEP Electronic Publishing Solutions GmbH.
REDDING, P. (1997). "Georg Wilhelm Friedrich Hegel", in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2006 Edition).
RENAU, I. & BATTANER, P. (2008). "Agrupación semántica de verbos de un Daele a través de sus relaciones hiperonímicas", in III Congreso Internacional de Lexicografía Hispánica, Málaga.
RILOFF, E. & SHEPHERD, J. (1997). "A corpus-based approach for building semantic lexicons", in Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 127–132.
ROARK, B. & CHARNIAK, E. (1998). "Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction", in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, pp. 1110–1116.
ROSCH, E. (1975). "Cognitive representation of semantic categories". Journal of Experimental Psychology 104, pp. 573–605.
RUSSELL, B. 1905. "On Denoting". Mind, vol. 14, pp. 479–493.

RYDIN, S. (2002). “Building a hyponymy lexicon with hierarchical structure”, in Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, July 2002, Association for Computational Linguistics, pp. 26–33.
SAGGION, H. & GAIZAUSKAS, R. (2004). “Mining on-line sources for definition knowledge”, in Proceedings of the 17th International FLAIRS Conference.
SAHLGREN, M. (2008). “The distributional hypothesis”. Italian Journal of Linguistics, vol. 20, no. 1, 2008, pp. 33–54.
SALTON, G. & MCGILL, M. (1983). “Introduction to Modern Information Retrieval”. McGraw-Hill, New York, NY.
SANFILIPPO, A. & POZNANSKI, V. (1992). “The acquisition of lexical knowledge from combined machine-readable dictionary sources”, in Proceedings of the third conference on Applied natural language processing, Morristown, NJ, USA. Association for Computational Linguistics, pp. 80–87.
SAUSSURE, F. de (1916/1985). “Cours de linguistique générale”. Paris, Payot.
SCHANK, R. (1975). “Conceptual Information Processing”. Amsterdam, North-Holland Pub.
SCHMIDT, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees”, in Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49.
SCHULTE IM WALDE, S & MELINGER, A. (2008). “An in-depth look into the co-occurrence distribution of semantic associates”. Italian Journal of Linguistics, vol. 20, no. 1, 2008, pp. 89–128.
SCHÜTZE, H. (1998) “Automatic word sense discrimination”. Computational Linguistics, MIT Press. Cambridge, MA, vol. 24, no. 1, pp. 97–123.
SCHÜTZE, H. & PEDERSEN, J. (1995). “Information Retrieval Based on Word Senses”, in Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175.
SCHÜTZE, H. & PEDERSEN, J.(1997). “A co-occurrence-based thesaurus and two applications to information retrieval”. Information Processing and Management, vol. 33, no. 3, pp. 307–318.
SHADBOLT, N.; BERNERS-LEE, T. & HALL, W. (2006). “The Semantic Web Revisited”, IEEE Intelligent Systems 21(3), May/June, pp. 96–101.
SHANNON, C. E. (1948). “A mathematical theory of communication”. Bell System Technical Journal, vol. 27, July and October, pp. 379–423 and pp. 623–656.
SHAPIRO, A. (2001) “TouchGraph Software”. Online Document. http://www.touchgraph.com/TGGoogleBrowser.html [accessed June 2010].

SHEREMETYEVA, S. (2009). "On Extracting Multiword NP Terminology for MT", in Proceedings of the EAMT Conference. Barcelona, Spain, May 13-15.
SIERRA, G.; ALARCÓN, R. & AGUILAR, C. (2006). "Extracción automática de contextos definitorios en textos especializados". Procesamiento del Lenaguaje Natural, no. 37, pp. 351–352.
SINCLAIR, J, (1991). "Corpus, concordance, collocation", Oxford, Oxford University Press.
SINCLAIR, J.(1966/2008) "Beginning the Study of Lexis", in P. Hanks (ed.) *Lexicology: Critical Concepts in Linguistics*. vol. IV: Syntagmatics, pp. 16–34.
SKINNER, B. F. (1957). "Verbal behavior", New York, Appleton-Century-Crofts.
SKIPPER, J. I. & ZEVIN, J. D. (2010). "How the brain predicts forthcoming words during sentence listening". Paper to be presented at the Annual Meeting of the Cognitive Neuroscience Society, Montreal, Canada.
SMADJA, F. & MCKEOWN, K. (1990) "Automatically extracting and representing collocations for language generation", in Proceedings of the 28th Annual Meeting of the Association for Comp. Ling, ACL, 1990, pp. 252–259.
SNOW, R.; JURAFSKY, D. & NG, A. (2006). "Semantic taxonomy induction from heterogenous evidence", in Proceedings of the 21st International Conference on Computational Linguistics (COLING), Sydney, Australia, pp. 801–808.
SOWA. J. (1991). "Principles of Semantic Networks", Morgan Kaufmann Publishers, San Mateo, CA, US.
SOWA. J. (1992). "Semantic networks", Encyclopedia of Artificial Intelligence, edited by S. C. Shapiro, Wiley, New York, 1987.
SOWA, J. (2000a). "Ontology, Metadata, and Semiotics", Darmstadt, ICCS.
SOWA, J. (2000b). "Knowledge representation logical, philosophical, and computational foundations", Pacific Grove Brooks/Cole cop.
SPARCK JONES, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, vol. 28, no. 1, pp. 11–21.
SPECIA, L. & MOTTA, E. (2006). "A Hybrid Approach For Extracting Semantic Relations From Texts". Workshop On Ontology Learning And Population: Bridging The Gap Between Text And Knowledge-
STRAWSON, P.F. 1950. "On Referring". Mind, vol. 235, pp. 320–344.
TANAKA, T. & MATSUO, Y. (1999). "Extraction of Translation Equivalents from Non-Parallel Corpora", in Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 109–119.
TAYLOR, M; MATUSZEK, C.; KLIMT, B. & WITBROCK, M. (2007). "Autonomous Classification of Knowledge into an Ontology", in

