Automatic Acquisition of Lexical Knowledge from MRDs

Tesi doctoral presentada al
Departament de Llenguatges i Sistemes Informàtics
de la Universitat Politècnica de catalunya

per optar al grau de
Doctor en Informàtica

per

German Rigau Claramunt
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3 08034 Barcelona. Catalonia
g.rigau@lsi.upc.es

sota la direcció del doctor
Horacio Rodríguez Hontoria

Barcelona, Maig 1998

Acknowledgments

This thesis wouldn't have been possible without the aid and collaboration of many people whom I wish to thank.

The hardest and longest task has been carried out by Horacio Rodríguez. His devotion and patience has gone far beyond what one expects from an advisor, a colleague or a friend.

A very special mention is reserved to Eneko Agirre for his contributions to this research. Working with him has been always productive and fun.

Additionally, I would like to express my gratitude to the (current and past) members of the Natural Language Research Group. I would like specially mention the persons who introduce me in the Artificial Inteligence Section: Núria Castell and Felisa Verdejo. And the linguistic team: Toni Martí, Irene Castellón, Mariona Taulé and Salvador Climent. And the software team: Alicia Ageno, Francesc Ribas, Lluís Padró, Lluís Marquez, Jordi Atserias, Jordi Turmo, Toni Tuells, Xavier Farreres, Gerard Escudero and Laura Benítez.

I also want to thank the Software Department at the UPC for facilitating my research. And the Computer Research Laboratory at NMSU where I performed a stay of three months during 1994. I would like specially mention: David Farwell, Jim Cowie and Louise Guthrie.

This research couldn't be possible without the lexical resources used. The monolingual and bilingual dictionaries provided by Biblograf and WordNet provided by the University of Princeton.

I am also in gratitude with our colleagues of the European projects. Specially with Ted Briscoe, Ann Copestake, John Carroll and Piek Vossen.

Above all, my deepest gratitude and love go to my fathers and brothers. And Montse, no one has sacrified more and complained less.

The author has been also supported by a grant awarded by the Ministry of Education and Science, 92-BOE-16392 and the work has also been partially funded by the following projects and initialives:

Acquilex-I (Esprit BRA 3030) UE project
Acquilex-II (Esprit BRA 7315) UE project
EuroWordNet (LE4003) UE project
ITEM (TIC96-1243-C03-03) Spanish Department of Research project.
CIRIT. Grup de Recerca de Qualitat 1995SGR-00566

Index

Summary of Contents

1. Introduction and Motivation

This chapter is devoted to motivate the work done on acquiring lexical knowledge from Machine-Readable Dictionaries (MRDs). It also introduces the methodology followed in this thesis for the automatic construction of a large and highly structured multilingual lexical knowledge base (MLKB) directly from monolingual and bilingual MRDs. After the first section, which presents the framework of this thesis, Section 2 is devoted to explaining the main facets related to the construction of massive lexicons that are useful for Natural Language Processing (NLP). Section 3 focuses on the construction of such lexicons using MRDs. Section 4 introduces, with a brief example, the methodology used by the *Sistema d'Extracció d'Informació Semàntica de Diccionaris* (SEISD, System for Extraction of Semantic Information from Dictionaries) and the subsystems it contains. Finally, Section 5 overviews the remaining chapters.

2. Words and Works

This chapter summarizes and discusses the main problems we have to face during lexical acquisition tasks. Different approaches related with such problems are described and the main results are presented. This takes the form of an in-depth study of the different lexical acquisition approaches, methodologies and experiments appearing in the literature, i.e.: a) what information/knowledge is needed in the lexicon? (Section 2), b) where is this information/knowledge located? (Section 3) and c) what procedures can be applied to extract this information/knowledge from its sources? (Section 4). In Section 5, we perform an in-depth study of lexical knowledge acquisition from MRDs. Section 6 studies the work on lexical knowledge acquisition from corpora and Section 7 on mixing structured and not structured resources. Finally, Section 8 accounts for the main international projects currently existing in the field of lexical acquisition.

3. The Methodology and SEISD

The purpose of this chapter is to describe the general methodology for creating a Multilingual Lexical Knowledge Base (MLKB) from monolingual and bilingual MRDs and to present SEISD, the software system that supports this methodology. After introducing, in Section 2, the main methodological considerations, in Section 3 the main objectives of the SEISD are explained. The components of the environment are briefly described in Section 4 and Section 5. Section 6 is presents the whole acquisition process using SEISD. This Section is divided also in four subsections. Subsection 6.1 is devoted to the semantic knowledge acquisition process and Subsection 6.2 to the mapping process of the acquired knowledge onto the Lexical Knowledge Base (LKB). In Subsection 6.3 the multilingual knowledge acquisition task is presented and finally, Subsection 6.4 describes the exploitation and validation process of the acquired lexical knowledge.

4. Main Issues of the Acquisition Process

The basic aim of this chapter is to describe briefly the main problems faced by SEISD in order to perform the acquisition of lexical knowledge from MRDs. Thus, it provides an in depth study of those subtasks SEISD is deal with. Section 2 explains different methodological approaches for classifying the concepts described within an MRD. Section 3 is devoted to several approaches for the construction of taxonomies from a monolingual MRD. Section 4 deals with the extraction of the main semantic relations from the dictionary definitions and their mapping onto a Lexical Knowledge Base (LKB). Section 5 focuses on the construction of multilingual Lexical Knowledge Bases and finally, Section 6 is devoted to the main mechanisms used for the validation and exploitation of the multilingual LKB.

5. Monolingual Lexical Knowledge Acquisition

This chapter covers the main experiments and results concerning the acquisition of lexical knowledge by using SEISD on the monolingual dictionary *Diccionario General Ilustrado de la Lengua Española* (DGILE). After a short introduction, Section 2 deals with the automatic selection of the main semantic primitives present in DGILE. Section 3 is devoted to the automatic acquisition of taxonomies from DGILE, and Section 4 describes the work done on the automatic acquisition of knowledge from the definitions contained in DGILE.

6. Multilingual Lexical Knowledge Acquisition

The purpose of this chapter is to present the work carried out for the automatic construction of the multilingual facet of the LKB. While Section 2 presents the complete framework and resources used by SEISD for linking lexical entries across languages, Section 3 shows the main techniques and results, applying this process by attaching Spanish lexical units to English ones.

7. Conclusions and Further Work

This chapter summarizes the work presented in this thesis, and also the results obtained. Thus, in Section 2 the main goals achieved during this work are shown. Section 3 lists the main lexical resources acquired from the MRDs during the work presented here and, at the end, Section 4 describe the further work we are planning to do.

Dictionaries

References

Appendix

Chapter 1

Introduction and Motivation

1.1 Setting

The automatic acquisition of knowledge, a central issue in artificial intelligence, is the main field of the research work presented here. In particular, the thesis deals with the acquisition of lexical and conceptual knowledge, a topic of increasing importance within Computational Linguistics (CL), Computational Lexicography (CLX) and Natural Language Processing (NLP).

In recent years many NLP systems have reach the level of industrial products such as Machine Translation (MT), Text Comprehension, Text Summarization, Information Retrieval (IR) or Natural Language Interfaces, leading to a significant rise in the need for linguistic resources. Knowledge of and about words plays a central role in all these applications. Thus, the lexicon[1], which represent lexical information reliably and precisely enough for automated use, has become the focus of a great deal of research in CL theory, CLX and NLP. There are both theoretical and practical reasons for this trend.

On the theoretical side, most current linguistic theories (perhaps starting with [Chomsky 70]) grant the lexicon a much larger role than before (e.g., Word Grammar [Hudson 84], Generalized Phrase Structure Grammar (GPSG) [Gazdar et al. 85] and Head-driven Phrase Structure Grammar (HPSG) [Pollard & Sag 94]). Much of the knowledge that in the classical theory resided in grammar rules now had to be based on the lexicon. Thus, the lexical dimension of the recent linguistic theories has grown.

On the practical side, as NLP systems become more sophisticated and potentially able to make the transition from laboratories to industry, the need for large lexicons becomes more pressing. The lexicon is recognized as one of the majors problems in NLP applications both because of the need for substantial vocabulary in habitable NLP systems and because of the increasing complexity. The term "lexical bottleneck" [Briscoe 91] describes the problem the lack of lexical resources causes in existing NLP technologies, and the problems of getting such resources into NLP systems. The task of constructing realistic lexicons for natural languages is formidable because of the enormous amount of words and knowledge to be dealt with. There are many words and many distinct types of information about words potentially relevant to different kinds of NLP tasks. Furthermore, the total amount of useful lexical resources for NLP is not the same for different languages. While for English several large-scale lexicons are available (e.g., WordNet [Miller 90], Alvey Lexicon [Grover et al. 93], Comlex [Grishman et al. 94], etc.) there are few Spanish wide-range lexicons available for NLP. The work presented here attempts to lay out some solutions to overcome or alleviate these problems.

The setting of the thesis thus having been determined, the next section deals with the aspects we need to take into account for the lexical acquisition. Section 1.3 focuses the lexical acquisition problem on Machine-Readable Dictionaries (MRDs), possibly one of the most

[1] *As in [Wilks et al. 96] by "lexicon" we mean a set of formalized entries to be used in conjunction with computer programs, and by "dictionary" the physical printed text giving lexical information, including meaning descriptions.*

useful on-line lexical resource available. Section 1.4 gives a brief overview of the methodology we have applied in this thesis and, finally, Section 1.5 outlines the thesis.

1.2 On lexical acquisition

In order to deal with the problem of lexical acquisition three central questions must be answered: a) what lexical information/knowledge is needed for a concrete NLP system? b) where is this information/knowledge located? and c) which procedure can be applied to extract/handle this information/knowledge from its sources? These three main questions, namely the information needed by the lexicon, the possible sources of lexical knowledge and the possible automatic methodologies (i.e., using a computer) that can be applied, frame the work presented in this thesis.

1.2.1 Information needed by the lexicon

The linguistic and conceptual information associated with each lexical entry placed in the lexicon depends on the NLP system. For each different NLP task, different information attached to each lexical entry is requested. Usually, the lexicon is used:

- to obtain the morphological inflexion, composition or derivation patterns.
- to assign morphologic, syntactic, semantic or pragmatic properties.
- to assign the (simple or complex) syntactic category.
- to obtain the possible translations.
- to obtain statistical properties (e.g., frequency, cooccurrence patterns, etc.).

At least six broad types of information which are potentially relevant to NLP systems might be placed in the lexicon:

- phonology: phonemes, stress, etc.
- morphology: parts of speech, cooncordance patterns, etc.
- syntax: syntactic category, subcategorization, predicate/argument structure, valences, cooccurrence patterns, etc.
- semantics: semantic class, properties of the class, selectional restrictions, etc.
- pragmatics: usage, registers, topic domains, etc.
- Translation Links: the architecture of the Machine Translation system determines the transfer level. A lexical driving transfer mechanism needs a complete different source and target lexicon than a more conceptual (close to the interlingua) Machine Translation system.

Obviously, a specific lexicon for a specific NLP system does not need all this information. The lexical information needed for a spell-checker system may be completely different from, for instance, a Natural Language Interface system. And, of course, most of the information that could be attached to a lexical entry depends on the part(s) of speech of that entry. For example, selectional restrictions are usually attached to verbs, while number can be attached to nouns, pronouns or verbs, etc. The duality between the lexical/conceptual information that most NLP systems need is an important issue to be borne in mind.

1.2.2 Sources of lexical information

Three main sources of information for building wide-coverage lexicons for NLP systems can be considered:

• **Introspection**, i.e., constructing the lexicon using the knowledge about the language and the world that the human builder of the NLP system owns (e.g., Linguistic String Project [Sager 81], CYC [Lenat & Guha 90], EDR [Uchida 90] and WordNet [Miller 1990][1]) .

• **Structured lexical resources** such as conventional monolingual and bilingual dictionaries form an excellent starting point for building substantial lexicons because they constitute a highly structured and relevant source of information about words and meanings (e.g., [Amsler 81], [Boguraev & Briscoe 89a], [Dolan et al. 93], [Wilks et al. 96], [Richardson 97]). Thesauri (e.g., [Yarowsky 92] or [Grefenstette 93]), encyclopaedias (e.g., [Gomez et al. 94]) or other lexical resources for human use could also be considered.

• **Unstructured lexical resources** such as corpora provide an additional though less organized source, relating to issues of usage, such as the relative frequency of word senses or the range and frequency of different patterns of syntactic realization (e.g., [Church & Hanks 90] and [Zernik 91]).

As could be expected, it is not realistic to obtain all the information needed for a lexicon from only one source. Consequently, these sources are often used in combination (e.g., [Carroll & Grover 89], [Grishman et al. 94], [Knight & Luk 94], [Ribas 94] or [Klavans & Tzoukermann 96]).

1.2.3 Methods of lexical acquisition

The literature shows two main alternative approaches to the lexical acquisition process: the prescriptive approach and the descriptive approach. In the prescriptive approach, a set of primitives is defined, or prescribed, prior to or in the course of designing and developing the whole system. The descriptive approach, on the other hand, allows a natural set of primitives derived from a natural source of data without any preexisting frame.

From the point of view of human intervention, the information attached to each lexical entry can be obtained by manual, automatic or semi-automatic approaches depending on the methods applied, sources used and the information needed for a particular application.

Three major approaches to lexical acquisition have been developed: machine-aided manual construction, (semi)automatic extraction from preexisting lexical resources and the combination of both.

1.3 Lexical acquisition from MRDs

One reason why the lexical capabilities of NLP systems has remained weak is because of the labour intensive nature of encoding lexical entries for the lexicon. It has been estimated that the average time needed to construct a lexical entry for a NLP system by hand is about 30 minutes [Neff et al. 93]. If we assume that the task of developing an adequate "core" lexicon is equivalent to that of developing a conventional advanced learners dictionary (containing typically between 40,000 and 50,000 entries), then the labour runs into tens of persons/year.

An interesting approach might be to take advantage of preexisting lexical resources. Dictionaries are texts whose subject matter is vocabulary and meaning. Machine-Readable Dictionaries (MRDs), the conventional dictionaries for human use on a computer support, usually "contains spelling, pronunciation, hyphenation, capitalization, usage notes for semantic domains, geographic regions, and propriety; etymological, syntactic and semantic information about the most basic units of the language" [Amsler 81]. In addition, the words are described in terms of senses (lexical concepts), and the concepts are described in terms of words.

This thesis will focus on the massive acquisition of lexical knowledge from MRDs using automatic methodologies. That is, considering the three factors mentioned in Section 1.2., we will show a) the different kinds of information that can be extracted from b) structured

[1] *See an overview of CYC, WordNet and EDR, with comments from the authors, in [ACM 95].*

lexical resources such as monolingual and bilingual MRDs by applying c) automatic procedures.

It is clear that different dictionaries do not contain the same explicit information. Despite this, we will prove in the course of this thesis that any conventional dictionary contains a great amount of implicit lexical knowledge that is useful for NLP tasks and can be extracted using automatic approaches. Compare, for instance, the following *Diccionario General de la Lengua Española VOX* (DGILE[1]) and *Longman Dictionary of Contemporary English* (LDOCE[2]) "lispified" entries corresponding to "paella":

```
((paella )
(ETIM cat., sartén; V. padilla )
(Sense 1)
(CA   f.)
(DEF  Plato de arroz seco, con carne, legumbres, etc., muy usado en la región valenciana.)
(Sense 2)
(CA   f.)
(DEF  Sartén donde se hace dicho plato.)
)

((paella )
(HN   0)
(SN   0)
(PS   n)
(GC   U)
(PC   FO--)
(SC   5)
(DF   rice cooked with pieces of meat, fish, and vegetables in, esp. in Spain)
)
```

These lexical entries are in Lexical Data Base form. That is, each piece of information contained in a lexical entry have been split into different labelled fields. For the DGILE lexical entry, ETIM stands for etymology, CA for part of speech and DEF for definition. The Spanish lexical entry contains morphological information (the two senses of "paella" have the code "female noun" as part of speech), no syntactic information coded and no explicit semantic information is provided. In addition, some other explicit information can be found in different DGILE lexical entries, such as flection, compounds, semantic relations (e.g., synonymy, antonymy, etc.), domain codes (e.g., music, military, etc.), geographical codes, usage codes, etc.

For the LDOCE lexical entry, HN stands for homonym number, SN for sense number, PS for part of speech (noun), GC for grammatical code (collective), PC for pragmatic code (primary code FOOD), SC for semantic code (organic materials) and DF for definition (using a predefined defining vocabulary of 2,000 words). Verbal entries in LDOCE contains other kinds of information such as subject preference, typical object, indirect object preference, etc.

It soon becomes clear that the amount of coded information per entry in LDOCE is greater than in DGILE[3]. This explicit information in LDOCE makes it easy to extract other implicit information (i.e., taxonomies [Bruce et al. 92]). Does this mean that only highly structured dictionaries such as LDOCE are suitable to be exploited to provide lexical resources for NLP systems? This thesis seeks to show we can extract, by means of automatic procedures, useful explicit and implicit information for NLP systems from any conventional dictionary (that is, with no explicit semantic codification).