Proceedings of the Twentieth International FLAIRS Conference, Key West, FL, May 2007.
TESNIÈRE, L. (1959/1965). "Éléments de Syntaxe Structurale", Paris, Librairie C. Klincksieck.
THINKMAP INC. (2004). "VisualThesaurus.com". Online Document. http://www.visualthesaurus.com [accessed June 2010].
TORRES-MORENO, J. M.; ST-ONGE, P.L; GAGNON, M.; EL-BÈZE, M. & BELLOT, P. (2009) . "Automatic Summarization System coupled with a Question-Answering System (QAAS)". Online Document. http://arxiv.org/pdf/0905.2990 (arXiv:0905.2990 - May 2009) [accessed on May 2010] .
TRUJILLO, R. (2006). "Lenguaje, individuo y sociedad", in Mercedes Sedano, Adriana Bolivar y Martha Shiro (eds.), *Haciendo Lingüística Homenaje a Paola Bentivoglio*. Caracas: Universidad Central de Venezuela.
UCHIDA, H. (1987). "ATLAS: Fujitsu Machine Translation System". Machine Translation Summit, September 17-19, 1987, Hakone Prince Hotel, Japan.
VAN DIJK, T. (1983). "La ciencia del texto", Barcelona; Buenos Aires, Paidos.
VAN DIJK, T. (1993). "Texto y contexto", Madrid; Cátedra.
VAN DIJK, T. (2000). "Estudios del discurso", Barcelona; Buenos Aires; Gedisa.
VAN OS, A. (2003). "Antiword" Linux Command README File. Online Document. http://www.winfield.demon.nl/index.html [accessed June 2010].
VEGA, M. & CUETOS, F. (1999). "Psicolingüística del español"; Madrid, Trotta.
VELARDI, P.; CUCCHIARELLI, A. & PETIT, M. (2007). "A taxonomy learning method and its application to characterize a scientific web community". IEEE Trans. on Knowl. and Data Eng., vol. 19, no. 2, pp. 180–191.
VÉRONIS, J. & Ide, N. (1991). "An assessment of semantic information automatically extracted from machine readable dictionaries". Fifth Conference of the European Chapter of the Association for Computational Linguistics, Berlin, pp. 227–233.
VÉRONIS, J. (2004). "HyperLex: Lexical Cartography for Information Retrieval". Computer Speech & Language, vol. 18, no 3, pp. 223–252.
VIGNAUX, G. (1986). "La argumentación". Buenos Aires. Hachette.
VIVALDI, J. (2001). "Extracción de candidatos a término mediante combinación de estrategias heterogéneas". PhD Thesis, IULA, Universitat Pompeu Fabra.
VIVALDI, J. (2009). "Corpus and exploitation tool: IULACT and bwanaNet". A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Actas del I Congreso

Internacional de Lingüística de Corpus (CICL-09), 7-9 Mayo 2009, Universidad de Murcia]. Murcia: Asociación Española de Lingüística del Corpus, pp. 224–239.
VON FOERSTER, H. (1984). "On Constructing a Reality", in Watzlawick, P. (Eds), *The Invented Reality: How do we know what we Believe we know?*, W.W. Norton, New York.
WANNER, L.; BOHNET, B. & GIERETH, M. (2006). "Making Sense of Collocations". Computer Speech & Language, vol. 20, no. 4, October 2006, pp. 609–624.
WATTS, D. & STROGATZ, S. (1998). "Collective dynamics of 'small-world' networks", Nature, vol. 393, pp. 440–442.
WATZLAWICK, P.; BEAVIN, J. & JACKSON, D. (1985). "Teoría de la Comunicación Humana", Barcelona, Herder.
WEIDER, B. (1998). "The Assassination of Napoleon". Lecture given at the International Military History Festival "Borodino Day", Borodino Russia - September 5-10, 1997 and at the Sandhurst Military Academy, London, England, February 18, 1998.
WETTLER, M.; RAPP, R. & SEDLMEIER, P. (2005). "Free Word Associations Correspond to Contiguities Between Words in Texts". Journal of Quantitative Linguistics; Routledge, Taylor & Francis Group, 2005, vol. 12, no. 2-3., pp. 111–122.
WIERZBICKA, A. (1996). "Semantics: primes and universals". Oxford. Oxford University Press.
WIDDOWS, D. (2004). "Geometry and Meaning", Center for the Study of Language and Information/SRI.
WIDDOWS, D. & DOROW, B. (2002). "A Graph Model for Unsupervised Lexical Acquisition", in Proceedings of the 19th International Conference on Computational Linguistics, Taipei, August 2002, pp. 1093–1099.
WIENER, N. (1948). "Cybernetics or Control and Communication in the Animal and the Machine". Cambridge. MIT Press.
WILLIAMS, G.C. (1998). "Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles". International Journal of Corpus Linguistics, John Benjamins Publishing Co., vol. 3, no. 1, pp. 151–71.
WILLIAMS, J.N. (2001). "Psycholinguistics", in K. Malmkjaer (Ed.) *The Linguistics Encyclopedia*. Second Edition. London, Routledge.
WOLFF, K. (1994). "A first Course in Formal Concept Analysis", in Proceedings of SoftStat'93. Gustav Fischer Verlag.
WÜSTER, E. (1979/1998). "Introducción a la teoría general de la teminología y a la lexicografía". Institut Universitari de Lingüística Aplicada, Barcelona.
YAROWSKY, D. (1992). "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora", in Proceedings of *Coling-92*, Nante, pp. 454–460.

YAROWSKY, D. (1995). “Unsupervised word sense disambiguation rivaling supervised methods”, in Proceedings of the 33rd Annual Meeting of ACL, pp. 189–196.
YULE G.U. (1944). “The Statistical Study of Literary Vocabulary”. Cambridge University Press, Cambridge.
ZIPF, G.K. (1949). “Human Behavior and the Principle of Least-Effort”. Addison-Wesley.

APPENDIX A

Table with the rest of the results from the experiment in section 7.3.3.

==========

EXPERIMENT 2. **Sign:** APCS

==========

Senses in Wikipedia:
Advanced Passenger Control System
Amyloid P Component, Serum, a human gene
Advanced Placement Computer Science
Argonne Premium Coal Sample
Assembly for the Promotion of Civil Society
Australian-Polish Community Services
ARM Procedure Call Standard
Avoyelles Public Charter School

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
Advanced Placement Computer Science	15	1	2	94	89
Antigen-presenting cells	4	4	4	50	50
Armored Personnel Carriers	7	0	6	100	54
mob burns	3	0	0	100	100
Ambulatory Payment Classifications	5	0	1	100	80

Clusters found by the human evaluator:
Australia Power Control Systems (4 documents)

Examples of clusters rejected for not having enough members:

Clusters of two documents:
adenomatous polyposis coli
Asia Pacific Cities Summit
Asynchronous Procedure Calls
Academy of Psychological Clinical Science

Isolated documents:

Amyloid P component serum
American Pencil Collectors Society
Automated PC Solutions
Associated Presbyterian Churches
Australian-Polish Community Service
Association For the Psychoanalysis of Culture and Society
APCS Power Clearing and Settlement

Academy of Psychological Clinical Science
Advance Purchase Commitments
Asia Pacific Cultural Studies
Atlantic Pacific Companion
Advance Panel Construction Services
Australian Payments Clearing Association
Australian Project and Consulting Services
Advanced Payments Credits
Amsterdam Power Classic Station
Approach Power Compensator System
Areas of Specific Concerns
Australian Pay and Classification Scale

==========

EXPERIMENT 3. **Sign:** ASG

==========

Senses in Wikipedia:
ASG (band)
Abstract semantic graph
Abu Sayyaf Group
Adaptive Services Grid
Airsoft gun
All Saints Greek Orthodox Grammar School, or All Saints Grammar
All-star game
Alsager railway station's national rail code
Alternating Step Generator, a cryptographically secure pseudorandom number generator
Australasian Seabird Group, a special interest group of Birds Australia
Australian Standard Garratt, a WWII-era Australian steam-engine locomotive
Austria Switzerland Germany, a geographic region used by companies to name the predominately German speaking countries in Europe
Labour and Social Justice Party (Arbeit & soziale Gerechtigkeit)
Atlanta Spirit Group, the governing body of the Atlanta Hawks, Atlanta Thrashers, and Philips Arena

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
ASG SOFTWARE	4	1	0	75	100
ASG (rock band)	3	0	0	100	100

Examples of clusters rejected for not having enough members:
Clusters of two documents:
American Sewing Guild
Advanced Surveillance Group
Australian Scholarships Group

Isolated documents:
Army Sub Group
Aids Services Group
Automated Solutions Group
Aplicaciones y sistemas de gestión
Administración de servicios generales

Austin Stitchery Guild
Adaptive Servide Grid
Alternative Splicing Gallery
Advance Succed Grow
Application Systems Group
Aronson Security Group
American Society of Genealogists
Ateneo School of Government
Academy of Striking and Grappling

==========

EXPERIMENT 4. **Sign:** BVM

==========

Senses in Wikipedia:

Blessed Virgin Mary, the mother of Jesus Christ and a main figure in Christianity
Sisters of Charity of the Blessed Virgin Mary, a religious order
Bag valve mask, a device used in resuscitation procedures to assist patients in breathing
Birla Vishwakarma Mahavidhyalaya, the first ever engineering college in Gujarat; part of Sardar Patel University situated in city of Anand, Gujarat state, India
Big Vinnie Mac; a nickname of Vince McMahon.

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
Blessed Virgin Mary	11	2	5	82	69
bvm apparitions	5	2	0	60	100
master monitor	6	0	1	100	74
bvm-tv kdoc	3	0	0	100	100

Examples of isolated documents:

Bolsa de Valores de Montevideo
BVM Engineering
Bildverarbeitung für die Medizin
Britt Vegetation Management
Badminton Video Magazine

==========

EXPERIMENT 5. **Sign:** CKD

==========

Senses in Wikipedia:

Chronic kidney disease, a slowly progressive loss of renal function
Complete knock down, a complete kit needed to assemble a vehicle
Count Key Data, a disk architecture used in IBM mainframe computers
CKD (Ceskomoravská Kolben-Danek), an engineering company in the Czech Republic
CKD, a Certified Kitchen Designer with the National Kitchen and Bath Association

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
chronic kidney disease	49	6	6	88	89
ckd galbraith	6	1	0	84	100

==========

EXPERIMENT 6. **Sign:** DDO

==========

Senses in Wikipedia:

David Dunlap Observatory Catalogue, the "A Catalogue of Dwarf Galaxies" compiled at the David Dunlap Observatory.

DDO Artists Agency, a talent agency

Defense Depot Ogden, a former U.S. military installation located in Ogden, Utah.

Deputy Director for Operations - former title for the head of the branch of the U.S. Central Intelligence Agency that conducts covert operations. The position is now known as Director of the National Clandestine Service See also: National Clandestine Service.

Diocese Director of Ordinands - The person in charge of discerning vocations in the Anglican Church.

Dollard-des-Ormeaux - A city in the Greater Montreal area of Quebec.

Dubberly Design Office - An interaction and service design firm in San Francisco, CA.

Dungeons & Dragons Online: Stormreach - a massively multiplayer online role-playing game based on Dungeons & Dragons 3.5 edition developed by Turbine, Inc.

Dynamic Drive Overlay

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
David Dunlap Observatory	8	1	3	88	66
Civic Honda D.D.O	3	0	0	100	100
Dungeons and Dragons	45	5	4	89	92

==========

EXPERIMENT 7. **Sign:** ETN

==========

Senses in Wikipedia:

Erythritol tetranitrate, an explosive chemical compound

Exchange Traded Notes, a synthetic financial derivative

Erie Times-News, the daily newspaper in Erie, Pennsylvania

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
Efficiency technical note	11	0	0	100	100
early transposon	3	0	0	100	100
AASC Talent Show	3	0	0	100	100
low-cost airlines	3	0	0	100	100

Eaton corporation	5	0	6	100	46
data provided	5	0	5	100	50
Exchange traded notes	5	1	5	80	50
Exchange traded notes II	4	1	4	75	50
EurTradeNet: customs administrations	3	0	0	100	100
Employment Training Network (espurio)	5				

==========

EXPERIMENT 8. **Sign:** FYI

==========

Senses in Wikipedia:

FYI, short for 'For Your Information', is commonly used in email or memo messages, typically as the first thing on the subject line, to flag the message as an informational message that does not require a response.

On the Internet, FYIs are a subset of the RFC series.

The following description is taken from FYI 1, the FYI on FYIs: The FYI series of notes is designed to provide Internet users with a central repository of information about any topics which relate to the Internet. FYIs topics may range from historical memos on "Why it was done this way" to answers to commonly asked operational questions.

FYI was also the name of a fictional news magazine on the television series Murphy Brown.

FYI was also the name of a 1980s daily information program on ABC.

Film Your Issue is an annual film-making competition.

FY_, a play on the common "FYI". Here you change the last letter of the acronym to match the first letter of your name. For instance, if your name is "Jeff", you would "FYJ" someone (that is to say, you would "For Your Jeff" them).

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
--> --> (spurious cluster)	7				
web site (spurious cluster)	5				

Note: There are no clusters in this experiments. There is only one prominent sense that is *For Your Information,* however the set that these pages conform are not related from the point of view of the content, therefore it is good that the program had reacted in this way.

==========

EXPERIMENT 9. **Sign:** IED

==========

Senses in Wikipedia:

Improvised explosive device, an explosive device often used in unconventional warfare

Istituto Europeo di Design, The European Institute of Design

Institution of Engineering Designers, an UK Engineering Council ECUK registered professional society

Indo-European Etymological Dictionary, by Julius Pokorny
Instantaneous electrical detonator
Interlingua-English Dictionary, the first major presentation of Interlingua to the public
Intermittent explosive disorder, a disorder of the brain.
International Endowment for Democracy, a U.S. non-profit organization.
Intelligent electronic device
Intelligent edge device, an interface device supporting complex protocols and packet encapsulations to manage diverse network topologies and deliver new services
Immigration Employment Document, in the context of UK Immigration Documentation.
Institute for Educational Development, at the Aga Khan University
Individual educational development, one name for a Special education class
Inter-ictal Epileptiform Discharge, a characteristic of Epileptic seizures.