---

[1] *A study of the information content of the DGILE dictionary can be found in [Castellón et al. 91].*

[2] *A detailed study of LDOCE dictionary information can be seen in [Boguraev & Briscoe 89] or [Wilks et al. 96].*

[3] *For instance, 44% of LDOCE senses contain pragmatic codes.*

Obviously, explicit information can be directly used to construct ad-hoc lexicons. For instance, inflection or part of speech in MRDs can be used straightforwardly as a lexical component of simple morphological analysers[1]. LDOCE grammar codes have been extensively used as lexical information for parsing systems (i.e., [Boguraev & Briscoe 89b], [Carroll & Grover 89], [Sanfillippo 94]). LDOCE semantic and pragmatic codes have been used to assist in the automatic extraction of implicit semantic information (i.e., [Copestake 90], [Bruce et al. 92]).

Many researchers believe that for effective NLP it is necessary to build a Lexical Knowledge Base which includes taxonomic information. This Lexical Knowledge Base should contain facts such as specializations (class/subclass relations) or instantiation (class/instance relations) and mechanisms for the inheritance of properties and other inferences. It is clear that monolingual MRDs contain knowledge about the language and knowledge about the world that is essential for NLP systems (e.g., [Byrd 89], [Vossen et al. 89], [Wilks et al. 93], [Dolan et al. 93], [Kilgarriff 93], [Wilks et al. 96], [Guthrie et al. 96], [Richardson 97]).

However, as is soon realized, an MRD does not offers an immediately usable resource as a computational lexicon. Dictionaries are usually built for human use, and not for machine use. Some researchers have concluded that dictionaries are inadequate as a source of semantic information to serve as the Lexical Knowledge Base for sophisticated semantic processing (i.e., [Walker & Amsler 86] and [Atkins et al. 86]). Definitions frequently fail to express even basic facts about word meanings. Consider, for instance, the lexical entry *flor_1_1* [2] (flower) in DGILE:

> **flor_1_1** Órgano complejo de la reproducción sexual en las plantas fanerógamas, procedente de la evolución de las hojas de un brote, ... (*literally, complex organ for sexual reproduction in phanerogamic plants, originating from the evolution of the leaves in a bud, ...*)

This definition lacks any detailed description of the physical structure of flowers, information about instances of flowers, and so on.

Given a dictionary in book form, i.e., for human use, the only way to find information about a given word is to look it up, then explore the semantic content of any words mentioned in its entry, and so on. This strategy could be called the forward-chaining model of dictionary consultation. Another possible strategy is be the backward-chaining model, that is, when looking up a word, consulting not only its own definition but also the definitions of any word which mentions it. The important point here is that much of the information about a given word's meaning is typically located not in the entry for that word itself, but rather in the entries for other words.

This possibility was noted, but not implemented, by [Amsler 81] and exploited for the first time by [Chodorow et al. 85], who found semantic links between the different lexical components of the *Webster's Seventh New Collegiate Dictionary* (W7N). These semantic links between words produce a huge and highly interconnected network of concepts linked by arcs labelled with semantic relations such as case relations (agent, patient, recipient, time, location, goal, cause, purpose, etc.), class/membership relations (hypernymy and hyponymy), part/whole relations (part-of, member-of, substance-of, etc.), which can be used as a Lexical Knowledge Base (LKB) of use for NLP systems (e.g., [Jensen & Binot 87], [Fox et al. 88], [Byrd 89], [Vanderwende 95]).

Let us consider this approach in more detail. Searching DGILE for entries which mention "flor" in their definitions allows us to construct a highly detailed picture of what a flower is: where we can usually find flowers, which are the parts of the physical structure of flowers, the fact that bees collect nectar from them, the places where they are sold, the people who sell flowers, a list of different instances of flowers, etc.

---

[1] *SegWord (see Section 3.5.1) and MACO [Acebo et al. 94] use a lexicon derived from Spanish MRDs.*

[2] *i.e., the first sense of the first homonym of the headword "flor" in the DGILE dictionary.*

Places where flowers are found:

jardín_1_1 Terreno donde se cultivan plantas y **flores** ornamentales. (*garden: extension of land where plants and ornamental flowers are grown*).
florero_1_4 Maceta con **flores**. (*vase: pot with flowers*).
ramo_1_3 Conjunto natural o artificial de **flores**, ramas o hierbas. (*bouquet: natural or artificial set of flowers, branches or herbaceous plants*).

Parts of flowers:

pétalo_1_1 Hoja que forma la corola de la **flor**. (*petal: leaf that forms the corolla of the flower*).
tálamo_1_3 Receptáculo de la **flor**. (*thalamus: receptacle of the flower*).
néctar_1_3 Líquido azucarado que contienen ciertas **flores**. (*nectar: sweet liquid contained in some flowers*).
polen_1_1 Polvillo fecundante contenido en la antera de los estambres de las **flores**. (*pollen: fecundated dust contained in theanthers on the flower's stamens*).

Products of flowers:

miel_1_1 Substancia viscosa y muy dulce que elaboran las abejas, en una distensión del esófago, con el jugo de las **flores** y luego depositan en las celdillas de sus panales. (*honey: viscous, very sweet substance produced by bees, in a distension of the oesophagus, with the juice of flowers and then deposited in the cells of their honeycombs*).

Place where flowers are sold and people who sell flowers:

florería_1_1 Floristería; tienda o puesto donde se venden **flores**. (*florist's shop: florist's; shop or stall where flowers are sold*).
florista_1_1 Persona que tiene por oficio hacer o vender **flores**. (*florist: person whose job is to make or sell flowers*).
floricultor_1_1 Persona que tiene por oficio cultivar las **flores**. (*floriculturist: person whose job is to grow flowers*).

Kinds of flowers:

camelia_1_1 Arbusto cameliáceo de jardín, originario de Oriente, de hojas perennes y lustrosas, y **flores** grandes, blancas, rojas o rosadas (Camellia japonica). (*camellia: camelliaceous garden shrub, native to Asia, having perennial glossy leaves and large white, red or pink flowers*).
camelia_1_2 **Flor** de este arbusto. (*camellia: flower of this shrub*).
rosa_1_1 **Flor** del rosal. (*rose: flower of the rosebush*).
orquídea_1_2 **Flor** de una planta orquidácea. (*orchid: flower of an orchidaceous plant*).

Consider furthermore the following verb definitions:

abrir_1_22 Salir en las **flores** [los pétalos del capullo] (*open: to come out in the flowers [the petals from the bud]*)
cerrar_1_19 Juntar las **flores** sus pétalos (*close: to join the petals of the flowers*)
florecer_1_1 Echar **flor**. (*bloom: to flower*)
florar_1_1 Dar **flor** una planta, florecer. (*bloom: to flower a plant*)
nacer_1_3 aparecer las hojas, **flores**, frutos o brotes en la planta. (*born: to appear the leaves, flowers or sprouts of the plant*)

pecorear_1_2 Salir las abejas a recoger el néctar de las **flores.** (*collect: to go out the bees to collect the nectar from the flowers*)
romper_1_16 Abrirse las **flores**. (*start: to open the flowers*)

were the typical actions related to flowers are described (ways the flowers grow, etc.).

Consider moreover the following verb definitions not related to flowers:

barrer_1_1 **limpiar** (el suelo) con la escoba (*sweep: to clean (a floor) with a broom*).
freír_1_1 **cocer** (un manjar) en aceite o grasa hirviendo. (*fry: to cook (a food) in boiling oil or lard*).
comprar_1_1 **adquirir** (una cosa) a cambio de cierta cantidad de dinero. (*buy: to purchase (a thing) in exchange for a certain amount of money*).
cazar_1_1 **buscar** o **perseguir** (a las aves, fieras, etc.) para cogerlas o matarlas. (*hunt: to search for or pursue (birds, wild animals, etc.) in order to catch or kill them.*)

where the typical objects (floor, dish, thing and animals), typical instruments (broom, oil and money) and a purpose (to take or kill) of some verbs are provided.

But all this great amount of knowledge collected for the use of human readers cannot be used directly for NLP tasks. First, it is necessary to extract the implicit information contained in the MRD non-systematically and represent this information formally and explicitly for future use in NLP systems.

Bilingual MRDs may contain phonetic, morphologic, syntactic and semantic knowledge, equivalent lexical translations, examples of usage, etc. Consider, for instance, the following two entries from the Spanish/English and English/Spanish bilingual dictionaries.

flor *f* flower. **2** (piropo) compliment: • **a ~ de piel**, skin-deep; *fig* **en la ~ de la vida**, in the prime of life; *fig* **la ~ y nata**, the cream (of society).

flower ['flaw@r] *n* flor *f*. - **2** *i* florecer • **~ bed**, parterre *m*.

Thus, the main goal of this thesis is not to demonstrate that an NLP lexicon can be built by collecting implicit and explicit lexical knowledge from monolingual and bilingual MRDs (many examples have proved this statement) but to show that this process can be done, even for less structured and coded MRDs, with comparable results. In order to achieve this goal, a complete methodology is proposed, and an integral system which supports the methodology is provided (SEISD, *Sistema d'Extraccció d'Informació Semàntica de Diccionaris*). Although the system implements automatic procedures for extracting lexical knowledge from any conventional MRD, we perform the whole process using DGILE.

1.4 Brief overview of the thesis

The work for this thesis has been carried out in the framework of the Acquilex-I (BRA 3030), Acquilex-II (BRA 7315) Esprit and EuroWordNet (LE 4003) projects. The major goal undertaken by these projects has been the construction of substantial Multilingual Lexical Knowledge Bases (MLKBs) from preexisting texts (MRDs and corpora) for use in NLP systems. The main aim of this thesis is to develop computational methods and techniques to allow the acquisition of lexical and semantic knowledge from MRDs.

In particular, this thesis provides a complete methodology and a whole system for extracting explicit and implicit information from monolingual and bilingual dictionaries to construct an MLKB using several preexisting structured lexical knowledge resources. **SEISD** environment (*Sistema d'Extracció d'Informació Semàntica de Diccionaris*) (see [Ageno et al. 91b], [Ageno et al. 92a] and [Ageno et al. 92b]), the system which implements the methodology, has been applied to extract implicit semantic information from the

monolingual Spanish dictionary DGILE. This system integrates the two representational formalisms used within Acquilex, i.e., the LDB and the LKB. The LDB [Carroll 92] is a database-like system which provides flexible access to dictionary entries via any of the information contained in the MRD, and the LKB [Copestake 92b] is a system developed to represent lexical entries by means of typed Feature Structures constrained by a Type System.

The purpose of this section is to provide an overview of the general methodology for creating an MLKB from MRDs and to present SEISD, the software system we implemented that supports this methodology. Thus, this thesis has two main objectives:

- Developing a methodology for extracting explicit and implicit knowledge from MRDs in order to build a MLKB.
- Applying this methodology to a huge, lossely structured, Spanish monolingual dictionary. Thus, another complementary objective has been the design and construction of the software environment for supporting the methodology.

Briefly, the approach applied to build an MLKB is the following. First, the monolingual and bilingual MRDs are loaded into a standardized Lexical Data Base (LDB). Once an MRD is transformed into a Machine-Tractable Dictionary (MTD) [Wilks et al. 96] and placed in a Lexical Data Base (LDB) the previously explained dictionary access strategies can be exploited. Secondly, separate semantic taxonomies are derived from these monolingual LDBs to create monolingual LKBs. Using bilingual LDBs, the monolingual LKBs are then linked together in order to create an MLKB. Figure 1.1 depicts the approach for constructing an MLKB from monolingual MRDs.

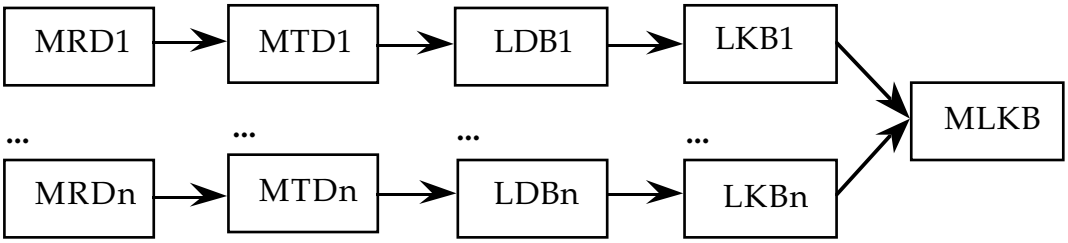


Figure 1.1, general approach for building MLKBs from monolingual MRDs.

A computer environment has been developed to support this methodology. The **SEISD** environment is a modular and interactive system for extracting semantic information from LDBs and providing ways to represent and exploit this information in an MLKB. The environment allows the lexical information to be processed in an incremental, interactive and (semi)automatic way[1], supplying lexicographers with complementary information to assist in the extraction process. Several lexical knowledge resources, some of them external and others derived from the MRD itself, are used for this purpose.

Four main issues were considered for designing the base methodology: the characteristics of the lexical resources used, the information to be extracted from them, how to carry out the process and how to represent and exploit the information extracted (see Section 3.2).

Although several lexical knowledge resources are involved in this methodology, the main one is the monolingual Spanish MRD *Diccionario General Ilustrado de la Lengua Española VOX* (DGILE). Most of the contents of the final methodology presented in this work are due to the specific features of this dictionary.

The most important semantic relation to be extracted from an MRD is the hypo/hypernym relation between senses. That is, the subclass-superclass or ISA relation. This implicit relation arises from the inherent structure of most usual dictionaries that allow us to construct concept taxonomies from dictionary definitions.

The organization of a lexicon in the form of a hierarchy offers several advantages as regards efficiency of information storage. What is stated for concepts at the highest levels

[1] *A minimal set of grammars, ontologies (the Type System), conversion rules, etc. must be provided by lexicographers.*

can be inherited by senses at more specific levels. In this way, information stored once is distributed to a whole class of concepts.

Thus, the hypo/hypernym relation between dictionary senses is not only important because it can be used as a backbone of taxonomies, but also because this relation acts as a support for the main inheritance mechanisms, thus facilitating the acquisition of other relations and semantic features [Cohen & Loiselle 88], providing formal structure and avoiding redundancy in the lexicon [Briscoe et al. 90]. Most of the effort reported in this work deals with this issue (see Sections 4.3 and 5.3).

From the point of view of sources of lexical information, rather than a single one, our methodology takes advantage of many sources of information. Although some parts of our methodology require introspection (mainly during the construction of the top ontology, that is, the Type System of the LKB) most of the lexical knowledge is acquired from structured lexical resources (monolingual or bilingual dictionaries).

Thus, rather than a purely descriptive or prescriptive approach we propose a combined strategy. Although some parts of our methodology require human intervention (mainly during the construction of the Type System of the LKB) the lexicographer can select the appropriate degree of interaction with the system, ranging from none (fully automatic but less accurate) to complete (manual and computer-aided, with a maximum degree of accuracy).

Consider, as an illustration of the acquisition process performed by **SEISD**, the lexical entry *fabada_1_1*, belonging to the taxonomy *alimento* (food). The methodology is divided into six partial steps.

1) First, the top dictionary senses that cover the semantic classes to be represented in the LKB are selected and assigned to the appropriate semantic type (see Sections 4.2 and 5.2 for a detailed discussion of this issue). At this stage, for the taxonomy derived from *caldo_1_1* (broth) the dictionary sense root is attached to the **c_art_subst** (comestible-artifact-substance) type[1].

Word sense:	caldo_1_1
Attached-to:	**c_art_subst** type.
Definition:	**alimento** que resulta de cocer en agua la vianda (*liquid which results from cooking food in water*).

2) Exploiting the implicit hypo/hypernym relation, the sense disambiguated taxonomies are generated, collected, validated and attached to the same semantic class represented in the Type System. This task is performed by the **TaxBuild** (Taxonomy Builder) module of SEISD (see Sections 3.6.1.2, 3.6.1.3, 4.3 and 5.3). An example of a hyponym dictionary sense of *caldo_1_1* following the hyponym chain through *potaje_1_1* is *fabada_1_1*:

Word sense:	potaje_1_1
Hyponym-of:	caldo_1_1
Definition:	**caldo** de olla u otro guisado (*pot broth or other stew*).
FPar:	((CLASS CALDO))

Word sense:	fabada_1_1
Hyponym-of:	potaje_1_1
Definition:	**potaje** de judías con tocino y morcilla, que se usa en asturias (*stew made with beans, lard and black pudding used in Asturias*).
FPar:	((CLASS POTAJE))

3) For each semantic class, the different conceptual nodes attached to it are processed in order to obtain deeper knowledge of the case roles relations and content relations appearing in the differentiae. This process is carried out by the **SemBuild** (Semantic Builder) of SEISD (see Sections 3.6.1.4, 4.4 and 5.4). For instance, the definition of *fabada_1_1* gives a meaning

---

[1] *See Section 3.5.2 for a description of the Type System supporting the LKB.*

of the word *fabada* as a kind of stew (the genus term) but different from the other stews because it is made with beans, lard and black pudding (the differentiae).

Word sense: caldo_1_1
Hyponym-of: bebida_1_1
Definition: **líquido** que resulta de cocer en agua la vianda (*liquid which results from cooking food in water*).
SinPar: [SN: [n: líquido],
SW: [p0r: que],
SV: [v0v: resultar],
SP: [r0d: de,
SV: [i0v: cocer]],
SP: [r0p: en,
SN: [n: agua]],
SN: [n: vianda]].