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
improvised explosive device	15	0	9	100	63
ied attacks	6	0	0	100	100
institute for educational development	3	0	0	100	100
Istituto Europeo di Design	8	2	3	75	73

==========

EXPERIMENT 10. **Sign:** JUB

==========

Senses in Wikipedia:
Jacobs University Bremen, a private research university in Germany. The use this abbreviation is actively discouraged by the administration of the Institution.
Jaiminiya Upanishad Brahmana, a Vedic text
JustUsBoys, a gay community website.

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
juba airport	3	0	0	100	100
j-u-b engineers	3	0	0	100	100
gene expression	3	0	2	100	60
Sir Jub-Jub	3	1	0	66	100
jub jub (spurious cluster)	30				
flash player (spurious cluster)	3				

Clusters found by the human evaluator:
JUB painting (4 documents)

==========

EXPERIMENT 11. **Sign:** KPS

==========

Senses in Wikipedia:

Kalamazoo Public Schools
Karnofsky performance status, a measurement of a cancer patient's wellbeing
Kilometers per second (see Metre per second)
Kosovo Police Service
K. P. S. Gill (Kanwar Pal Singh Gill), a former police officer and sports administrator in India
League of Communists of Slovenia or Komunisticna partija Slovenije
Kilobits Per Second
Kilobytes Per Second
Kamikaze Peep Squad
Kawasaki Production System
Keratic Precipitates (ophthalmology)
Kappa Pi Sigma (Kings Point Spirit (Kings Point Sucks))
Korean Physical Society
Kranji Primary School (Singapore)
Keming Primary School (Singapore)
Kent Place School (Summit, New Jersey)
Kingscliff Public School (Gold Coast, Australia)
Kingsgrove Public School (Kingsgrove, Australia)

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
kps gill	8	0	2	100	80
kps health	5	2	0	60	100
kps consortium	3	0	0	100	100
file type	3	0	0	100	100
kps (a youtube user)	3	0	0	100	100
hong kong (spurious cluster)	6				
web site (spurious cluster)	3				

==========

EXPERIMENT 12. **Sign:** KSP

==========

Senses in Wikipedia:

Karolinska Scales of Personality, a personality questionnaire
Kentucky State Police
Kerala Socialist Party
Keyword Services Platform
Korea Socialist Party
Kassettbolaget Svart Pyramid
Key signing party
Solubility product constant (Ksp)
KSP Sound Player

The filename extension of the KSpread spreadsheet application (.ksp)

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
KSP Lite (software)	3	0	0	100	100
K-Sea Transportation Partners LP	8	4	1	50	89
The Hindu (India's Newspaper)	3	0	0	100	100
KSP inhibitor	5	0	1	100	84
calcium carbonate (spurious cluster)	5				

Clusters found by the human evaluator:
Kentucky State Police (4 docs)

==========

EXPERIMENT 13. **Sign:** LEP

==========

Senses in Wikipedia:
Lancashire Evening Post, an English newspaper
Large Electron-Positron Collider, one of the largest particle accelerators
Liberal Egyptian Party
Licensed Educational Psychologist
Light emitting polymer
Limited English proficiency, a term within English language learning and teaching
local ecumenical partnership or local ecumenical project
local education partnership
Lower Elements Police, a police squad found in the Artemis Fowl series of children's novels.
Lepidoptera, the category for Butterflies, Moths and Skippers.
Laterally extended parametrectomy
Lepus (constellation), an astronomical constellation

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
Limited English Proficiency	28	1	5	97	85
Large Electron Positron (particle physics)	8	0	2	100	80
hip hop artist	5	2	0	60	100
Local Employment Partnerships	3	0	0	100	100

Clusters found by the human evaluator:
lep gene (leptin) 3 documentos

==========

EXPERIMENT 14. **Sign:** Nafta

==========

Senses in Wikipedia:

An acronym for the North American Free Trade Agreement
An acronym for the New Zealand Australia Free Trade Agreement
The town of Nafta, Tunisia
Naphtha, various liquid hydrocarbon intermediates produced from fossil fuel
The Polish name for Kerosene/paraffin oil
The Ukrainian name for Petroleum
Nafta, a Soviet Union oil company operating abroad

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
North American Free Trade Agreement	79	1	7	99	92

==========

EXPERIMENT 15. **Sign:** NCO

==========

Senses in Wikipedia:

NAFTA certificate of origin, a customs document for certificate of origin
NCO Group, an international corporation that provides customer service contracting
Net Capital Outflow, an economic metric measuring the amount of money from a country holding assets elsewhere
NetCDF Operators, a suite of programs for manipulating NetCDF files
Network-centric operations, an emerging theory of war in the information age
Non-commissioned officer in the rank structure of many armed forces worldwide
Numerically-controlled oscillator, a clock divider for microprocessors
Nuova Camorra Organizzata, a defunct Italian Camorra criminal organization in Naples founded by Raffaele Cutolo.

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
nco group (servicios financieros)	22	2	2	91	92
Noncommissioned officers	15	1	8	94	66
nco file extension	4	0	0	100	100
netCDF Operators (software)	5	0	4	100	66
contact nco (spurious cluster)	3				

==========

EXPERIMENT 16. **Sign:** NPN

==========

Senses in Wikipedia:
NPN transistor, a type of bipolar junction transistor
Na Progu Nieznanego, a UFO research organization in Poland
National Party of Nigeria, the dominant political party in Nigeria during the Second Republic
National Peer Network, a network of support groups that meet with amputees
New Politics Network, a United Kingdom independent political and campaigning think tank
Non-protein Nitrogen, an animal feed component

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
npn transistor	36	4	0	89	100
National Phenology Network	5	1	0	80	100

Clusters found by the human evaluator:
national performance net
docs: 4 (one in German)
the new plan network
docs: 3

==========
EXPERIMENT 17. **Sign:** OLA
==========

Senses in Wikipedia:
Our Lady of the Abandoned
Online Authorisation
Operating Level Agreement
Ontario Lacrosse Association
Ontario Legislative Assembly
Ontario Library Association
Office of Legal Affairs - legal departments of many national governments and the UN.
Online Advertising
Overlap-add method

No clusters (M and the evaluator agree).

==========
EXPERIMENT 18. **Sign:** OOV
==========

Senses in Wikipedia:
OOV is a Dutch abbreviation for "Openbare Orde en Veiligheid", which translates as "Public Order and Safety"
This three-letter acronym refers to the combination of the Police force, Fire department and various (para)medical services (such as ambulance and SAR (Search and Rescue) services) that provide their services for the public order and safety.
The emergency telephone number in the Netherlands is 1-1-2.

OOV in text processing stands for 'out-of-vocabulary', i.e. a word that is not known in the computer's online dictionary. OOV words are often a problem for automatic spell checkers and voice recognition software.
OOV can also stand for 'out-of-vision', such as in film scripts when an actor is talking and is audible, but not visible in the shot

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
oov (artist)	6	3	1	50	86
out of vocabulary (information retrieval)	16	0	2	100	89
oov vs. (vs. other artists)	8	0	0	100	100
t-shirts	4	1	0	75	100
oov (aircraft)	4	2	1	50	75

==========

EXPERIMENT 19. **Sign:** PCR

==========

Senses in Wikipedia:
the Communist Party of Réunion (Parti Communiste Réunionnais)
Passive Covert Radar, a radar system exploiting broadcast transmitters
Passport Carrier Release, the name of the operating system that runs on Nortel Multiservice switches in carrier configurations
Peak Cell Rate, a source traffic characteristic on ATM networks
Periodic current reversal a technique used in electrochemistry
Platform Configuration Register, a Trusted Platform Module register storing platform configuration measurements
Polymerase chain reaction
Practical Chinese Reader, a book series designed to teach students Chinese
Principal component regression, a statistical technique
Program Clock Reference, a feature of MPEG
Process capability ratio
the Romanian Communist Party (Partidul Comunist Român)
XM PCR, a satellite receiver

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
polymerase chain reaction	52	3	10	95	84
pcr primer tools (sub-cluster from previous)	3	0	2	100	66
midi keyboard	3	1	0	66	100

==========

EXPERIMENT 20. **Sign:** PTW

==========

Senses in Wikipedia:

PTW Architects, best known for designing the Beijing National Aquatics Center
Pottstown Limerick Airport
Poison the Well, a musical group from Florida

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
wep key (software)	9	0	1	100	90
systema ptw (airsoft gun)	7	0	2	100	78
ptw architects	7	0	1	100	88
vantage point (television wall mounts)	3	0	0	100	100
civilization iii (pc game)	5	2	0	60	100
paper thin walls	3	0	0	100	100

==========

EXPERIMENT 21. **Sign:** RLA

==========

Senses in Wikipedia:

Railway Labor Act, an American law governing labour relations in the railway and airline industries.
The IATA code for Rolla Downtown Airport in Rolla, Missouri, United States
Restricted Landing Area, an American type of private or otherwise limited-use aerodrome.
The Royal Lao Army, the armed forces of the Kingdom of Laos.

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
real estate	13	4	1	70	93
macro photography	9	2	0	78	100
railway labor act	6	1	0	84	100
file extension	5	0	0	100	100
indigenous peoples	3	1	0	66	100

==========

EXPERIMENT 22. **Sign:** SEU

==========

Senses in Wikipedia:

A single event upset is a change of state caused by a high-energy particle strike to a sensitive node in a micro-electronic device, such as in a microprocessor, semiconductor memory, or power transistors.

subjective expected utility is a way to make decisions (decision theory) in the presence of risk.

Southeastern University, a private, non-profit university located in Washington, D.C.

Southeast University (Bangladesh) is a private, non-profit university located at Banani, Dhaka, Bangladesh.

Southeast University (China) is a university located in Nanjing, Jiangsu Province, China.

Saint Mary of Valencia Cathedral, as it is known in Catalan

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
seu jorge	24	3	0	88	100
Seu D'Urgell	7	0	1	100	88
l'Almoina del Pa de la Seu de Girona	3	1	0	66	100

==========

EXPERIMENT 23. **Sign:** TDD

==========

Senses in Wikipedia:

Technical design document, a low-level type of design document

Telecommunications device for the deaf, a device for text communication along a telephone line

Test-driven development, a type of software development model

Trick Daddy Dollars, American rap artist

Three Doors Down, an American rock band

Transdermal Drug Delivery, a method of delivering drugs through the skin and into the bloodstream

Time-division duplex, the application of time-division multiplexing to separate outward and return signals

Teniente Jorge Henrich Arauz Airport, Trinidad, Bolivia, IATA airport code

Ternary Decision Diagram

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
test-driven development	63	0	2	100	97
Telecommunications Device for the Deaf	6	0	4	100	60

==========

EXPERIMENT 24. **Sign:** TLA

==========

Senses in Wikipedia:

Temporal Logic of Actions, a logic used to describe behaviours of concurrent systems
Textile Labour Association Indian trade union formed in 1917 with the aid of Gandhi, precursor of SEWA
Theater of the Living Arts, colloquially known as the The TLA, Philadelphia
TLA Entertainment Group, a movie retailer and distributor based in Philadelphia, USA
TLA Releasing, film distribution division
Golden Sun: The Lost Age, a 2003 video game
Tom Lord's Arch, a component of the GNU arch revision control system

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
tla (gay movies)	5	0	6	100	46
tla video (sub-cluster of the previous)	4	0	0	100	100
Text Link Ads	5	1	1	80	84
GNU Arch 1 (also known as tla)	5	0	0	100	100
Sub-TLA Assignments	4	0	0	100	100
three-letter acronym	5	0	1	100	84
Temporal Logic of Actions	3	0	1	100	75
web site (spurious cluster)	9				

Clusters found by the human evaluator: texas library association docs: 4

==========

EXPERIMENT 25. **Sign:** WTF

==========

Senses in Wikipedia:

WTF is an internet slang acronym for "What the fuck?"
World Taekwondo Federation, a member of the International Olympic Committee
Waking the Fallen, an album by Avenged Sevenfold
Werewolf: The Forsaken, a role-playing game developed by White Wolf Game Studio
Work Time Fun, a game for the PlayStation Portable
Working Title Films, a United Kingdom film production company
Walking Through Fire, album by April Wine
Weekly Top Five, a video program by World Wrestling Entertainment on wwe.com
Warren T. Furutani, a California State Assemblyman
The Daily WTF, a blog about information technology perversions

Clusters found:

Cluster	Documents	Errors	Omissions	%Precision	%Recall
john mccain	6	1	0	84	100
The Daily WTF	3	0	0	100	100
wtf microsiervos	3	0	0	100	100
videos de youtube	4	1	0	75	100
email address (spurious cluster)	4				

Clusters found by the human evaluator: taekwondo tournament docs: 4

APPENDIX B

Table with the different senses and examples of the sample of 10 polysemous words in the dataset of Véronis (2004) manually disambiguated by a native French Speaker.