Word sense: potaje_1_1
Hyponym-of: caldo_1_1
Definition: **caldo** de olla u otro guisado (*pot broth or other stew*).
SinPar: [SN: [n: caldo,
SP: [r0d: de,
SN: [n: olla,
n: guisado]]]].

Word sense: fabada_1_1
Hyponym-of: potaje_1_1
Definition: **potaje** de judías con tocino y morcilla, que se usa en Asturias (*stew made with beans, lard and black pudding in Asturias*).
SinPar: [SN: [n: potaje,
SP: [r0d: de,
SN: [n: judía]]],
SP: [r0p: con,
SN: [n: tocino,
n: morcilla]],
ORIGIN: [w: asturias]].

4) This enriched taxonomy is then represented in the LKB formalism in order to exploit the inheritance and other inferential mechanisms that make explicit, for instance, the inherited properties of the hypernym lexical entries. This task is performed mainly by the **CRS** (Conversion Rule System) of SEISD (see Sections 3.6.2, 4.4.2 and 5.4). In our example, when the analysed *fabada_1_1* is placed as a lexical entry into the LKB lexicon all the local and inherited information acquired (or represented in the Type System) is available. That is, the special ingredients of *fabada_1_1* cooked in water (from *caldo_1_1*) are explicitly represented in the qualia structure of *fabada_1_1*.

fabada_x_1_1
< lex-noun-sign rqs > < potaje_X_I_1< lex-noun-sign rqs >
< lex-sign sense-id : sense-id dictionary > = ("VOX")
< lex-sign sense-id : sense-id word > = (“fabada”)
< lex-sign sense-id : sense-id homonym-no > = (“1”)
< lex-sign sense-id : sense-id sense-no > = (“1”)
< rqs : constituency > = (“judía”, “tocino”, “morcilla”).
< rqs : origin-area > = (“asturias”)

Steps 3 and 4 has been regarded as being evolutionary (e.g., [Vanderwende 95], [Arranz et al. 95]). That is, in contrast with single shot techniques, our methodology acquires knowledge

as a result of stepwise refinement allowing the lexicographer to inspect each step cycle of new knowledge being acquired.

5) Once a semantic class of lexical entries have been generated and placed in the LKB for the different languages, the acquisition of multilingual lexical information by means of the multilingual connection between lexical entries can be performed. This process is carried out by the **TGE** (Tlinks Generation Environment) module of SEISD (see Sections 3.6.3 and 4.5 and Chapter 6). For instance, using the knowledge placed in the bilingual dictionaries, the following links can be generated for *fabada_x_1_1* and lexical entries of LDOCE (a similar mechanism has also been used, see 6.3.3, for linking lexical entries to WordNet synsets):

fabada_x_1_1	linked to	stew_l_1_1	(by means of parent tlink).
fabada_x_1_1	linked to	broth_l_0_1	(by means of grandparent tlink).
fabada_x_1_1	linked to	stock_l_1_12	(by means of grandparent tlink).

6) Finally, when the extraction process ends, the lexical knowledge acquired must be validated and tested in order to look for incompleteness (for instance, daughter lexical entries with no differences between them), in order to perform further acquisition cycles or add new information manually. This process is aided by the **LDB/LKB system** enhancement (see Sections 3.6.4 and 4.6).

1.5 Outline of the thesis

This thesis focuses on the massive acquisition of lexical knowledge from monolingual and bilingual conventional dictionaries (on-line dictionaries or Machine-Readable Dictionaries, MRDs). A complete productive methodology for acquiring useful lexical knowledge from MRDs has been designed. SEISD, a powerful, complete and flexible software system allowing us to acquire massive lexical knowledge from on-line monolingual and bilingual dictionaries and to represent and validate the lexical knowledge acquired in a Multilingual Lexical Knowledge Base, has been implemented. Finally, we propose, implement and experimentally test various techniques in different methodological steps, obtaining improvements for several of them.

Thus, in this thesis we set out to achieve the massive automatic acquisition of lexical knowledge from conventional dictionaries allowing an easy construction of a large set of rich lexicons (from MTDs to MLKBs) suitable for use in a wide range of NLP systems (morphological analysers, syntactic analysers, Information Retrieval systems, Machine Translation applications, etc.). While for English a huge set of rich lexical resources are available (highly coded MRDs such as LDOCE, Lexical Data Bases such as Comlex, Lexical Knowledge Bases such as WordNet, etc.) this is not the case for the majority of languages (even for an widely spoken language such as Spanish). However, a great deal of monolingual and bilingual dictionaries are available for many languages. The possibility of obtaining large computational lexicons for NLP tasks from them using automatic techniques (even for less coded and structured dictionaries than LDOCE) is explored in this thesis.

In particular, we designed a complete methodology to build and validate an MLKB from a set of monolingual MRDs using bilingual MRDs to aid the linking process between languages. We applied this methodology to a concrete set of monolingual and bilingual MRDs (with their own particular characteristics: size, encoding, information content, etc.) without loosing generality. However, our methodology can be applied to any monolingual descriptive dictionary in any language.

As the majority of MRDs are not built for computational purposes, we designed a mixed methodology. We prescribed a set of semantic primitives using the LKB and we derived a natural classification of the concepts represented implicitly in the MRD definitions.

We covered the whole methodology designning and implementing a complete modular computer system named **SEISD** (*Sistema d'Extracció d'Informació Semàntica de Diccionaris*) which provides a user-friendly interface with several subsystems and also a way of

integrating these subsystems with the management of the multiple sources of heterogeneous data used by the system. SEISD was designed as a medium for the extraction methodology and integrates the Acquilex representational formalisms and their supporting software tools. SEISD covers the main functions of the proposed methodology, that is, the extraction of semantic information implicitly located in DGILE (performed by the **TaxBuild** and **SemBuild** modules), the mapping process of the information extracted to the LKB (covered by the **CRS**), the multilingual acquisition process (performed by the **TGE**) and the validation and exploitation of the lexical knowledge acquired (carried out by the **LDB/LKB** System).

A central guideline was to build the whole system so as to perform each process semi-automatically. Once the whole system was finished, each module was tested in order to analyse its performance (the results are reported in [Castellón 93] and [Taulé 95]). After this first test, some improvements in both methodology and techniques applied were introduced into some modules for efficacy (to obtain more information) and efficiency (to obtain this information more easily, that is, applying fully automatic techniques). A second test was performed to compare the results with the previous ones, obtaining improvements in both aspects (efficacy and efficiency). Thus, following the methodology presented in this thesis and using SEISD we are able to acquire more knowledge with less effort from conventional dictionaries.

The rest of the thesis is structured as follows. After this introductory chapter, Chapter 2 presents a general overview and discussion of the main alternatives, problems and results regarding lexical acquisition. That is, an in-depth study is provided of the different lexical acquisition approaches, methodologies and experiments appearing in the literature.

Chapter 3 is intended as a general vision of the global methodology and the SEISD environment. Thus, firstly, the main methodological aspects of the lexical acquisition process are explained. This includes the characteristics of the lexical resources used, the information to be extracted from them, how to carry out the process and how to represent and exploit the information extracted. Secondly, the main objectives, architecture and subsystems of SEISD are described. This second part makes a description of the different functions covered by SEISD as well as the different systems used to represent and exploit the lexical knowledge, the different analysers used to perform the acquisition and the different modules of SEISD architecture.

Although Chapter 2 covers the current state of the art on lexical acquisition, Chapter 4 explores in depth the main problems that each of the SEISD modules is required to deal with. That is, Chapter 4 performs a deep study of the different approaches appearing in the literature to carry out the different functions we designed for SEISD modules. Then, this Chapter focuses on the definition of the main semantic subsets, the construction of taxonomies having no explicit semantic information, the deeper semantic acquisition process, the multilingual lexical acquisition and the validation of the lexical knowledge extracted (this last task, carried out by the **LDB/LKB** system of SEISD).

In Chapter 5 we explains the main results achieved applying the main methodology using the SEISD environment. A detailed study and evaluation of a concrete subset is carried out. The methodology and system have been tested on the task of acquiring as much lexical semantic knowledge as possible from a monolingual dictionary for a significant semantic domain. Section 5.2 describes two new automatic techniques: The first for detecting (and/or selecting) the main semantic subsets underlying MRD definitions and the second for discovering the main top dictionary senses representative of a given semantic subset. Section 5.3 reports the (semi)automatic and fully automatic taxonomies we obtained using **TaxBuild**. Section 5.4 reports the results obtained first by the **SemBuild** module of SEISD analysing in depth the definitions of DGILE and second by placing this knowledge in the LKB by means of the **CRS**.

Chapter 6 deals with the multilingual lexical knowledge acquisition, applying different strategies with preexisting LKBs and bilingual dictionaries using the **TGE** module of SEISD. While Section 6.2 deals with the general multilingual framework of SEISD, Section 6.3 is devoted to illustrating the main results: a) linking bilingual dictionaries to monolingual English dictionaries and b) linking monolingual Spanish dictionaries to monolingual English dictionaries using bilingual dictionaries.

Finally, Chapter 7 addresses the main contributions, results and conclusions of this thesis and possible directions of further work.

Chapter 2

Words and Works

2.1 Introduction

This chapter overviews and discusses the main problems we have to face in lexical acquisition tasks. Different approaches to deal with such problems are described and main results are presented here. That is, what follows tries to be a general study of the different lexical acquisition approaches, methodologies and experiments appearing in the literature. That is, the current state of the art on Lexical Acquisition. As we have shown in section 1.2, in order to face the lexical acquisition problem three central questions must be answered, a) what information/knowledge is needed? b) where this information/knowledge is located? and c) which procedures can be applied to extract this information/knowledge from the sources?

Then, after this introduction, the first sections of this chapter try to answer these fundamental questions. As this thesis mainly focuses on the descriptive approach (see section 1.2), we perform an in depth study of lexical knowledge acquisition from on-line resources. Thus, section 5 summarises the lexical knowledge acquisition from MRDs, section 6 the work on lexical knowledge acquisition from Corpora and section 7 combining structured and unstructured lexical resources. Section 8, finally, accounts for the main international projects in the field of lexical acquisition.

2.2 What information is necessary in the Lexicon?

Of course, the first issue to be addressed when dealing with lexical acquisition refers to the type of information we need to include in the lexicon and must be acquired from the available knowledge sources.

We must consider both the domain (how many lexical entries would be present in the lexicon) and the range (the amount of information that should be attached to each entry). Both aspects are strongly dependent on the specific application for which the lexicon is building (which application domain, which tasks, etc). Roughly speaking, an analysis of the tasks to be carried out will determine the range -the content information- of each entry while the application domain will determine the domain of the lexicon (when the lexical entries could be considered word forms or stems or lexemes or when the information could be factorised through hierarchies or other lexical organisation will be addressed later).

We discussed, in section 1.2.1, the kind of NLP tasks where lexicons generally take place and we derived from it six types of information potentially relevant to NLP systems that would be present in the lexicon (phonology, morphology, syntax, semantics, pragmatics, translation links). We must point out that this information covers all the levels of description (from phonology to pragmatics) usually taken into account in NLP tasks. Several topics must be considered as regards the kind of information to be placed in the lexicon.

2.2.1 Granularity of the information.

Both the range and the domain of each piece of information present in the lexicon can be addressed at different level of granularity. Some information, for instance, is attached to word forms, other to lemmas, other to senses.

The distinction between word-entry and sense-entry is specially important dealing with the semantic content of the lexicon (i.e. word taxonomies [Nakamura & Nagao 88] vs. word sense taxonomies [Bruce et. al. 92] or between both the coarse grained sense distinctions made in [Gale et al. 93] and [Yarowsky 92], also called homographs in [Guthrie et al. 93], that can be compared to that of the file level in WordNet [Agirre & Rigau 96a]). Determining the number of senses for a given word at a given level of granularity and attaching to each sense its specific information detecting at a time commonalties that can be factored at a entry-level or attached to a collection of senses, are different tasks that must be carried-out during acquisition. Obviously, this issue is closely related to the WSD (Word Sense Disambiguation) problem that will be addressed later in section 4.3.

The kind of allowable values to assign to each feature presents, too, different levels of granularity. In the case of information expressed as symbolic labels (e.g., part-of-speech POS) the number of allowable labels differs from one approach to another (e.g., in POS tagging the cardinality of the tagset can vary from fine-grained ones -more than 200 tags- to rather coarse-grained ones -less than 40 tags-. Forms of subcategorizations can be used as well.

2.2.2 Representation of the information.

An important issue to be addressed is the way to represent the lexical information. It is difficult to select a unique representation formalism for lexical information due to the great number of features to be represented and its heterogeneity. Most of the features we have considered differ on their form: attribute-value pairs, binary or n-ary relations, on their realisation: optional vs. obligatory, cardinality, default values, on the degree of the imposed constraints: exact values, preferences, stochastic assignment, etc.

For most attribute-value features (e.g. POS) a database-like organisation seems to be appropriate. The relational model has been used widely. In this model, the lexical entries are represented as tuples in one or more relations, each of which includes several attributes ranging over the appropriate domains [Ide at al. 91].

When dealing with text-based features the previous model is clearly inadequate and we can move to the so-called text models. In these models, adequate for representing, for instance, the definition of an entry or examples of use, the information can be seen as a, possibly marked or tagged, stream of characters [Ide and Véronis 95].

Both previous approaches lack expressivity for dealing with common deductive requirements, for instance property inheritance, that appear frequently in lexicons. Other issues not covered by these conventional approaches are the need of procedural (or assertional) capabilities and the treatment of exceptions. Object Orientation is, of course, the obvious answer to these objections. OO systems cover largely both the terminological part (that describes the data) and the assertional part (that describes the procedures and functions associated with the data).

For limited deductive capabilities OO database systems or deductive database systems can be used. If the expressive capacities we need are over the scope of these general-purpose systems a possible solution is to turn over frame-based representation formalisms. These systems provide extended and more powerful capacities than the OO systems at a higher cost. Frequently, these extra capacities imply severe operational limitations as the need of residing in primary memory instead of a secondary storage.

If deeper inference capacities are requested, like more sophisticated inheritance mechanisms, lexical or morphological rules, reentrancy, disjunctive values, constraints, etc. we must move to non-standard representation formalisms. In this area, most systems can be labelled as “ad-hoc systems”. Anyway, we can see that a great number of such approaches fall into the “unification paradigm”.

Lexicons can be represented as MRDs (Machine Readable Dictionaries, i.e. dictionaries for human use in electronic support), MTDs (Machine Tractable Dictionaries, i.e. the same after some limited processing for allowing easier access to NLP programs), LDBs (lexical Databases) owning the same data but organised in a database-like fashion for allowing more flexible querying and LKBs (Lexical Knowledge Bases) where the content, possibly derived from previous LDBs is organised from a semantic, rather than lexical, point of view.

2.2.3 Scope of the information.

In some cases lexicons owns general knowledge about words and in other cases more specific domain-dependent information (e.g., terminological information). The way of acquiring the information and the way of accessing it can depend on these differences.

An important point to consider is whether the lexicon will be word-based or concept-based, assuming, of course, that both linguistic and concept information must be included [Cavazza & Zweigenbaum 95]. This issue is closely related with the descriptive or prescriptive approaches for lexical acquisition as discussed in section 1.2.3 that will be addressed in detail later.

2.2.4 Way of accessing the information.

The way of accessing the information contained in a lexicon depends, of course, on the form this lexical information has been represented: structured (lexicons derived from corpora, Thesaurus, MRDs, MTDs, LDBs, LKBs, semantic nets, etc.), or unstructured (raw corpora, processed corpora).

There are, usually, three forms of accessing: a) directly by headword, word, word form, sense, e.g. getting all the available information for a given word, b) by content, e.g. getting all the entries satisfying a given constraint and c) by relation, e.g. getting all the entries related with a given one to a particular relation. Frequently, lexicon access mechanisms lacks for allowing this last form for accessing in an efficient way.

For instance, SemCor [Miller et al. 93], a part of the Brown Corpus semantically tagged using WordNet synsets, provides an interface allowing the simultaneous access to the corpus and WordNet at a word form, word or sense level. Nevertheless, complex queries by content or by relation are not provided by the interface. Simultaneous access to both resources by content or relation only can be done using ad-hoc programs on the source data.

2.3 Where is the information needed for the Lexicon?

Three main sources of information to build wide-coverage lexicons for NLP systems have been considered in section 1.2: introspection, structured lexical resources and corpora.

Introspection. Obviously, the construction of the lexicon using the knowledge about the language and the world the NLP system human builder owns, should guarantee the quality of the resulting data[1]. However, large-scale lexicons constructed in this way needs a huge human labour during a large period of time. Many lexicons for NLP have been developed by introspection. Among others, Word Dictionary, a lexicon constructed for the Linguistic String Project (LSP) contains over 10,000 entries [Fox et al. 88], WordNet [Miller 90] currently (version 1.6) represents 123,497 different content words and 99,642 senses grouped into synsets and related by a set of semantic relationships, Comlex [Grishman et al. 94], a computational lexicon providing detailed syntactic information for approximately 38,000 English words, Cyc Ontology [Lenat 95] contains over 100,000 terms and has consumed a person-century of effort.