Word	Contexts	Senses	Examples
barrage	100	routier	Barrages non autorisés. Il y a un mois, le ministre d'Etat, ministre de l'Intérieur, de la Sécurité et de la Décentralisation, Emile Boga Doudou, se réjouissait de la baisse de la criminalité en Côte d'Ivoire. Il expliquait cette baisse par la levée des barrages et par l'instauration des patrouilles mobiles. Un mois plus tard, le constat est toujours alarmant. Catastrophique, même. Non seulement les barrages illégaux sont toujours en vigueur mais les forces de l'ordre redoublent de férocité. S'estimant mal payées, elles revendiquent un droit de passage auprès des automobilistes. 500 F CFA (5 FF) par conducteur. Les chauffeurs de taxi et les transporteurs routiers sont les premiers à se plaindre des agissements des policiers.
		frontiere	La défense des frontières par les barrages terrestres est une des préoccupations majeures du commandement depuis que le Maroc et la Tunisie sont devenus indépendants, c'est-à-dire depuis 1956 et jusqu'au jour de l'indépendance algérienne. Les barrages sont une création continue ; ils évoluent en fonction du développement des armées de libération nationale (ALN) se constituant à l'extérieur des frontières. Si les passages à travers les barrages se tarissent, les harcèlements se développent et la crainte d'un passage en force d'une ALN de Tunisie bien armée est entretenue jusqu' en 1962. La construction, l'entretien et la défense de ces barrages sont une mission interarmées puisque y participent marins (1) et aviateurs aux côtés de toutes les armes de l'armée de Terre, comme l'illustrent les études et témoignages ici réunis.
		match	1-VI Barrages, 1ère journée Profitant de la victoire aux tirs au but, qui ne rapporte que deux points, de Saint-Brieuc face à Segré, le FC Mantois prend la tête du mini-championnat après son succès 2-1 face à la réserve de Sedan. Il ne fallait pas arriver en retard au stade Aimé-Bergeal... ni partir avant la fin du match! Après 120 secondes de jeu, les Ardennais débloquent la marque sur une tête lobée de Harreau. Piqués au vif, les Mantois se ruent à l'assaut des buts sedanais et égalisent logiquement sur un coup de tête de Rameau (23e). La rencontre est haletante avec de nombreuses occasions de part et d'autres ; la plus belle est pour les Mantois avec un tir de Preira sur la barre (70e). La dernière poussée des locaux est la bonne : lancé en profondeur par Challouf, Ictoi inscrit le but de la victoire à l'entame des arrêts de jeu! Saint-Brieuc est passé tout près de la victoire dans le temps réglementaire : alors que le score est de 1-1, les tentatives de Morin (79e) et Desriac (80e) trouvent les montants segréens!! Grâce à deux parades de leur gardien Druguet, les Griffons remportent la

			séance de tirs au but, 4-2. 1-VI Mantes 2, Sedan Rés. 1 Rameau (23e), Ictoi (90e+1) pour Mantes ; Harreau (2e) pour Sedan Rés. 800 spec. Saint-Brieuc 1, Segré 1 Saint-Brieuc vainqueur aux tirs au but, 4-2. Allainmat (50e) pour Saint-Brieuc ; Vallais (p 31e) pour Segré. 500 spec.
		eau	Les barrages mobiles Les barrages mobiles sont des ouvrages qui sont construits dans la partie aval (basse) du cours des rivières ou la pente est faible. Ils sont constitués de bouchures (grandes vannes) et d'une infrastructure en béton. L'avantage des barrages mobiles : On peut règler l'ouverture des vannes et ainsi réguler l'arrivée d'eau selon le niveau de crue de la rivière . D'autre part , on les associe à des écluses permettant le franchissement, pour la navigation, de la chute crée par le barrage . Pour favoriser une production optimale , il faut donc non seulement étudier duquel on tire différentes édifications mais il faut aussi étudier les hauteurs de chute pour éviter les risques tout en produisant au maximum. Pour montrer ces différentes hauteurs de chutes nous présenterons le tableau ci-après énumérant plusieurs hauteurs selon le type de construction ou de matériaux avec lesquels ils sont édifiés.
deten tion	100	animal	Directives relatives à l'élevage et à la détention d'animaux de compagnie ainsi qu'à l'exploitation des pensions et refuges pour animaux: définition du caractère professionnel, obligation d'annoncer et emploi de gardiens d'animaux.
		provisoir e	Après les mots : « une indemnité », la fin de l'article est ainsi rédigée : « est accordée, à sa demande, à la personne ayant fait l'objet d'une détention provisoire au cours d'une procédure terminée à son égard par une décision de non-lieu, de relaxe ou d'acquittement devenue définitive, afin de réparer le préjudice moral et matériel qu'elle a subi à cette occasion. Toutefois, aucune indemnisation n'est due lorsque cette décision a pour seul fondement la reconnaissance de son irresponsabilité au sens de l'article 122-1 du code pénal, une amnistie postérieure à la mise en détention provisoire, ou lorsque la personne a fait l'objet d'une détention provisoire pour s'être librement et volontairement accusée ou laissé accuser à tort en vue de faire échapper l'auteur des faits aux poursuites. » ;
		arme	Sont punis d'un emprisonnement de 5 à 10 ans et d'une amende de 2.000.000 à 10.000.000 de francs ou de l'une de ces deux peines seulement, ceux qui contreviennent aux dispositions du présent Code relatives à l'offre, la mise en vente, la distribution, le courtage, la vente et la livraison à quelque titre que ce soit, l'envoi, l'expédition, le transport, l'achat, la détention ou l'emploi des drogues à haut risque du Tableau I.
		detenu	Les deux cellules étaient de dimensions plutôt réduites (+ 5 m²), mais néanmoins adéquates pour une détention individuelle de courte durée. Elles étaient équipées d'un bât-flanc en béton, d'un W.-C. asiatique, et matelas et couvertures étaient à la disposition des détenus. L'éclairage naturel - assuré par des blocs de verre translucide - et artificiel était satisfaisant. De plus, les cellules étaient équipées d'un système

			de conditionnement d'air, récemment installé. Elles n'étaient dotées d'aucun système d'appel, mais l'officier de gendarmerie présent a indiqué que des rondes régulières étaient effectuées. Enfin, la délégation a noté l'état de propreté et d'entretien scrupuleux des lieux.
form ation	25	professio nnel	Ce groupe aura pour but de mettre en place une collaboration structurée entre prestataires de formation, employeurs et institutions de façon à favoriser la réactivité du dispositif de formation par rapport aux évolutions des politiques et des pratiques de l'action sociale. Il devra proposer des dispositifs permettant de mieux articuler formations initiales, stages, pratiques de terrain et formations continues autour de l'évolution des compétences et des qualifications. Il devra notamment promouvoir une approche des stages de terrain permettant aux étudiants d'acquérir une vision globale et diversifiée des modalités d'intervention.
lance ment	50	satellite	Avant d'analyser les multiples textes juridiques qui interviennent dans cette construction, un mot sur le CSG, Port Spatial de l'Europe et sur les activités de lancement. Un lancement c'est d'abord un ensemble constitué d'une zone de lancement et d'installations-sol, de préparation du lanceur et des satellites. Le département français de la Guyane a été retenu comme site de lancement en raison d'un certain nombre d'avantages comme la proximité de l'équateur qui permet de profiter de la vitesse de rotation de la Terre et de réduire les manoeuvres en orbite, d'où un gain de poids et un allongement de la durée de vie du satellite; l'ouverture sur l'océan réduit les risques en cas de retombée et permet les lancements sur un large éventail d'azimut (du Nord à l'Est). l'absence de cyclones et de tremblements de terre. Le site retenu en 1964, opérationnel en 1968 a vu d'abord l'installation par le CNESd un pasde tir Diamant et de fusées-sondes puis d'Europa, la fusée développée par le CECLES/ELDO. Le CSG c'est aussi une surface de 900 square chilometres avec 50 km de côtes, un effectif de quelque 1100 personnes.
		produit	Atlantica International la communication par le voyage. Séminaire, événement d'entreprise, séminaires, événements d'entreprise. Seminaires, evenements d'entreprise, seminaire, evenement d'entreprise. Séminaire de direction, séminaire force de vente, séminaires de direction, séminaires force de vente. Lancements de produit, séminaires de team building, lancement de produit, séminaire de team building. Séminaire de cohésion d'équipe, anniversaire d'entreprise, séminaires de cohésion d'équipe, anniversaires d'entreprise. Atlantica International, agence-conseil est spécialisée dans le tourisme d'affaires et l'organisation à la carte de séjours, circuits, voyages et événements d'entreprise (incentives, séminaires, conventions, colloques, lancement de produits, stimulation, team building, anniversaires d'entreprise, etc.) en France et à l'étranger.
orga ne	165	don	En tout et partout, une fois la mort cérébrale déclarée, les équipes médicales ne disposent que de 24 heures pour agir et procéder à toutes les étapes énumérées ci-dessus. Près de 8 heures sont consacrées uniquement aux tests de compatibilité.