---

[1] *If a strong control on the produced data is performed in orden to avoid inconsistencies and other errors. For instance, WordNet1.5 still contains a large number of them [Fischer 97].*

The latest version of LDOCE, LDOCE3-NLP, with 80,000 senses has been specially created by Longman lexicographers for computational linguistic research.

MRDs. As dictionaries are special texts whose subject matter is a language (or a pair of languages in the case of bilingual dictionaries) they provide a wide range of information about words (see section 1.3) by giving definitions of senses of words, and, doing that, supplying knowledge not just about language, but about the world itself. Thus, conventional monolingual and bilingual dictionaries form an excellent starting point to construct substantial lexicons because they constitute a highly structured and relevant source of information about words and meanings.

From the earliest attempt to convert a paper-printed dictionary into a MRD performed by hand in the late 1960s with the W7N until now, a large set of dictionaries have been exploited as lexical resources (see [Wilks et al. 96] chapter 6, for an account of such early approaches). Although the most widely used monolingual MRD for NLP is LDOCE (for details see [Boguraev & Briscoe 89a]) which was designed for learners of English with only limited facility in the language, COBUILD and OALD has been widely used too. MRDs, usually ranging between 30,000 and 50,000 entries, contain structured information on spelling, stress, pronunciation, hyphenation, capitalisation, usage notes for semantic domains, geographic regions; etymological, syntactic and semantic information about the most basic units of the language (and translation correspondences to other languages in the case of the bilingual ones). Currently, many dictionaries in machine-readable form are becoming available from publishers, thanks to the initiatives such as the Consortium for Lexical Research (CRL), the Linguistic Data Consortium (LDC), the Oxford Text Archive (OTA), European Linguistic Resources Association (ELRA), etc.

Other structured lexical knowledge resources (in machine-readable format) for human use like thesauri or encyclopaedia may be also considered for lexical knowledge acquisition. Thesauri like Roget's International Thesaurus which separates 60,071 words into 1,000 semantic categories (used by [Yarowsky 92], [Grefestette 93] or [Resnik 95]), Roget's II: The New Thesaurus and "The New Collins thesaurus" (used both by [Byrd 89]), Macquarie's thesaurus (used by [Grefenstette 93]), or the Spanish thesaurus contained into *Diccionario Ideologico de la Lengua Espa–ola* J. Casares (converted to machine-readable format by [Sanchez 91]), or the *Bunrui Goi Hyou* Japanese thesaurus (used by [Utsuro et al. 93]). Encyclopedia like Grolier's Encyclopaedia (used by [Yarowsky 92]) or The World Book Encyclopedia (used by [Gomez et al. 94]).

Other more specifics sources has been used as Onomasticon Telephonic Guides, collections of proper names, terminological data banks, etc.

Corpora provide an additional, though less structured source, relating to issues of usage, such as the relative frequency of words or the range and frequency of different patterns of linguistic realisation. A variety of corpora with different levels of annotation has been collected in recent years. Annotated corpora range from pos-tagged corpora, lemmatised corpora, syntactically analysed corpora, bracketed corpora, semantically marked corpora (with very different granularities: senses, WordNet synsets, WordNet lexicographic files, Roget's categories, etc.). Perhaps, the most used English text collections in the research community are the Brown Corpus (around 1 million words [Francis & Kucera 82]) and the Wall Street Journal materials (different releases of these materials range from 1 to 3 million words). Both collections have been annotated by the Penn Treebank Project for part of speech (POS) tags [Marcus et al. 93] and for skeletal syntactic structure [Marcus et al. 94]. Brown Corpus have also been annotated partially (over 250,000 words [Miller et al. 93]) by Princeton automatically for POS tags [Brill 92] and manually for WordNet sense tags. [Ng & Lee 96] tagged manually 192,800 word occurrences of 191 nouns and verbs, which appears in both the Brown corpus and the Wall Street Journal. The most used bilingual text collection in the research community is the Hansard corpus (around 85 million English words corresponding to 3.5 million sentences and 97 million French words corresponding to 3.7 million sentences from the Canadian Parliamentary Proceedings). This corpus has been aligned by sentence [Brown et al. 91b]. Today, many locations have samples of text running into the order of millions, or even

tens or hundreds of million words. Collections of this magnitude are becoming available, thanks to data collection efforts such as the ACL's Data Collection Iniciative (ACL/DCI), the European Corpus Iniciative (ECI), the British National Corpus (BNC), the Linguistic Data Consortium (LDC), *El Instituto Cervantes* (IC, for Spanish), *L'Institut d'Estudis Catalans* (IEC, for Catalan), etc. Readers interested in issues related to the collection of such corpora may refer to [Atkins et al. 92] or [Alvar & Villena 94]. [Souter & Atwell 94] present a survey on currently available syntactic analysed corpora.

Not only balanced general-purpose corpora can be used as source of lexical information. Narrow domain specific corpora has been used as well for extracting terminological information.

Mixing resources. As could be expected, it is not realistic to obtain all the information needed for a lexicon from only one source. Existing wide-coverage computational lexicons built by hand or extracted from MRDs may suffer from incompleteness problems. For instance, [Walker & Amsler 86] compared entries in W7N to a 3-month sample of stories from the New York Times newswire. They found only 64% of the news wire words were not in the dictionary. Their breakdown of these results revealed one fourth to be inflected forms, one fourth were proper nouns, one sixth were hyphenated forms, one twelfth were miss-spellings and one fourth were unresolved (due to were new words since the dictionary was published).

In addition, [Briscoe & Carroll 93] report that half the failures of the wide-coverage parsing system utilising a lexicalist framework was due to incorrect subcategorization for predicate valency. Furthermore, lexical information is often tied to particular domains not reflected into general dictionaries. A final problem is that MRDs rarely record the relative frequency of lexical information in language usage as corpora do. So, often these sources are used in a combined way for acquiring lexical knowledge not present in only one source. Works combining resources may be classified in a) those which combine structured lexical knowledge sources (MRDs, ontologies, thesaurus, etc.) among them (i.e. [Knigth & Luk 94]), and b) those which combine structured and non-structured (corpora) on-line lexical resources[1] (i.e. [Klavans & Zoukermann 96]).

An important point as regards the use of lexical resources is availability. How the information is represented in the original source and, so, should be it extracted will be discussed in next section. Another problem derives from copyright. Most of the data present in lexical resources are protected by copyrights and researchers have limited right of access. Frequently the choice of using a resource is constrained for this reason.

2.4 How to extract that information?

We stated in section 1.2.3 that two main alternative approaches could be used in the lexical acquisition process: the prescriptive and descriptive approaches. That is, in the prescriptive approach, a set of primitives is defined, or prescribed, prior to or in the course of designing and developing the whole system. The descriptive approach on the other hand, allows a natural set of primitives derived from a natural source of data without any pre-existing frame.

The information attached to each lexical entry can be obtained by manual, automatic or (semi)automatic approaches depending on the methods applied, sources used and the information needed for a particular application.

From the point of view of the human intervention, three major approaches to lexical acquisition have been developed: machine-aided manual construction, (semi)automatic extraction from pre-existing lexical resources and the combination of the two previous ones. Of course, each of these techniques has advantages and disadvantages.

---

[1] *Calzolari (see [Wilks et al. 96] page 97) distinguished the different relations that can be stablished between corpora and lexicons examining which informations present in corpora could be extracted for enriching lexicons and which one, present in lexicons, could be used for enriching corpora.*

An important issue is the relation between conceptual acquisition and lexical acquisition. Frequently, the acquisition task involves both kind of knowledge as well as the relations between them.

Most of the systems built following the prescriptive approach start building a conceptual framework for attaching later the corresponding lexicalizations to each concept.

Following the prescriptive approach several tools have been developed making quicker and easier the manual construction and maintenance of the lexicons. Manual construction is the most reliable technique but suffers of a very time-consuming problem. The system presented in [Nirenburg & Raskin 87] allows in the first step the creation of an Ontology of concepts (the system controls the process in order to create the concepts and maintain its consistency) and in the second one to connect the words with the ontological concepts. One sentence express perfectly this methodology: "The world first, the words later". This first approach has been taken in large-scale MT systems, see for example, the description of METAL in [Hutchins & Somers 92]. Other system that follows this approach are, for instance, the projects Cyc [Lenat & Guha 89], Upper Model [Bateman 90], ONTOS [Carlson & Niremburg 90], WordNet [Miller 1990] and EDR [Yokoi 95].

The descriptive approach is intended to obtain lexical knowledge in an automatic or semiautomatic way from pre-existing texts. Such lexical resources include a wide range of data available in computer access format as MRDs, lexicographic databases, terminological data banks, (morphological, syntactically and semantically tagged) monolingual and bilingual corpora and already existing lexicons. This approach might be seen as the contrary of the previous one: "The words first, the concepts later". For instance, using structured lexical resources, [Bruce et al. 92] built the complete taxonomy of the nominal part of LDOCE linking top dictionary senses to subject codes (no external primitives were added). Using corpora, [Pereira et al. 93] create new semantic hierarchies using distributional clustering.

Other works follow a combination of the two previous approaches. They propose first, to prescribe a minimal part of primitives in order to provide coherence and structure to the lexicon and, second, following descriptive approaches extract lexical data to be attached to the previously prescribed part (e.g. [Ageno et al. 92a], [Hovy & Knight 93]). Thus, [Resnik 93], [Ribas 95] or [Li & Abe 95] combine the use of monolingual corpora and WordNet with distributional statistics to obtain surface semantic restrictions for predefined syntactic positions.

On the other hand, under the generic label of Information Extraction (IE) a huge number of systems and techniques has been developed recently (i.e. Fastus [Appelt et al. 93] or [Hobbs et al. 93], AutoSlog [Riloff & Shoen 95]). See [Grishman & Sundheim 96] for an overview of the current Message Understanding Conference MUC-6.

2.5 Lexical Knowledge Acquisition from MRDs

As we stated in the previous section, manual construction of lexicons (by expert people) is the must reliable technique for obtaining structured lexicons but is costly and highly time-consuming. This is way many researchers have focussed on trying to extract lexical knowledge and semantic information from pre-existing structured lexical resources in an as automatic as possible way. This section deals with previous work on the acquisition of lexical information from structured sources (i.e. MRDs, Thesauri, etc.) while section 2.6 is devoted to the work using non-structured sources (i.e. Corpora) as source of lexical information and section 2.7 to computational methods for obtaining this information from a combination of lexical resources (i.e. structured and non-structured).

Of course, MRDs -the machine tractable versions of conventional dictionaries- have been the primary source of lexical knowledge and semantic information that can be used for automating the task of lexicon construction.

As MRDs usually appear in a special format for printing purposes, much of the information they contain consists of special codes to produce a readable document for humans. Thus, parsing the dictionary entries, a fundamental preprocessing step for producing a MTD from

the MRD, uses to be necessary. For exploiting the data contained into the dictionary, the loading of the MTD into a LDB uses to be performed.

Much of the knowledge needed for NLP lexicons can be found explicitly or implicitly in a conventional dictionary (see section 1.2.1, 1.3). Explicit information (i.e. part-of-speech categories for a given lemma, etc.) can be extracted straigthfully. Problems arise dealing with implicit information usually contained in definitions and examples (or in translations in bilingual dictionaries). Most of the relevant research on MRDs is devoted to extract knowledge appearing implicitly in dictionary definitions.

Although early attempts were made in the later sixties and seventies (see [Wilks et al. 96]), the seminal work on acquiring implicit lexical knowledge from dictionaries was done by [Amsler 81] building "tangled hierarchies" of lexical units.

Following the descriptive approach several works have been made using different methodologies that differ in lexical knowledge used, coverage, tools, purposes and results. [Boguraev & Briscoe 87] collect the grammar codes from LDOCE for building a large lexicon to perform syntactic analysis. [Boguraev & Briscoe 89a] describe different techniques and experiments exploiting MRDs (mainly with LDOCE) in order to construct lexical components for NLP. [Veronis & Ide 91] provide a quantitative evaluation of the information extracted from several MRDs. [Artola 93] shows a dictionary help system that take advantage of the automatically extracted information of the Le plus petit Larousse (LPPL) dictionary. [Castellón 93] describes the linguistic research made on nouns and [Taulé 95] on verbs using SEISD environment with DGILE dictionary. [Wilks et al. 93] describe three methodologies to construct in a combined way a robust linguistic database. [Dolan et al. 93] describe an automatic methodology that exploits LDOCE to construct a highly structured lexical knowledge base. [Knight & Luk 94] describe a system for building in a semiautomatic way a large-scale ontology merging several and heterogeneous lexical knowledge sources to support semantic processing in the Pangloss knowledge-based machine translation system. [Vossen 95] studies the ways nouns "name" the things, and how this information should be stored in a lexicon. [Wilks et al. 96] integrate and synthesise different methods in the SPIRAL and ARC procedures with LDOCE for building new lexicons.

As we have seen, the lexical knowledge contained into MRDs have been widely used and exploited. In order to provide the range of the knowledge contained into MRDs we provide a review of the work using MRDs as a lexical knowledge source for several NLP tasks:

Syntactic disambiguation. [Jensen & Binot 87], [Ravin 90] (using respectively noun and verb Webster's dictionary definitions) and [Dolan et al. 93] (using LDOCE) propose to resolve prepositional phrase attachments by using preferences obtained by applying a set of heuristic rules to dictionary definitions. The rules match against lexico-semantic patterns in the definitions in order to evaluate separately each possibility. Furthermore, [Jensen & Binot 88] propose the use of the information contained into MRDs to help determine the proper attachment of prepositional phrases and relative clauses, pronoun reference and the interpretation of dangling references. Rather than a simple stepwise approach, [Vanderwende 95] proposes an incremental approach were the semantic relations extracted from LDOCE added to the LKB in initial steps allow to disambiguate ambiguous patterns, enabling the identification of more semantic relations during subsequent steps.

Semantic Processing [Byrd 89] defends the increase of the semantic capability of NLP systems creating a LKB derived from several MRDs. In that sense, the LKB derived from LDOCE by [Dolan et al. 93] could be used for resolving semantic ambiguities in text, such as the correct attachment of prepositional phrases, anaphora, etc. [Ravin 90] reports the use of the knowledge contained in Webster's dictionary definitions to disambiguate the multiple semantic relations holding between the head and a prepositional phrase in a subset of verbal definitions in the same dictionary. [Vanderwende 95] uses the semantic information present in LDOCE captured in previous steps to extracts more accurate semantic information in current steps.

Word Sense Disambiguation. Although work on WSD will be studied in depth later[1] (see section 4.3), we can quote here two different methodological approaches. The first, statistically-based (i.e. [Yarowsky 92]) and the second one knowledge-based (i.e. [Agirre & Rigau 96a]).

Since [Lesk 86] many researchers have used MRDs as a structured source of lexical knowledge for the WSD problem. That is, attaching a set of prescribed dictionary senses to words in context. He proposed a method for guessing the correct word sense in context by counting word overlaps between each dictionary definition and the context. [Veronis & Ide 90] propose a similar method but uses a spreading activation network (see [Hirst 88] and [Hayes 77]) constructed from Collins Dictionary of English Language. [Sutcliffe & Slater 94] compare the both previous methods using the Merrian-Webster dictionary. [Slator 91] propose a system for preferring word senses according to context using a restructured version of the LDOCE subject codes. [Cowie et al. 92] and [Wilks & Stevenson 97] use the simulated annealing technique for overcoming the combinatorial explosion of Lesk method using LDOCE.

More sophisticated techniques exploiting dictionary definitions have been also carried out. Thus, [Wilks et al. 93] use coocurrence data extracted from LDOCE for constructing word-context vectors and thus, word sense-vectors. The similarity between those vectors can then be compared by means of several formulas.

[Yarowsky 95] proposes the use of MRDs to collect seed words in the first step of his cycling procedure, which collect local features. [Karov & Edelman 96] describe another cycling procedure for learning from a corpus a set of typical usage for each of the senses of the polysemous word listed in a MRD.

[Guthrie et al. 91] propose the use the information located in the subject semantic codes of LDOCE for partitioning the dictionary and collect neighbours (or salient words [Yarowsky 92]) for WSD in a Lesk style.

[Harley & Glennon 97] perform an ad-hoc weighting mechanism using the different sources of lexical knowledge present into the completely coded Cambridge International Dictionary of English (CIDE).

[Rigau et al. 97] also use implicit information contained into MRDs for constructing content vector representations and testing different techniques and similarity measures for assigning the correct hypernym genus sense. In this approach, we also use a bilingual MRD to assign semantic categories from WordNet to word senses and performing, in a similar way than [Yarowsky 92], an unsupervised training process for collect salient words for each semantic category (see [Rigau et al. 98]).

Information Retrieval. MRDs have been used also as structured lexical knowledge resources to support several tasks in Information Retrieval systems. [Fox et al. 88] describe the construction of a large LKB from several MRDs to support interactive query expansion and search for information. [Krovetz & Croft 92] propose to index documents by word senses taken from an MRD and [Voorhees 93] from WordNet.

Bilingual Lexicons for Machine Translation. Although most of the effort for the extraction of implicit knowledge has been carried out on monolingual MRDs, valuable contributions have been also performed using bilingual MRDs. [Rizk 89] discuses the problem of ambiguous sense references in the Collins Robert French/English dictionary. [Knight & Luk 94] use the Collins Spanish/English bilingual dictionary for linking Spanish words to the Sensus Ontology (where concepts are represented by means of English words). [Tanaka & Umemura 94] use two intermediate Japanese/English and French/English bilingual dictionaries to construct automatically a Japanese/French bilingual dictionary. [Ageno et al. 94] use a Spanish/English bilingual dictionary for linking in a (semi)automatic way Spanish and English taxonomies extracted from DGILE and LDOCE. In a similar approach, [Rigau et al. 95] propose an automatic approach for linking Spanish taxonomies extracted from DGILE to

---

[1] *In order to build sense disambiguated taxonomies from MRDs (a central issue in the work presented here, see Section 3.6.3 and 5.3.2) a Genus Sense Disambiguation task (a closely related problem) must be performed.*

WordNet synsets. [Okumura & Hovy 94] describe (semi)automatic methods for associating a Japanese lexicon to an English ontology using a bilingual dictionary. In a similar approach, [Rigau & Agirre 95] propose several complementary techniques for attaching directly Spanish and French words extracted from the bilingual dictionaries to WordNet synsets. [Atserias et al. 97] combine several lexical resources and techniques to map Spanish words from a bilingual dictionary to WordNet in order to build a parallel in structure semantic net. [Farreres et al. 98] propose also the use of the taxonomic structure derived form a monolingual MRD to aid this mapping process.

Enriching semantically MRDs. MRD information could be also used to enrich semantically others or the same MRD. [Yarowsky 92] propose to use the salient words collected from a encyclopaedia and thesaurus for assigning semantic categories to dictionary definitions. Using this coocurrence-based technique and the notion of conceptual distance, [Rigau et al. 98 perform a three step semantic tagging of DGILE with semantic labels collected from WordNet. [Rigau 94] perform a knowledge-based technique for assigning semantic tags to the Spanish monolingual dictionary DGILE using a bilingual dictionary and WordNet. Using also knowledge-based methods, [Knight 93] propose several algorithms to attach LDOCE and WordNet and then, transfer the lexical knowledge from one to the other. [Chen and Chang 98] propose LinkSense, a simple matching algorithm to label LDOCE to respect the semantic codes collected from LLOCE and Roget's thesaurus.

Building LKBs. Following the descriptive approach, several works have been made for the construction of LKB from MRDs useful for NLP. These works varies on the different degree of human intervention during the construction process, the lexical components represented in the LKB, the degree of syntactic or semantic analysis of the dictionary definitions. Thus, while [Fox et al. 88], [Dolan et al. 93] or [Barrière & Popowich 96] describe the construction of a large semantic network of words from several English dictionaries, [Byrd 89] proposes the creation of an LKB from MRDs in which word senses were clearly identified, endowed with appropriate lexical information, and correctly related to one another for increasing the semantic capabilities of NLP systems.

A central problem that will be treated in depth in following chapters consists of extracting taxonomies from the implicit knowledge, which appear in dictionary definitions. This main problem can be divided in two different subproblems. First, the location of the genus terms in the definitions. As the genus term appears not as a sense but simply as a word, the next subproblem consist on the selection of the correct sense (which usually appears in the same dictionary) for that genus term. The Genus Sense Disambiguation (GSD) problem can be considered as special case of the most general Word Sense Disambiguation (WSD) problem.

The most in serious attempt dealing with the automatic correct genus sense selection has been performed by the New Mexico State University NLP group at the Computing Research Laboratory with LDOCE (see [Bruce & Guthrie 92] or [Bruce et al. 92]).

Thus, some researchers have focussed on the automatic construction of taxonomies as a backbone of the LKB (i.e. [Copestake 90], [Bruce et al. 92], [Rigau et al. 97], [Rigau et al. 98]). Rather than acquiring taxonomies only from monolingual and bilingual MRDs, [Knight & Luk 94] describe several techniques for building a large scale LKB attaching monolingual and bilingual MRDs to several and heterogeneous ontologies. In the same way, [Rigau et al. 97] and [Rigau et al. 98] use an English ontology and a bilingual dictionary as a lexical knowledge source for several heuristics and mapping processes (see Sections 5.2 and 5.3 for further details).

Instead of limiting itsefs to taxonomies, some approaches perform an in depth analysis of the dictionary definitions taking profit of the defining formulae which are "significant recurring phrases" [Markowitz et al. 86]. Some early works perform a string pattern matching approach (i.e. [Chodorow et al. 85], [Markowitz et al. 86]) while others prefer structural patterns that match the syntactic analysis (i.e. [Jensen & Binot 87], [Alshawi 89], [Ravin 90], [Klavans et al. 90], [Artola 93], [Castellón 93], [Dolan et al. 93]). While some approaches only consider a one-to-one relation between the defining formulae and the type of lexical information it identifies (i.e. [Jensen & Binot 87]), later studies (i.e. [Ravin 90], [Klavans et

al. 90], [Vanderwende 95]) have shown that some defining formulae can convey several types of semantic information. While some researchers prefer general purpose parsing tools (i.e. [Jensen & Binot 88], [Dolan et al. 93], [Vanderwende 95]) some others prefer partial or adapted parsing tools designed for dictionaries (i.e. [Alshawi 89], [Artola 93], [Castellón 93]). Some attempts have been performed processing completely controlled or small dictionaries (i.e. [Artola 93], [Dolan et al. 93], [Barrière & Popowich 96]). Others preferred to process subsets of closely related dictionary senses (i.e. [Castellón 93]). Rather than a single shot process, [Vanderwende 95] proposes a cycling methodology improving the analysis each cycle is performed.

Related works constructing the LKB from noun definitions includes [Vossen 95], for verb definitions [Klavans et al. 90], [Ravin 90] or [Taulé 95] and for adjective definitions [Soler 96].

2.6 Lexical Knowledge Acquisition from Corpora

Although MRDs seems to be the more adequate texts for extracting lexical knowledge because they offer a vast size and ready highly-structured source of lexical knowledge they do not contain all the data needed for constructing a generic lexicon useful for any NLP system.

The growing availability of large on-line resources encourages the study of word behaviour directly from accessible raw texts. However, the methods by which lexical knowledge should be extracted from plain texts are still matter of debate and experimentation.

Corpus-based lexical acquisition knowledge is based on Firth[1] distributional hypothesis. Thus, the acquisition process proceeds from the analysis and synthesis of the lexical properties through distributional contexts where the interesting lexical items occur in raw texts. Both processes may be done in a manual or (semi)automatic fashion depending on the existing resources/tools and the difficulties to make the necessary generalisations (see [Church & Hunks 90]). Recently, several automatic systems to collect context of use and to analyse them in order to obtain appropriate lexical, syntactic and semantic generalisation have been proposed (work progress overviews on corpus processing and lexical acquisition may be found in [Zernik 91], [Charniak 93], [Oostdijk and deHann 94] and [Boguraev & Pustejovsky 95]). In this section we review several of these approaches and techniques in order to grasp some ideas about the current tenets on automatic acquisition of lexical information from corpora. Here we classify them by the kind of information extracted:

Proper Nouns. Proper nouns cause problems due to their high frequency in many types of text, their poor coverage in conventional dictionaries (old fashioned may appear in encyclopedia) and their importance in text understanding process. [Coates-Stephens 92], [Nani & McMillan 95], [Hearst & Schütze 95] and [Paik et al 95] describe different systems for Proper Noun knowledge acquisition. While [Coates-Stephens 92] describe a complete module based for text understanding, [Nani & McMillan 95] and [Paik et al 95] describe two systems focusing on Information Retrieval. While [Coates-Stephens 92] and [Nani & McMillan 95] approaches rely on context information, [Paik et al. 95] approach rely more on built-in knowledge bases. On the other hand, [Hearst & Schütze 95] use lexical coocurrence statistics in combination with a set of flat categories derived from WordNet to classify proper nouns in text.

Idiosyncratic Collocations. That is, the extraction of word coocurrence restrictions (predicative relations, rigid noun compounds, phrasal templates, etc.) that are lexically driven. Idiosyncratic collocations make that although "powerful" and "strong" are semantically almost equivalent (3 senses of "powerful" appear as a direct hyponym of "strong" senses in WordNet 1.5), they restrict for different words (e.g. you may prefer "strong tea" rather than "powerful tea" [Church et al. 91]). Several methods have been applied to corpora for the extraction of recurrent collocations. [Smadja 91a], [Smadja91b] and [Smadja 93] extracts words collocations detecting word sequences whose relative frequency of occurrence is

[1] *"you shall know a word by the company it keeps" [Firth 57].*

significant while [Smadja 92] compile a bilingual lexicon of collocations from a bilingual corpora. Most approaches differ on 1) the amount of context used for searching the collocation, ranking from the very local context (five words length windows) to global ones; 2) the statistical measure used for word association (the most frequently used is Mutual Information, [Church & Hanks 90a], although other measures, like the association ratio or the relative entropy has been used too); 3) the use of a dispersion measure in parallel with he association one (as in [Calzolari & Bindi 90]) and 4) the measures for computing the statistical significance of the detected association (usually Chi-square).

Preposition preferences. [Calzolari & Bindi 90] also propose to detect the lexical preference of verbs/nouns for specific prepositions that introduce complements. [Hindle & Rooth 93] use coocurrence of verbs and nouns with prepositions in a large body of text as an indicator of lexical preference. They compared the list of lexical preferences detected with those provided by COBUILD founding that the coverage obtained by the automatic procedure over-passed that of the MRD. [Charniak 93] proposes a general framework for extracting this kind of information and discusses the viability of the different models taking into account the availability of the sources. [Resnik and Hearst 93] combine the lexical association strategy with the use of noun class information.

Subcategorization structures consisting of patterns of lexical preferences that predicates show for the syntactic realisation of their arguments, is far beyond of capturing individual lexical preferences. [Poznanski & Sanfilippo 93] present a method for individuating dependencies between the semantic class of predicates and their associated subcategorization frames. [Briscoe & Carroll 93] propose a system that uses global syntactic information and linguistically guided filters on the patterns accepted. [Briscoe & Carroll 97] describe a novel technique and implemented system for constructing a subcategorization lexicon from textual corpus improving the accuracy of a parser in a appreciable amount.

Selectional restrictions. Methods for acquiring selectional restrictions from on-line resources combine the use of syntactically analized corpora, pre-existing thesaurus (which provides conceptual generalisation to lexical occurrences) and different kinds of conceptual similarity estimation. Thus, [Resnik 93], [Ribas 95] or [Li & Abe 95] combine the use of monolingual corpora and WordNet with distributional statistics to obtain surface semantic restrictions for predefined syntactic positions. [Utsuro et al. 93] and [Wu & Palmer 94] combine the use of bilingual corpora and a Japanese and Chinese thesaurus for acquiring selectional restrictions and preferences for Machine Translation.

Thematic structure can be acquired using the functionality of the verb and the co-ocurrence of data. The co-ocurrences are matched against subcategorization patterns allowing the thematic structures be recovered. [Basili et al. 92a], [Basili et al. 92b] or [Pazienza 94] describe the ARIOSTO system, which acquire class-based selectional restrictions annotated with thematic labels.

Word semantic classes. Under this epigraph, two complementary acquisition tasks are considered: a) assignment of pre-existing semantic categories to unknown words (e.g. [Zernik 89], [Grefestette & Hearst 92]) or assignment of known words to new pre-existing semantic categories for specific domains (e.g. [Basili et al. 95]), and b) creation of new semantic hierarchies using distributional clustering (e.g. [Brown et al. 92], [Pereira et al. 93], [Dagan et al. 94]). The second task, includes the first one: new classes are somewhat assigned to the words that originated them. Instead of sophisticated statistical techniques, [Hearst 92] describe a simple method to capture hyponymy relations from corpora.

Bilingual lexicon acquisition. Parallel corpora are useful resources for acquiring a large variety of linguistic knowledge [Dagan et al. 91]. Bilingual corpora can be used for many purposes. Among others, for acquiring new lexical correspondences word to word [Smadja 92], or new lexical knowledge for disambiguating word senses across languages [Gale et al. 93].

Mainly, lexicon compilation methods attempt to extract pairs of words or compounds that are translations of each other from previously sentence-aligned parallel texts (e.g. [Eijk 93], [Kumano & Hirakawa 94] or [Utsuro et al. 94]). Bilingual corpora alignment can be performed at character, word or sentence level (e.g. [Brown et al. 91b], [Gale & Church 91], [Church 93] or [Kupiek 93]). Furthermore, [Fung 95] proposes an algorithm for bilingual lexicon acquisition that bootstraps off the corpus alignment process.

2.7 Lexical Knowledge Acquisition Combining Resources

Since [Byrd 89] proposed the integration of several structured lexical knowledge resources derived from monolingual and bilingual MRDs and Thesaurus, many researchers have proposed several techniques for taking advantage from more than one lexical resource.

Working with several structured lexical resources. Working on monolingual dictionaries, [Veronis & Ide 91] show that MRDs can be reliable sources of lexical knowledge if we are able to combine information from them. They provide a quantitative evaluation of the information extracted merging five monolingual MRDs showing that for any one dictionary, 55-70% of the extracted dictionary is garbled in some way. However, these results can be dramatically reduced to about 6% by combining the information extracted from all of them. [Grishman et al. 94] compared Comlex subcategorization information from those appearing in Word Dictionary (from the linguistic String Project [Sager 81]), OALD and LDOCE dictionary. Comlex is a good example of this approach. The lexicon was built using as sources LDOCE, the Acquilex type system for verbal structure, the Brandeis classification of verbs, etc. and manual tunning. [Knight 93] provides a definition match and hierarchical match algoritms for linking WordNet synsets and LDOCE definitions (placed in a taxonomy).

[Byrd 89] and [Risk 89] using similar techniques performed, respectively, a mapping process between two thesaurus and two sides of a bilingual dictionary. [Tanaka & Umemura 94] produce a new Japanese/French bilingual dictionary using a Japanese/English and French/English bilingual dictionaries.

Exploiting also bilingual dictionaries for building a multilingual large-scale lexical knowledge base, [Knight & Luk 94] describe the algorithms for merging complementary structured lexical resources from WordNet, LDOCE and a Spanish/English bilingual dictionary. They focus on the construction of Sensus, a large knowledge base for supporting the Pangloss machine translation system, merging ontologies (ONTOS [Nirenburg & Defrise 93] and Upper Model [Bateman 90]) and WordNet and monolingual and bilingual dictionaries. [Okumura & Hovy 94] describe (semi)automatic methods for associating a Japanese entries to an English ontology using a Japanese/English bilingual dictionary. [Rigau & Agirre 95] describe several methods for linking Spanish and French words from bilingual dictionaries to WordNet synsets and [Atserias et al. 97] using also information collected from DGILE show that combining the results provided by each method the total amount of useful data grows out of 40%. Furthermore, [Farreres et al. 98] propose a way to take profit of the taxonomy structure acquired from a Spanish monolingual MRD as a knowledge source to the mapping process.

[Ageno et al. 94] describe a semiautomatic environment for linking DGILE and LDOCE taxonomies using also a bilingual dictionary. Following these ideas, [Rigau et al. 95] describe a more complex, complete and automatic mechanism for linking LDOCE and DGILE taxonomies using also a Spanish/English bilingual dictionary and the notion of Conceptual Distance between concepts.

Using also a Spanish English bilingual dictionaries and WordNet, [Rigau 94] describe an automatic method to enrich semantically the monolingual Spanish dictionary DGILE. [Rigau et al. 97] describe eight different heuristics for the Genus Sense Disambiguation problem using knowledge acquired from the monolingual dictionary, the bilingual Spanish/English dictionary or even WordNet. [Rigau et al. 98] describe a three step method to collect accurate taxonomies from monolingual MRDs using also Spanish/English bilingual mappings and WordNet.

Not only monolingual and bilingual MRDs have been used as structured lexical knowledge resources. [Byrd 89] uses the New Collins Thesaurus and Roget's II: The New Thesaurus and several MRDs for creating a knowledge base and increasing the semantic capability of NLP systems. [Yarowsky 92] uses the Roget's Thesaurus for collecting coocurrence data from the Grolier Encyclopaedia. [Chen & Chang 98] use LLOCE and Roget's thesaurus to label LDOCE.

Using both structured and non-structured lexical resources. Combining thesaurus and corpora, [Greffenstette 93] uses the Roget's and Macquarie's thesaurus as standards for evaluating automatic semantic extraction techniques from corpora. [Utsuro et al. 93] use a Japanese online thesaurus for describing as semantic categories the semantic restrictions of case slots for acquiring surface case frames of Japanese verbs from bilingual corpora.

[Pustejovsky 92] proposes a method for sublanguage corpus analysis beyond that available from the seeding of MRDs. [Pustejovsky et al. 93] present an approach to acquisition of lexical semantic knowledge using firstly a lexicon obtained from MRDs and afterwards refined by inspecting a corpus of text processed by several tools. They propose systems to detect and extract several types of semantic information: metonymy, taxonomic relations, noun's qualia structure, coercive environments, etc.