			Il faut aussi compter que le donneur et le receveur peuvent se trouver à différents endroits du Québec; il faut donc prévoir du temps de transport pour toute cette démarche. Finalement, le temps de conservation maximal des organes après le prélèvement est de : 4 à 5 heures pour le c ▪ur et les poumons, 12 à 15 heures pour le foie, 6 à 12 heures pour le pancréas et jusqu'à 36 heures pour le rein. Plus l'organe est transplanté rapidement, plus les chances de réussite de l'opération sont élevées.
		representant	Chaque organe de concertation paritaire constitué peut déroger aux règles qui précèdent dans le cadre du protocole.
		consultatif	Citoyenneté européenne Gouvernance européenne Institutions européennes Législation européenne Observatoire du marché unique OMU Comité économique et social européen, CES Organisations non gouvernementales ONG Société civile organisée Organe Consultatif Européen
		parti	Concernant les nombreux titres de la presse sénégalaise, on peut noter une originalité, à savoir l'existence de plusieurs quotidiens : Le Soleil, quotidien gouvernemental, Sud quotidien et Walfadjri. Peu de titres paraissent régulièrement. On note Le Politicien, mensuel satirique, Promotion, bimensuel indépendant, Le Cafard libéré, Le Rénovateur, Le Devoir, Républicain, Faggaru, Xarebi, Le Témoin, Le Soutien, etc. L'Unité pour le socialisme est l'organe mensuel du Parti socialiste et Sopi est le quotidien du Parti démocratique sénégalais.
		technique	Fondée en 1917, la Revue de Musicologie est l'organe de la Société française de musicologie. Elle a pour objectif de rendre compte de l'activité scientifique internationale dans les domaines de la science de la musique, l'histoire des techniques et des langages musicaux, ainsi que de l'étude des conditions sociales et culturelles de la pratique musicale. Ouverte à toutes les périodes de l'histoire, elle ne publie que des articles originaux, des dossiers de synthèse, des comptes rendus de lecture ou de congrès. ISSN : 0035-1601.
		reglement	L'Organe d'appel, constitué par l'Organe de règlement des différends, est un organe permanent composé de sept membres et largement représentatif de la composition de l'OMC. Ses membres sont nommés pour quatre ans. Il doit s'agir de personnes dont l'autorité est reconnue en matière de droit et de commerce international et qui n'ont aucune attache avec une administration nationale.
		maladie	Le coeur: Le coeur est un organe vital dont la tâche est d'assurer la circulation du sang dans l'organisme. Il effectue en permanence un travail de pompe, maintenant une irrigation continue de tous les autres organes.
		deliberant	Notification au maire de la commune de la délibération de l organe délibérant de l établissement public de coopération intercommunale.
passage	210	euro	Les modalités d'application aux systèmes de paiements français des dispositions de la réglementation européenne

			relatives aux conversions et aux arrondis ont été détaillées dans un document (▪Le passage à l'euro, les arrondis, recommandations, mai 1997 ▪) élaboré par la mission interministérielle de préparation des administrations publiques à l'euro, l'AFECEI, le conseil national de la comptabilité et la Banque de France. La profession bancaire a apporté sa contribution à cet accord dans le chapitre "Procédures de conversion" des cahiers des charges.
		niveau	Lorsque, en raison de la situation des lieux, des véhicules ayant initialement circulé en files parallèles sont obligés de poursuivre leur marche sur une chaussée rétrécie sur laquelle une circulation en files parallèles ne s'avère plus possible et que les files de véhicules sont obligées de se confondre, le conducteur circulant le plus à droite bénéficie de la priorité de passage. Dans ces conditions, le fait que les véhicules circulant le plus à droite se portent vers la gauche ne constitue pas une manoeuvre au sens de l'article 12.4 du code de la route.
		parametre	C++, comme Pascal, offre deux modes de passage des arguments à une fonction : par valeur, et par référence. Par contre, C12, Java et CAML, par simplicité n'en offrent qu'un, le passage par valeur, tandis que Fortran fait le choix inverse.
		an_2000	Bien que ces installations soient arrêtées pour le passage à l'an 2000, le redémarrage des installations devra également prendre en compte les risques de perturbation du réseau EDF dus aux variations de charge attendues pour cette période : l'Autorité de sûreté et son appui technique ont ainsi sensibilisé les exploitants sur la vérification des protections des équipements contre les surintensités et la disponibilité des alimentations de secours en cas de déclenchement de ces protections.
		terre	Bien que cette formule soit approximative, on peut en déduire en fonction des valeurs de L0 pour chaque passage, des informations concernant la fréquence des passages observables. Ainsi pour Mercure les passages observés au nud ascendant (L0 = 232.5´) seront environ deux fois plus nombreux que les passages observés au nud descendant de mai (L0 = 109.4´). Pour Vénus les critères sont très proches, donc on peut s'attendre à avoir un peu plus de passages au n ▪ ud descendant qu'au n ▪ ud ascendant. Par contre la valeur de L0 étant très faible les passages de Vénus seront très rares.
		mort	Alors ressaisissons-nous. Ne gâchons pas ce temps précieux qui nous est offert pour aimer. Cela seul peut nous faire expérimenter la Pâque de Jésus, le "passage" ouvert par Jésus. Celui qui, durant sa vie, a su aimer les autres, celui-là n'a d'ailleurs rien à craindre de la mort. Sa vie, qui était déjà passage, l'emmène plus loin. L'aventure continue, autrement sans doute, mais - et c'est bien là l'essentiel - toujours avec Jésus!
		libre	Dans une ordonnance rendue en l'affaire du Passage par le Grand-Belt (Finlande c. Danemark), la Cour a dit, à l'unanimité, que les circonstances, telles qu'elles se présentaient à la Cour, n'étaient pas de nature à exiger