The most successful combination between structured and non-structured lexical resources for acquiring lexical knowledge from corpora (due to the large amount of studies carried out) seems to be using in combination to WordNet. Thus, [Greffenstette & Hearst 92] use WordNet for acquiring hyponymic relations from corpora and [Resnik & Hearst 93] for resolving the prepositional phrase attachments from sparseness data. [Resnik 93], [Ribas 95] and [Wu & Palmer 95] use WordNet for acquiring selectional restrictions from phrasal analysed corpora. In a similar approach, [Li & Abe 95] propose as a generalisation method the MDL (minimum Description Length) principle rather than the MI (Mutual Information) or AR (Association Ratio). [Basili et al. 95] use WordNet verbal taxonomy for testing acquired taxonomies from corpus. [Hearst & Schütze 95] use coocurrence statistics in combination with a set of flat categories derived from WordNet to classify proper nouns in text.

WordNet has also been used widely for disambiguating words in context. That is, enriching words in a text with its corresponding senses. Thus, [Miller & Teibel 91] propose the use of WordNet to estimate the semantic distance for polysemous words in context. SemCor [Miller et al. 93], a semantically tagged (with WordNet synset tags) version of the Brown Corpus, was provided by the Princeton group as a benchmarks for the automatic sense identification [Miller et al. 94]. [Resnik 95] uses the notion of semantic similarity of WordNet synsets to disambiguate groups of closely related nouns. [Sussna 93] describes an ad-hoc weighting mechanism on WordNet for disambiguating nouns in a text. In a similar approach, [Agirre & Rigau 96b] describe an unsupervised algorithm using the Conceptual Density formula for disambiguating nouns in SemCor. Using the most frequent sense per word information placed currently in WordNet1.5, [Peh & Ng 97] report an overall accuracy of 73.61%.

Other researchers have disambiguated words in contexts using MRDs rather than WordNet. Thus, [Cowie et al. 92] use the simulate annealing technique to disambiguate sentences against LDOCE senses. [Wilks et al. 93] use the coocurrence data extracted from LDOCE for constructing word-context vectors and thus, word sense-vectors for disambiguating non-POS disambiguated ocurrences of the word bank. [Liddy & Paik 92] use the LDOCE subject semantic codes and the Wall Street Journal corpus, for computing a subject-code correlation matrix among them. [Harley & Glennon 97] use an ad-hoc weighting mechanism to the different sources of lexical knowledge present into the completely coded Cambridge International Dictionary of English (CIDE).

Several attempts have been proposed combining bilingual corpora and bilingual MRDs. Thus, [Utsuro et al. 93] describe a method for acquiring surface case frames of Japanese verbs from bilingual corpora using bilingual MRDs; [Utzuro et al. 94] describe a unified framework for bilingual text matching by combining bilingual dictionaries and statistical techniques. [Klavans & Tzoukermann 96] present the Bicord (Bilingual Corpus-enhanced Dictionaries) system which involves linking entries from Collins Robert bilingual French/English MRD to the Hansard corpus for the creation of a bilingual LDB.

2.8 Main International Projects on Lexical Acquisition

In the research community there are an increasing interest on the lexicon. There are many projects and research groups in Europe, United States and Japan undertaking research on the construction of large-scale lexical resources for NLP as well as with various organisations and working groups created to provide the infrastructure to support such work[1]. A short overview of some of these projects is presented bellow.

• Japanese Projects

EDR. The Japan Key Technology Centre (a government agency) and eight private computer manufacturers established, in 1986, the Japan Electronic Dictionary Research Institute, Ltd. (EDR) [Yokoi 95] for developing during nine years a set of large-scale multilingual dictionaries mainly oriented to Machine Translation. The EDR architecture is divided in three layers. The data layer contains the English and Japanese Corpus with 250,000 words each. The surface layer contains monolingual, bilingual and coocurrence dictionaries for both languages. The monolingual dictionary has 200,000 general vocabulary and 100,000 technical terminology. The deep layer contains the concept dictionary with 400,000 concepts.

• American Projects

Comlex [Grishman et al. 94] is a broad coverage English lexicon (with about 38,000 lemmas) developed at New York University under LDC sponsorship. It contains detailed information about the syntactic characteristics of each lexical item, and is particularly detailed in its treatment of subcategorization (complement structures). It includes 92 different subcategorization features for verbs, 14 for adjectives, and 9 for nouns. These features distinguish not only the different constituent structures which may appear in a complement, but also the different control features associated with a constituent structure.

WordNet [Miller 90] is an on-line lexical database for English developed in Princeton. Current version, 1.6 contains more than 123,000 different words and more than 99,000 different word senses. Furthermore, WordNet includes eight different semantic relations (synonymy, antonymy, hyponymy, meronymy, troponymy and entailment) represented as links (more than 116,000) between word senses. Princeton group also provides SemCor, a sense tagged (with WordNet senses) part of the Brown Corpus [Miller et al. 93].

Pangloss. This research addresses the creation and use of large concept taxonomies and Ontologies for natural language processing and other applications by combining online resources such as dictionaries and thesauri, statistical methods over text, and traditional human knowledge acquisition interfaces. In particular, creating and organizing 70,000-item concept taxonomy for use in the Pangloss Machine Translation systems, the Penman sentence generation system, and eventually other systems as appropriate.

The topmost levels of the Ontology, called the Ontology Base (OB), consist of approx. 400 terms. The OB is a merge of the Penman Upper Model (based on Systemic-Functional Linguistics), the top-level ONTOS ontology (a semantic network; see [Nirenburg & Defrise 93]), and, for nouns, the LDOCE semantic categories. The function of the Ontology Base and its relation with the Interlingua are described in [Hovy & Nirenburg 92].

The primary source for the Ontology body is the semantic database WordNet [Miller 90]. To construct the main body of the Ontology, work was performed to automatically connect WordNet concepts and English lexical items by discovering pairs of corresponding senses (see [Knight 93] and [Knight & Luk 94]).

---

1 *American institutions such as the Consortium for Lexical Research (CRL) or the Linguistic Data Consortium (LDC), the Japanese EDR consortium or the European Language Resources Association (ELRA).*

In addition to housing the symbols to represent semantic meaning, the Ontology contains pointers from each symbol to appropriate lexical items in various languages (mainly, Japanese and Spanish). The Penman English lexicon currently contains about 50,000 spelling forms (corresponding to approx. 90,000 words); the Japangloss Japanese lexicon contains over 120,000 words.

Cyc [Lenat 95] started in 1984 a long term project (two person-century of effort) for codifying manually the common-sense knowledge needed for representing and using appropriately the knowledge contained in a 1-volume encyclopaedia. Currently, Cyc contains 100,000 concepts and 1,000,000 common-sense axioms describing human reality.

• European Projects

Multilex (Esprit II project 5304) [McNaught 90] was focused on stabilising a standard for multilingual and multifunctional lexicons for the European languages.

Genelex (Eureka) [Normier & Nossim 90] produce a generic and application specific lexicon according to unified lexical models.

Acquilex I and **II** (Esprit projects 3030 and 7315) were directed towards the acquisition of lexical information from monolingual and bilingual MRDs and from text corpora for NLP applications, and to create a prototype of a lexical knowledge base formalism.

LE-Parole will produce a large-scale harmonised set of corpora and lexicons for all European Union Languages. The resources will be produced in a standard format supporting selection and customisation. The lexicons (20K entries per language) will conform to a model based on Eagles guidelines and Genelex results, underlying a common lexical tool adapted from the Genelex project. The corpus part of the project will produce large (at least 20 million words) monolingual harmonised corpora in a common mark-up conventions. Part of the corpus will also be morphologically tagged, with tagsets compatible with the lexicons.

Sparkle (LE project 2111). One of the main goals of this project is to develop a lexical acquisition system capable of learning the aspects of word knowledge from free text, which are needed for NLP. This system will work on the output of the shallow parsers buit in first place of the project to extract lexical knowledge about semantic classes of predicates, subcategorizaion, argument structure, preferentional selectional restriction and diathesis alternation for the language of focus.

EuroWordNet (LE project 4003) [Vossen in Press]. The aim of this project is to develop a multilingual database with basic semantic relations between words (that is, WordNets) for several European languages (Dutch, Italian and Spanish). These European WordNets will as much as possible be built from available existing resources and databases with semantic information developed in various national and EU-projects (Acquilex, Sift, etc.). This will not only be more cost-effective but will also make it possible to combine information from independently created resources, making the ultimate database more consistent and reliable, while keeping the richness and diversity of the vocabularies of the different languages. The WordNets will be stored in a central lexical database system and the word meanings will be linked to meanings in the Princeton WordNet1.5. Furthermore, we will merge the major concepts and words in the individual wordnets to form a common language-independent ontology, while language specific properties are maintained in the individual WordNets. This will guarantee compatibility and maximise the control over the data across the different wordnets while language-dependent differences can be maintained in the individual WordNets. The database will be used for multilingual information retrieval that will be demonstrated by Novell Linguistic Development.

Chapter 3

The Methodology and SEISD

3.1 Introduction

This chapter provides a global picture of the whole methodology for creating the MLKB from monolingual and bilingual MRDs and overviews the arquitecture of SEISD (*Sistema d'extracció d'informació Semàntica de Diccionaris*), the software system we designed and developed for supporting this methodology. Thus, the main aim of this Chapter is to provide a clear vision of the tasks performed by the SEISD environment. Each task is described in detail. The problems related to each task are faced in Chapter 4 and the solutions we provide are described in Chapter 5 and 6. All the examples that ilustrate the steps of the methodology are taken from DGILE (*Diccionario General Ilustrado de la Lengua*), the specific dictionary we selected to perform the whole acquisition process. Some decisions that affect the design of the methodology are justified from the analysis of such dictionary. After introducing, in Section 2, the main methodological considerations, Section 3 explains the main objectives of SEISD. The components of this environment are briefly described in Sections 4 and 5. Section 6 is devoted to the semantic knowledge acquisition process and Section 7 to the mapping process of the acquired knowledge onto the LKB. In Section 8, the multilingual knowledge acquisition task is presented, and finally Section 9 describes the exploitation and validation process of the acquired lexical knowledge.

3.2 Methodology

Four main issues were considered when designing the base methodology: a) the characteristics of the lexical resources used, b) the implicit and explicit information to be extracted from them, c) how to represent and exploit the information extracted and d) how to carry out the acquisition processes.

3.2.1 Lexical knowledge sources used

A brief description of the main lexical knowledge sources used is provided below.

• **DGILE.** The main characteristics of the **DGILE** dictionary are detailed in [Castellón et al. 91]. This MRD contains 89,043 entries, 157,842 senses (1.77 senses per entry) and more than 1.4 million words in definitions and examples. For instance, the lexical entry for the word "vino" (wine) in DGILE is:

vino (l. vinu) *m*. Zumo de uvas fermentado; ... *2* fig. *Bautizar* o *cristianizar*, el ~, echarle agua. *3* fig. *Dormir uno el* ~, dormir mientras dura la borrachera; *tener uno mal* ~, ser pendenciero en la embriaguez. *4* p.ext. Zumo. | HOMOF.: vino (v.) , bino (v.) .
REL. **Enológico**, **enólogo**, **enotecnia**, derivados de *enología*, ciencia de la vinicultura, formada del gr. *oinos*.

This MRD has been processed in order to produce several MTDs [Wilks et al. 96] used in subsequent steps (detection of the genus term, morphological analysis, Genus Sense Disambiguation, etc.). Among others, the lispified version loaded into the LDB system [Carroll 90a], the part-of-speech lexicon used by the SegWord morphological analyser [Sanfillipo 90] (one of the morphological modules of MACO [Acebo et al. 94]) and MACO+ [Carmona et al. 98], a huge lexicon containing the relatedness between word forms, enriched with some statistical measures (such as Mutual Information (MI) [Church & Hanks 90] or the Association Ratio (AR) [Ribas 94]) on the bidirectional cooccurrence of pairs of words in all the definition fields of DGILE, and several other frequency lexicons generated from the definition field of DGILE, such as the frequency of each word form, bigrams, trigrams, etc. (see appendix).

• **WordNet**, a large public domain on-line lexicon based on psycholinguistic theories [Miller 90] which attempts to organize lexical information in terms of word meanings, rather than word forms. In this respect, it resembles a thesaurus or a Lexical Knowledge Base more than a dictionary. Currently, WordNet1.6[1] represents 123,497 different content words and 99,642 senses related by a set of semantic relationships (among others: hypo/hypernymy, mero/holonymy, etc.). Senses in WordNet are represented by means of synonym sets or synsets. For instance, the lexical overview for wine is:

The noun wine has 2 senses (first 2 from tagged texts)

1. {05916701} <noun.food> **wine**, vino -- (fermented juice (of grapes especially))
2. {03880346} <noun.attribute> **wine**, wine-colored -- (a red as dark as red wine)

The verb wine has 1 sense (no senses from tagged texts)

1. {00809609} <verb.consumption> **wine** -- (drink wine)

From left to right, synset number (location in the datafile), semantic file, synonym set and gloss.

• **Bilingual Dictionaries**. Several MTDs have also been obtained from the two sides of the bilingual dictionary used in this thesis. The lispified version of both has been loaded into the LDB [Hastings et al. 94]. Briefly, the Spanish/English dictionary **EEI** contains 16,463 entries and 28,002 translation fields, while the English/Spanish **EIE** contains 15,352 entries with 27,033 translation fields. For instance, the lexical entries for vino (EEI) and wine (EIE) are:

vino *m* wine. • ~ **de Jerez**, sherry; ~ **tinto**, red wine.

wine *n* vino

By merging both directions of the nominal part of the bilingual dictionary we preduced another MTD we called **HBil** (from harmonized bilingual). This dictionary contains 28,129 connections between 14,879 Spanish nouns and 15,848 English nouns.

• **Type System**. The multilingual conceptual representation language used in the LKB [Copestake 92b] is based on a Feature Structure formalism for representing lexical entries constrained by appropriateness conditions fixed by a preexisting Type System (TS). As the aim of our methodology is to build such lexical entries in terms of the classes of lexical items represented in the Type System, it constitutes an obvious knowledge source for this task. Currently, the Type System consists of 527 types and 196 different features.

---

[1] *Although WordNet 1.6 is now available, the version we used in this work is WordNet1.5.*

3.2.2 Lexical knowledge to be extracted

The basic aim of this study is to examine the feasibility of automatically obtaining large scale semantic information from MRDs that contain this information in explicit or implicit form [Atkins et al. 86]. It is clear that the explicit information attached to each lexical entry could be obtained quite straightforwardly (e.g., part of speech, topic domains, uses, etc.), but more sophisticated techniques are necessary to deal with other, implicit information placed in the MRDs. There are some interesting relations between lexical items (see [Calzolari 88] or [Byrd 89]) which could be extracted automatically (or at least semi-automatically) from MRDs in order to build a complete MLKB of use for NLP systems:

- Hypernym/hyponym relations (class-subclass relations).
- Synonymy/antonymy relations (equivalent/contrary relations).
- Meronym/holonym relations (part/whole relations, element/set relations, etc.).
- Case roles relations (agentive role, telic role, etc.).
- Content relations (qualia structure, form description, constitutive components, etc.).
- Collocational relations (compound words, collocations, idioms, etc.).
- Selectional restrictions (typical subject, typical object, etc.).
- Translation equivalences (in bilingual dictionaries).

As was said in Section 1.4, the most important relation to be extracted from an MRD is the hypo/hypernym relation (e.g., [Amsler 81], [Vossen & Serail 90], [Bruce & Guthrie 91], [Copestake 92b]) between dictionary senses, not only because of its own importance but also because this relation acts as a support for the main inheritance mechanisms, thus facilitating the acquisition of other relations and semantic features [Cohen & Loiselle 88], providing formal structure and avoiding redundancy in the lexicon [Briscoe et al. 90]. This implicit relation emerges from the inherent structure of conventional dictionaries. Thus, following the natural chain of dictionary senses described in DGILE from **rosa_1_1** (rose), we can discover that a rose is a part of a living vegetable organism.

rosa_1_1	**flor** del rosal. (*rose: flower of the rosebush*)	CLASS
flor_1_1	**órgano** reproductor de las plantas... (flower, organ for the reproduction of the plants...)	CLASS
órgano_1_3	conjunto de unidades funcionales de un **organismo** celular... (organ, functional units of a celular organism...)	PART-OF
organismo_1_2	**ser** viviente (*organism, living thing*)	CLASS
ser_1_2	**ente** (que vive) (*live entity*)	CLASS
ente_1_1	**algo** que es, existe o puede existir. (*something which exists or can exists*)	CLASS

The hypo/hypernym relation appears between the entry word (e.g., *rosa*) and the head word, genus term or core of the phrase (e.g., flor) of the definition, in which other case roles relations can be found from the differentia (prepositional modifiers, adjectives, adverbs, etc.) [Calzolari 91]. Thus, in the ideal case, a dictionary definition is written to employ a genus term combined with a differentia which distinguishes the word being defined from other words with the same genus term.

Usually, the information for a lexical unit contained in an MRD cannot be gathered from one entry but by processing the whole dictionary [Wilks et al. 96]. That is, the information is not located in one piece of information but is distributed throughout the dictionary (e.g., [Fox et al. 88], [Dolan et al. 93], [Vanderwende 95]) or even several dictionaries (e.g., [Byrd 89], [Veronis & Ide 91], [Klavans & Tzoukermann 96], [Richardson 97]).

3.2.3 Lexical knowledge representation

In order to represent the LKB we used LAUREL [Copestake 92b] which was developed as the Lexical Representation Language (LRL) for the Acquilex LKB system. LAUREL is a graph unification based representation language that offers the flexibility to represent both syntactic and semantic information formally in a way which could be easily integrated with much current work on unification grammar, parsing and generation.

The type Feature Structure formalism, based on [Carpenter 92], has been augmented to allow the lexicon to be structured hierarchically using default inheritance mechanisms (that which is stated for lexical entries at the highest levels can be inherited by concepts at more specific levels), lexical and phrasal rules, multilingual relations, etc.

Although the LKB provides a representation language and defines valid operations on entries, the LKB system capabilities have been augmented to allow flexible access to classes of LKB lexical entries by any information contained in it for consulting and testing the information extracted and represented in the lexicons [Rigau et al. 94].

3.2.4 General methodology

Most of the relations between lexical items described in Section 3.2.2 are semantic relations. That is, relations between word meanings, and relations between word meanings and word forms. Although the division of word meanings into dictionary senses and their classification are frequently arbitrary (e.g., [Atkins & Levin 88], [Kilgarriff 93], [Kilgarriff 97]) and dictionary entries for polysemous words are usually very closely related and encode fine-grain semantic distinctions that are unlikely to be of practical value for NLP [Dolan 94], dictionary senses have been considered as the semantic units because they represent different conceptual entities (with the corresponding descriptions) of the same word. Other approaches have used words instead of senses (e.g., word hierarchies [Nakamura & Nagao 88], semantic networks [Jensen & Binot 87], [Fox et al. 88], [Ricardson 97], the Pathfinder network [Wilks et al. 93], multidimentional space vectors [Niwa & Nitta 94]), avoiding the Genus Sense Disambiguation (GSD) problem (a central issue of this thesis, see Section 5.3).

The methodology we applied attempts to derive an MLKB from the monolingual and bilingual dictionaries, using mainly the descriptive approach (see Section 1.4 and 2.4) [Ageno et al. 92b] and roughly performs as follows. First, the monolingual and bilingual MRDs are loaded into a standardized Lexical Data Base (LDB). Once the MRD is transformed into a Machine-Tractable Dictionary (MTD)[1] and placed in a Lexical Data Base (LDB) the different dictionary access strategies could be exploited. Secondly, separate semantic taxonomies are derived (semi)automatically from monolingual LDBs for common subsets of vocabulary. Furthermore, the complete noun taxonomy of DGILE can also be derived without human intervention (see Sections 5.2 and 5.3). Later on, by parsing the dictionary definitions attached to the taxonomy senses, a richer knowledge about the defined concept is obtained. In this process the lexical knowledge acquired in previous steps (i.e., sense disambiguated taxonomies) is also used to perform a deeper semantic analysis (see Section 5.4). The lexical knowledge acquired from dictionary senses could then be placed in monolingual LKBs. Using bilingual LDBs, the monolingual LKBs can be linked to create an MLKB (see Chapter 6). Figure 3.1 depicts this approach to construct MLKBs from MRDs.

[1] *Most dictionaries are only available as machine-readable in typesetting format for editing purposes. In this case, automatic analysis must be performed to obtain a machine-tractable format. The transformation process done for DGILE is described in [Castellón et al. 91].*

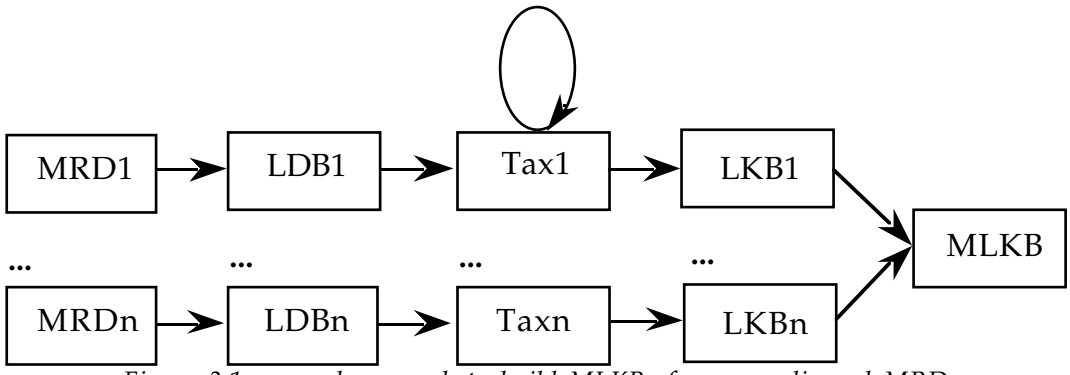


Figure 3.1, general approach to build MLKBs from monolingual MRDs.

Nevertheless, the methodology applied in this work is not completely descriptive. Our methodology also prescribes a minimal set of primitives to represent the main ontological concepts of the LKB. Thus, our approach is a mixed one, taking advantage of those described in Section 1.2.3. Chunks of taxonomies, derived from the dictionary, are assigned to the prescribed ontological concepts. This approach allows us to avoid the circularity problem in dictionary definitions [Amsler 81], assigning intermediate concepts from the dictionary to top ontological structures from the Type System and providing structure, knowledge and inheritance mechanisms to the class of concepts (see Sections 4.2 and 5.2 for more details).

The semantic relations contained in dictionary definitions [Boguraev & Pustejovsky 89a] can be extracted by means of a semantic analysis of such definitions (e.g., [Jensen & Binot 87], [Fox et al. 88], [Briscoe & Carroll 91], [Vossen 92], [Bruce et al. 92], [Dolan et al. 93], [Vanderwende 95]) using a broad-coverage morphological, syntactic and semantic parsing software. For English several such broad-coverage parsing engines exist (e.g., PEG [Jensen 86], Core Language Engine [Alshawi 92], Alvey Natural Language Tools [Grover et al. 93], etc.). Woefully, this is not the case for Spanish (and other many languages), where available wide-range tools for NLP are limited to morphological analysers and part-of-speech taggers (e.g., [Acebo et al. 94], [Farwell et al. 95], [Sanchez & Nieto 95], [Padró 98]). Building such tools is beyond the scope of this work, as it would require the creation of several specialized grammars ([Ageno et al. 91a], see also [Hagman 92]) for parsing entries belonging to different semantic classes (e.g., SUBSTANCE, FOOD, PERSON, INSTRUMENT, etc.).

Consider, as an illustration of the acquisition process performed by **SEISD**, the lexical entry *rioja_1_1*, belonging to the taxonomy *bebida* (drink). The methodology is divided into six partial steps.

1) First, the top dictionary senses that cover the semantic classes to be represented in the LKB are selected and assigned to the appropriate semantic type (see Sections 4.2 and 5.2 for a detailed discussion of this issue). At this stage, for the taxonomy derived from *zumo_1_1* (juice) the dictionary sense root is attached to the **c_art_subst** (comestible-artifact-substance) type[1].

Word sense: zumo_1_1
Attached-to: **c_art_subst** type.
Definition: **líquido** que se extrae de las flores, hierbas, frutos, etc. (*liquid extracted from flowers, herbs, fruits, etc.*).

2) Exploiting the implicit hypo/hypernym relation, the sense disambiguated taxonomies are generated, collected, validated and attached to the same semantic class represented in the Type System. This task is performed by the **TaxBuild** (Taxonomy Builder) module of SEISD (see Sections 3.6.1.2, 3.6.1.3, 4.3 and 5.3). An example of a hypernym dictionary sense of *rioja_1_1* following the hypernym chain through *vino_1_1* is *zumo_1_1*:

[1] *See Section 3.5.2 for a description of the Type System supporting the LKB.*

Word sense: vino_1_1
Hypernym: zumo_1_1
Definition: **zumo** de uvas fermentado (*fermented juice of grapes*).
FPar: ((CLASS ZUMO))

Word sense: rioja_1_1
Hypernym: vino_1_1
Definition: **vino** de Rioja (*wine from Rioja)*.
FPar: ((CLASS VINO))

3) For each semantic class, the different conceptual nodes attached to it are processed in order to obtain deeper knowledge of the case roles relations and content relations appearing in the differentiae. This process is carried out by the **SemBuild** (Semantic Builder) of SEISD (see Sections 3.6.1.4, 4.4 and 5.4). For instance, the definition of *rioja_1_1* gives a meaning of the word *rioja* as a kind of wine (the genus term) but different from the other wines because it is made in a particular region (the differentiae).

Word sense: zumo_1_1
Attached-to: **c_art_subst** type.
Definition: **líquido** que se extrae de las flores, hierbas, frutos, etc. (*liquid extracted from flowers, herbs, fruits, etc.*).
SinPar: [SN: [n: líquido],
SW: [p0r: que],
SV: [x: se,
v0v: extraer],
SP: [r0d: de,
SN: [n: flor.
n: hierba,
n: fruto]]].

Word sense: vino_1_1
Hypernym: zumo_1_1
Definition: **zumo** de uvas fermentado (*fermented juice of grapes*).
SinPar: [SN: [n: zumo,
SP: [r0d: de,
SN: [n: uva,
a: fermentado]]]].

Word sense: rioja_1_1
Hypernym: vino_1_1
Definition: **vino** de Rioja (*wine from Rioja)*.
SinPar: [SN: [n: vino],
ORIGIN: [w: rioja]].

4) This enriched taxonomy is then represented in the LKB formalism in order to exploit the inheritance and other inferential mechanisms that make explicit, for instance, the inherited properties of the hypernym lexical entries. This task is performed mainly by the **CRS** (Conversion Rule System) of SEISD (see Sections 3.6.2, 4.4.2 and 5.4). In our example, when the analysed *rioja_1_1* is placed as a lexical entry into the LKB lexicon all the local and inherited information acquired (or represented in the Type System) is available. That is, a rioja is a fermented liquid derived from grapes and produced in Rioja.

rioja x_1_1
< lex-noun-sign rqs > < vino_X_I_1< lex-noun-sign rqs >
< lex-sign sense-id : sense-id dictionary > = ("VOX")
< lex-sign sense-id : sense-id word > = (“rioja”)
< lex-sign sense-id : sense-id homonym-no > = (“1”)
< lex-sign sense-id : sense-id sense-no > = (“1”)
< rqs : origin-area > = (“rioja”)

5) Once a semantic class of lexical entries have been generated and placed in the LKB for the different languages, the acquisition of multilingual lexical information by means of the multilingual connection between lexical entries can be performed. This process is carried out by the **TGE** (Tlinks Generation Environment) module of SEISD (see Sections 3.6.3 and 4.5 and Chapter 6). For instance, using the knowledge placed in the bilingual dictionaries, the following links can be generated for *rioja_x_1_1* and lexical entries of LDOCE (a similar mechanism has also been used, see 6.3.3, for linking lexical entries to WordNet synsets):

rioja_x_1_1	linked to	wine_l_1_1	(by means of parent tlink).
rioja_x_1_1	linked to	drink_l_2_1	(by means of grandparent tlink).

6) Finally, when the extraction process ends, the lexical knowledge acquired must be validated and tested in order to look for incompleteness (for instance, daughter lexical entries with no differences between them), in order to perform further acquisition cycles or add new information manually. This process aided by the **LDB/LKB system** enhancement (see Sections 3.6.4 and 4.6).

Finally, our methodology has been regarded as being evolutionary [Vanderwende 95]. That is, our methodology acquires knowledge as a result of a stepwise refinement (i.e., by allowing the user to inspect each step cycle of new knowledge being acquired).

3.3 The Main objectives of SEISD

The main reason for designing and implementing SEISD (*Sistema d'Extracció d'Informació Semàntica de Diccionaris*) was to build a modular system capable of performing the whole process of exploiting monolingual and bilingual MRDs by creating an MLKB following the general methodology outlined above. The system design involves both methodological and technical considerations. The most important features of SEISD are:

- SEISD was designed to support the main methodology (see Section 3.2).
- The underlying methodology for semantic extraction from MRDs has been developed taking into account the characteristics of the MRDs and other lexical resources used.
- SEISD was built to extract lexical knowledge from MRDs with minimal effort and minimal human intervention. In fact, several subsystems of SEISD allow both interactive and batch modes with different level of human intervention ranging from (semi)automatic to fully automatic.
- The modular design of SEISD allows each module to be enriched with different approaches and techniques.
- System performance is controlled by a set of informed heuristics.
- Great attention has been paid to the reusability of software and lexical resources. In fact, the system has fully integrated previous Acquilex representational formalisms and their supporting NLP software tools.

A central guideline was to build the whole system so as to perform each process semi-automatically. A first version of SEISD was built and used within the Acquilex project. Results of the (semi)automatic use of SEISD are reported in [Castellón 93] and [Taulé 95]. We

report now in this thesis improvements both in methodology and techniques for efficacy (to obtain more information) and efficiency (to obtain this information more easily).

3.4 SEISD architecture

SEISD[1] supplies a user-friendly interface to several subsystems and multiple sources of massive and heterogeneous data, and also a way of integrating them. SEISD was designed as a medium for the acquisition methodology, and is fully integrated with the Acquilex lexical representational formalisms and their supporting software tools[2]. The whole system has been implemented in Common Lisp for Macintoch.

Figure 3.2 shows the most important modules of SEISD (right) and the knowledge sources (left) used by the system.

The SEISD environment provides full coverage to the general approach described in Section 3.2 to build MLKBs from monolingual and bilingual MRDs and exploit them. In this section, after introducing the common subsystems used in the various stages of the acquisition process, the main functions covered by SEISD are described. To date, the subsystems included in SEISD are:

- **LDB** (Lexical Data Base) [Carroll 90a] or [Carroll 92], a database-like system providing flexible access to dictionary entries via any of the information contained in the MRD.
- **LKB** (Lexical Knowledge Base) [Copestake 92a], a system developed to represent lexical entries by means of typed Feature Structures constrained by a Type System.
- **PRE** (Production Rules Environment) [Ageno et al. 93], a rule-oriented general purpose interpreter adapted to natural language applications.
- **LispWN**, a Lisp interface to WordNet allowing access to all the lexical knowledge stored in it.
- **TaxBuild** (Taxonomy Builder), a system that allows the (semi)automatic (see [Ageno et al. 92b]) or fully automatic (see [Rigau et al. 97] and [Rigau et al. 98]) construction and validation of taxonomies of senses from the LDB (selection of the genus term and resolution of the lexical ambiguity). That is, process 2 of those presented in Section 1.4. Currently, this system uses:

  - **SegWord** [Sanfilippo 90], a morphological analyser based on the morphological validations of headwords contained in the LDB.
  - **FPar** (Flexible Parser) [Carroll 90b], a syntactic-semantic analyser based on [Alshawi 89] that allows partial analysis of the dictionary definitions.