			l'exercice de son pouvoir d'indiquer des mesures conservatoires en vertu de l'article 41 du Statut.
		galerie	Créée en 1826 et aussitôt éclairée au gaz, grande nouveauté, elle attire par la grande splendeur de sa décoration concentrée sur les devantures des boutiques qui allient le bois sombre et le bronze au laiton. Avec le carrelage à damiers, l'alternance des plafonds et des verrières, ces boutiques donnent à la galerie une allure de vestibule précieux. 9. Passage Brady, 46 Faubourg Saint-Denis(métro Strasbourg Saint-Denis, Château d'eau) .
		cheval	L'apprentissage proprement dit est terminé. Votre cheval donne maintenant quelques foulées de passage.
restauration	165	rapide	A côté de la restauration dite traditionnelle et des chaînes de restaurants thématiques existent aussi la restauration collective et la restauration rapide. Dans ce type d'établissements, il n'y a pas de service en salle et la cuisine se fait plus en quantité qu'en qualité ou se résume au conditionnement de plats déjà préparés.
		meuble	Atelier Bois Arts repond a vos besoin en restauration de meubles anciens, decoration, ameublement d interieur, achat et vente de meubles. Villeurbanne.
		pierre	Emile Bonnet, né en 1905 et ayant une certaine expérience de la restauration et la construction d'aménagements et de bâtisses en pierre sèche.
		hebergement	Dans l'Allier en Auvergne le Val de Sioule tourisme hébergement restauration chambres d'hotes auberges hotel restaurant musée gites ruraux activité sportive artisanat manifestation
		conservation	La sauvegarde monumentale, qui se définit aujourd'hui comme une sauvegarde des traces historiques, doit souvent assumer la responsabilité de sacrifier certaines couches ou d'accepter des pertes pour pouvoir assurer la continuité de fonction d'un monument; seule une démarche scientifique permettra de peser les profits et les pertes, de surmonter les problèmes que pose le choix des mesures à prendre, et de remplir les obligations de documentation envers les parties, qu'on se voit obligé d'abandonner. L'article 2 de la charte de Venise précise que «la conservation et la restauration des monuments constituent une discipline qui fait appel à toutes les sciences et à toutes les techniques qui peuvent contribuer à l'étude et à la sauvegarde du patrimoine monumental».
		fichier	La restauration sélective permet de recharger un article ou le contenu d'un fichier dans une base de données déjà existante, depuis une bande de sauvegarde générale, de sauvegarde de md, de sauvegarde incrémentale ou de bande de transaction.
		hotelerie	GRETA FREYMING MERLEBACH: EMPLOYE DE RESTAURATION (NIVEAU V), EN CENTRE DE FORMATION (INTER) - Durée : 975 heures.
solution	75	gestion	IFS Distribution est la base de toute solution de Supply Chain Management. Ce produit est basé sur quatre principes

			fondamentaux : la simplicité, pour une meilleure visualisation des flux de produits et une facilité accrue de l'utilisation du système ; la souplesse, pour s'adapter facilement aux différents modèles de distribution et processus métiers ; l'évolutivité qui permet une montée en charge aisée du système et une meilleure adaptation aux évolutions de l'entreprise ; et enfin, l'ouverture, qui facilite la communication avec les autres acteurs de la chaîne de distribution.
		jeu	La solution ne contient pas toutes les actions facultatives du jeu. Pour les connaître, suis les liens vers la page Trucs et Astuces.
		injectable	SORBITOL DUBERNARD 10% solution injectable (arrêt de commercialisation)
station	200	primagaz	STATION ESSO OUSTRIC S.A.R.L. PRIMAGAZ Aire Carcassonne Arzens Sud - A61 - 11290 MONTREA
		meteo	WS 2300 station Météo Professionnelle PROMO
		spatial	Les troubles qui secouent la Russie risquent d'entraîner la perte du complexe orbital Mir et peut-être même celle de la Station spatiale internationale ISS. Si c'est le cas, de sérieux problèmes sont en vue.
		radio	Le seul moyen d'être en contact avec un satellite en orbite est d'utiliser les ondes radio. Le rôle d'une station de poursuite satellite est d'assurer la connexion radio entre le satellite et son centre de contrôle. Cette liaison est bidirectionnelle (émission et réception) à la fréquence de 2 GHz.
		eau	L'épuration des eaux usées élimine les matières organiques consommatrices d'oxygène grâce aux bactéries qui les assimilent et les utilisent dans la station plutôt que dans la rivière. Pour leur besoin, notamment pour la synthèse de protéines, elles assimilent également de l'azote et du phosphore mais en quantité insuffisante, pour qu'une fois extraites de la station sous forme de boues, la diminution de ces substances soit suffisante.
		ski	Situés dans la Vallée de Chamonix Mont-Blanc, Les Houches, village et station sportive, et Servoz, village de montagne, vous souhaitent la bienvenue au coeur du site prestigieux du massif du Mont-Blanc. En été comme en hiver, toutes les activités de montagne à votre portée : ski, snowboard, randonnée, escalade, eaux-vives et de nombreuses animations pour les enfants comme pour les grands.
		ligne	Porte de Versailles (ancien terminus ligne A du Nord-Sud) - station oubliée .
		travail	Pour les débutants, la station de travail Sun Blade 100 permet de développer et de déployer des logiciels dans un environnement 64 bits puissant pour un prix d'entrée de gamme intéressant.
vol	100	libre	Hanggliding, Delta, hang gliding, deltaplane, para gliding, paragliding, vol libre, aile delta, parapente,voler, fly, wing, wings, aile, delta, flying, ixbo, IXBO,rigid wing, rigide, aile

			rigide, performance, pilot, record, price, security, innovation,tecma, malinjoud, french, france, France, français, constructeur, vol libre, aviation
		avion	Billets d'avion et vols : comparez et trouvez des tarifs discount pour votre billet d'avion / prix de votre vol pour Madagascar et 400 autres destinations. Réservation en ligne. (billet degriffé, promotions, tarifs négociés) .
		vole	Cette base de données est en libre accès (pour consultation seulement) sur nos différents services télématiques (3615 Photim, Kiosque-Micro et www.photim.com via Internet). Les listes de matériel volé sont également transmises aux différents réparateurs et services après-vente des grandes marques, dans l'espoir que ces ateliers puissent, éventuellement, "arrêter" un matériel suspect s'il revient en réparation.
		voile	cipvvs Centre de Vol à Voile de Strasbourg