- **SemBuild** (Semantic Builder), a system that allows the acquisition and validation of lexical knowledge from the differentiae. That is, process 3 of those presented in Section 1.4. This system currently uses the MACO+ morphological analyser [Carmona et al. 98] and the Relax tagger [Padró 98] and includes:

---

[1] *SEISD is a software environment that was used (mainly by the Spanish team) within Acquilex and Acquilex II projects. SEISD was designed by the author of this thesis although other researchers participate in the construction of several modules of the system as builders and/or users (so must of the references of SEISD are co-authored). If no explicit indication of the contrary is made, the author is the designer (and also the main programmer) of all the components of the system.*

[2] *The subsystems developed within the Acquilex project by the Computer Laboratory of Cambridge University (CCL) and reused in the SEISD environment are the* ***LDB*** *System, the* ***LKB*** *System, the* ***SegWord*** *morphological analyser and the* ***FPar*** *syntactic-semantic analyser.*

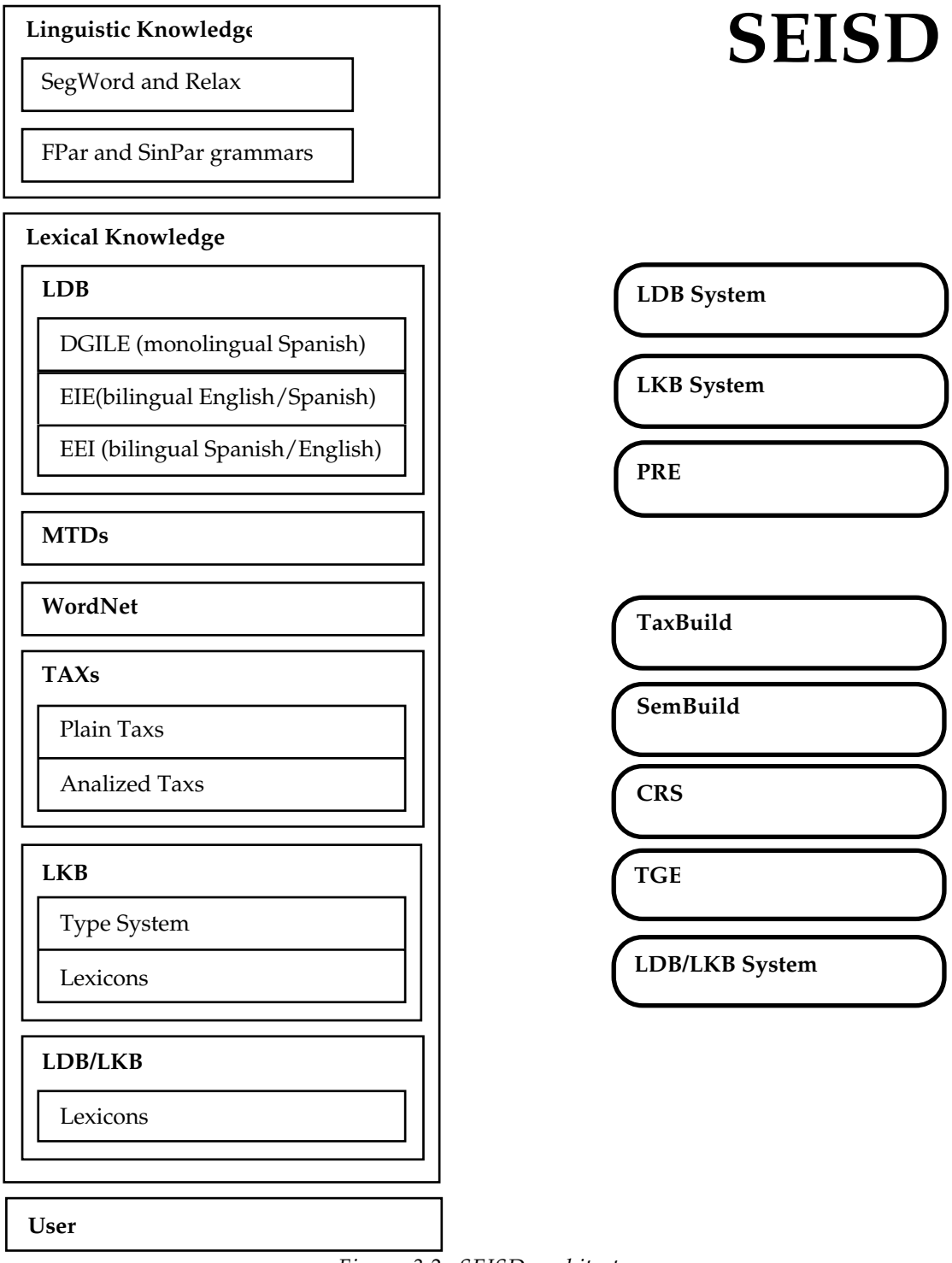


Figure 3.2, SEISD architecture.

• **SinPar** (Sintagm Parser), a shallow parsing tool implemented as a DCG grammar [Pereira & Warren 80] for parsing completely all dictionary definitions which provides for an input definition a fully analysed set of chunks (nominal, verbal and prepositional sintagms).

• **CRS** (Conversion Rule System) [Ageno et al. 92d], an interactive system that allows the (semi)automatic conversion of the information extracted from the LDB to lexicons included in

the LKB system. This module performs process 4 of the general methodology presented in Section 1.4.

• **TGE** (Tlinks Generation Environment) [Ageno et al. 94], an interactive system that allows the creation of Translation Links (tlinks) between lexical entries of several languages (semi)automatically. That is, process 5 of the general methodology presented in Section 1.4 (see Chapter 6 for more details).

• **LDB/LKB integration** [Rigau et al. 94], a system that increases the functions provided by the LKB system with flexible LDB-like access to classes of lexical entries via any of the information they contain. This module performs process 6 of the general methodology presented in Section 1.4.

3.5 Common subsystems used in SEISD

Four subsystems are used in several modules of SEISD. The LDB and LKB systems have been used by all the Acquilex partners and were developed by the Computer Laboratory of Cambridge University (CCL). Both systems were designed for representing lexical entries. PRE is a rule-oriented general purpose interpreter designed to be used in natural language applications and was developed by the Natural Language Group of the LSI department to provide flexible processing. PRE has been used in several parts of SEISD (CRS and TGE). PRE has been also used by the MACO morphological analyser [Acebo et al. 94], and other NLP systems [Gatius & Rodríguez 96] and [Turmo 97].

3.5.1 LDB

Within Acquilex, the Lexical Data Base (LDB), which implements the two-level dictionary access model [Boguraev et al. 91], was implemented to provide flexible access to MRDs. The LDB is endowed with a graphic interface which provides a user-friendly environment for query formation and Information Retrieval. It allows several dictionaries to be loaded and queried in parallel.

3.5.2 LKB

The two main components of the LKB are the Type System and the Lexicon. The Type System represented as a type hierarchy defines a partial order ("is more specific than") on the types and establishes consistency conditions. The operations that the LKB supports are (default) inheritance, (default) unification and lexical rule application.

Thus, the LKB provides facilities for creating Type Systems, loading lexicons and displaying fully expanded Feature Structures, type checking, and so forth. A brief description of the LKB system can be found in [Copestake 92a] and a complete one in [Copestake 92b].

3.5.3 PRE

The Production Rules Environment or PRE [Ageno et al. 93] is a rule-oriented general purpose interpreter designed to be used in natural language applications. The PRE follows the philosophy of most production rules systems (e.g., OPS5 [Brownstom et al. 86]). A set of objects is placed in an active data storage device (the Working Memory, WM) and a set of rules manage the WM objects. PRE has been used to implement a flexible and incremental mechanism in the CRS and TGE systems.

PRE rules are grouped into rulesets. Rulesets are identifiable sets of rules with specific control behaviour. The system performance is conducted by the action of a set of control user-defined mechanisms at ruleset or rule level. The capabilities of the system include consultation, modification and creation of WM objects and an expressive unification mechanism that has been added to the system in order to access the objects stored in the WM in a more flexible way.

The PRE system has been implemented to perform some complex functions and particular strategies with multiple and heterogeneous knowledge sources in the most declarative way. The PRE offers a powerful (in terms of both expressiveness and performance) rule application mechanism and provides the possibility of defining higher level mechanisms and control strategies to the mapping process (performed by the CRS, see Section 3.6.2) and multilingual acquisition (performed by the TGE, see Section 3.6.3).

For instance, the following descriptions define two rulesets of CRS, the **top** ruleset and the ruleset **extraction** as a subclass of top. **Standard-sort-proc** refers to a lisp function that performs a default rule sort procedure. All the other slots are self-explained.

```
(ruleset top
      control one-cicle
      sort-proc standard-sort-proc
      sort-type static
      final-cond nil)

(ruleset extraction
      isa top)
```

In order to ilustrate the PRE consider the following TGE rule:

```
(rule rule-1-all
 ruleset all
 control forever
 priority 1
 (translation-in ^trans-records (?translation *rest))
 ->
 (modify 1 ^trans-records (*rest))
 (create translation ^trans-psorts nil ^trans-record ?translation ^tlink-type nil
      ^checked nil))
```

In this rule the pattern-condition is the occurrence of an object named **translation-in** in the WM. This object must contain a **^trans-records** attribute whose value will be matched against the pattern **(?translation *rest)**. If the matching succeeds then **translation** variable will be unified with the first element of the list and **rest** variable with the remainder elements as a list. The action part of the rule consists of two actions. The former is the modification of **translation-in**, popping its first element, and the latter performs the creation of another object, named **translation**. **Rule-1-all** rule is applied until all the objects named "translation-in" have emptied the list contained in their slot **^trans-records**.

3.6 SEISD as a support of the extraction methodology

3.6.1 Semantic knowledge acquisition

The semantic knowledge acquisition function of SEISD is performed by TaxBuild and Sembuild. **TaxBuild** [Ageno et al. 91b], [Ageno et al. 92b] is one of the most important mechanisms of SEISD. This module produces complete disambiguated and partially analysed (using SegWord morphological analyser and FPar syntactic-semantic analysers) dictionary sense taxonomies from DGILE. **SemBuild** acquires semantic knowledge analysing completely the diferentia (using MACO+ morphological analyser, the Relax tagger and the SinPar parsing tool).

3.6.1.1 Analysers used by TaxBuild and SemBuild

As stated above in Section 1.3, the main source of information for obtaining semantic knowledge from MRDs is the dictionary definitions. Therefore, parsing definitions is a fundamental task of the semantic acquisition process. That is, building taxonomies from the genus term and performing an in-depth semantic acquisition from the differentia. Thanks to the modular design of SEISD several morphological and syntactic analysers have been tested. Results of the (semi)automatic use of SEISD reported by [Castellón 93] and [Taulé 95] were performed using SegWord and FPar, while those presented in this thesis use also SegWord and FPar to select the genus term and the Relax tagger and SinPar analyser to acquire semantic knowledge from the differentia.

Thus, in order to detect the genus term of dictionary senses, two different analyses were carried out: the morphological analysis (performed by SegWord), and the syntactic-semantic analysis (performed by FPar).

The most important feature of SegWord is the use of the part-of-speech information located in the LDB to perform the morphological analysis. Every word to be analysed is segmented into a set of pieces and then matched against a set of rules to build a possible lemma which is looked up (if it exists) in the LDB. This program returns for every word a set of possible morphological analyses. There are three kinds of rules, compound, prefix and suffix rules. Consider for instance the next morphological rule:

```
(suffix-rule
    restrictions        final-only
    surface-form        ido
    basic               (er ir)
    category            (V > PARTI))
```

This suffix rule can be applied only to the second and third conjugation regular Spanish verbs producing a participle. Thus, analyzing the Spanish word form **comido** the analyzer detects that is derived from the verb entry **comer** (changing the ending suffix **ido** for **er**) looked up from the LDB.

FPar (Flexible Parser) [Carroll 90b], a syntactic-semantic analyser based on a proposal made in [Alshawi 89], uses a grammar in the form of a hierarchy of patterns. The more general patterns at the top of the hierarchy provide an interpretation of the dictionary sense if more specific, complete and detailed ones lower down fail. In this sense, FPar is a robust dictionary definition parsing tool[1].

Every FPar grammar contains a set of analysis rules and a set of structure building rules. In this case, using the specialized grammar for substances, two main rules have been launched, the analysis and structure rules n-95:

```
(n-95
    (n +p && +noun *0s-adj *0pp-mod &&))

(n-95
    ((class +noun)
    (properties *0s-adj)
    (prep-mod *0pp-mod)
    (r-95)))
```

where && means an arbitrary list of words, +noun stands for exactly one noun and *0s-adj for zero or more adjectives.

These two simple analysers were able to select the genus term of all noun and verb definitions correctly with an accuracy of 97%.

---

[1] *The same schema was used for analysing the definitions of the LPPL dictionary [Artola 93].*

However, for an in-depth analysis of the dictionary definitions we used more powerful analysers. First, we used the MACO morphological analyser [Acebo et al. 94] (currently, we are using MACO+ [Carmona et al. 98]) and the Relax tagger [Padró 98] to analyse and assign a unique morphological tag and lemma to all word forms in DGILE definitions. The system allows for the provision of statistical and manually coded morphological rules for disambiguation. Second, we used SinPar (Sintagm Parser) a robust parser which provides complete syntagm analysis of dictionary definitions. We implemented this parser using a DCG grammar which analyses nominal, verbal and prepositional syntagms[1].

3.6.1.2 Selecting the correct genus term

In order to select the correct semantic head or genus term for noun and verb definitions, a specialized grammar[2] has been developed. Frequently, the genus term for noun definitions is the first noun present in it and for verb definitions the first verb [Amsler 81]. Obviously, there are many cases in which this simple rule for genus detection does not hold [Ageno et al. 91a] (for instance, the linkers [Meijs 90] between headword and the genus term).

In order to ilustrate the whole acquisition process performed by SEISD consider *ojén_1_1* dictionary definition.

The genus detection process of the taxonomy construction carried out by TaxBuild selects *aguardiente* as the genus term of *ojén_1_1* .

Word sense:	ojén_1_1
Definition:	**aguardiente** dulce, anisado (*sweet, anise flavoured liquor*)
FPar:	((CLASS aguardiente))

3.6.1.3 Genus Sense Identification

The correct selection of the genus term of the dictionary definitions makes it possible to build taxonomies of words [Nakamura & Nagao 88], but in order to build semantic hierarchies taking as the semantic units dictionary senses (meanings) rather than words, a costly disambiguation task must be undertaken because dictionary entries for polysemous words encode fine-grain semantic distinctions and typically differ only in some slight shade of meaning [Byrd 89], [Dolan 94]. Then, in order to construct disambiguated hierarchies, once Taxbuild has selected the correct genus term of a dictionary, if the genus term is polysemous, this candidate must be disambiguated against one of the senses of the genus term. This is the Genus Sense Disambiguation (GSD) process [Amsler 81], a particular case of Word Sense Disambiguation (WSD) also called Lexical Ambiguity Resolution (LAR) [Miller & Teibel 91], Word Sense Discrimination (WSD) [McRoy 92], Word Sense Selection (WSS) [Kilgarrif 93] or Word Sense Identification [Miller et al. 94].

Although the WSD process against dictionary definitions is a very difficult task even for humans[3] and the (semi)automatic techniques for GSD have been widely used (e.g., [Amsler

---

[1] *Perhaps a full grammar parser of Spanish could lead to better results, but, because of the sublanguage used in dictionaries and the acquisition goals, a partial analysis (no dependencies between syntagms) does not seem to be a serious limitation.*

[2] *Although it specialised to obtain the genus term, this grammar is domain independent and covers almost all noun and verb definitions.*

[3] *[Wilks et al. 93] say that disambiguating 197 occurrences of the word bank in LDOCE"was not an easy task, as some of the usages of the bank did not seem to fit any of the definitions very well". [Jorgenssen 90] shows interesting psycholinguistic experiments using the Agreement-Disagreement ratio to assign words in contexts to dictionary senses; an error rate of around 10% is found for polysemous words, tagging semantically SemCor by hand [Miller et al. 93]; [Sussna 93], manually assigning WordNet synsets to 544 nouns with context, reports that 22% had more than one sense applicable. [Ng & Lee 96] estimate an error of 10-20% tagging manually 192,800 word occurrences with WordNet synsets. Furthermore, as this*

81], [Vossen & Serail 90], [Ageno et al. 92b], [Calzolari et al. 93], [Artola 93], [Castellón 93]), some attempts of automatic GSD using the semantic codes of the dictionary (e.g., [Copestake 90], [Bruce & Guthrie 92]) or using cooccurrence data extracted from the dictionary itself (e.g., [Wilks et al. 93] and [Schütze 92c]) have been performed. However, for the more general problem of WSD other automatic approaches have also been proposed (see Section 4.3).

Although the first versions of TaxBuild only performed (semi)automatic construction of taxonomies (see the use of this version in [Castellón 93] and [Taulé 95]), we present in this thesis a new version that builds semantic taxonomies without human intervention. Our approach (see Section 5.3) for constructing fully automatically taxonomies from DGILE combines multiple methods (overlapping between definitions, content vectors, conceptual distance, etc.) and structured lexical resources (monolingual and bilingual MRDs, WordNet, etc.).

Considering our piece of hypernym chain again. There is no need for a GSD process for *ojén_1_1* because the genus, *aguardiente*, has only one possible sense in DGILE *aguardiente_1_1* , while for the genus term of *aguardiente_1_1* , *bebida*, a GSD process is necessary in order to select the correct hypernym of *aguardiente* from among four possible senses of *bebida*.

Word sense: aguardiente_1_1
Hyponym-of: bebida_1_3
Definition: **bebida** alcohólica que por destilación se obtiene del vino (*alcoholic drink obtainedby distillation from wine*).

Word sense: ojén_1_1
Hyponym-of: aguardiente_1_1
Definition: **aguardiente** dulce, anisado (*sweet, anise flavoured liquor*).

3.6.1.4 The analysis of the differentiae

Once a disambiguated taxonomy is created by the taxonomy acquisition module of TaxBuild and all the dictionary senses included are connected by hypernym links (except the top ones, which are connected to the Type System) and hyponym links (except the terminal dictionary senses), a further semantic enrichment process can be performed.

Knowledge appearing in the differentia [Calzolari 91] of the definition has to be extracted and assigned to the appropriate semantic roles in the LKB. This task implies a more in-depth analysis of such definitions. Domain-specific grammars have been developed to allow an in-depth acquisition of such semantic information placed in the differentia. The grammars involved must be more complete and complex than those grammars used in the taxonomy acquisition process, which are specialized for the genus detection[1].

Since the information acquired in the semantic acquisition process may be incomplete owing to the partial analysis carried out, the user is provided with an iterative process for improving incrementally the semantic information extracted (see Section 5.4). Viewing the results of the various analyses, the lexicographer can determine how tune the grammar grammar in order to reach richer and more complete results. This cycling process can be carried out as many times as desired.

The following examples are the partially syntactic analysed dictionary senses of *aguardiente_1_1* and *ojén_1_1* using SinPar where SN stands for nominal sintagm, SA for adjectival, SP for prepositional SV for verbal and SW for chunks of words.

---

tagging was also performed on the Brown Corpus (as well as the Wall Street Journal) they compare the subset of the occurrences that overlap. They found only a 57% of agreement with respect to SemCor [Miller et al. 94].

[1] *Currently,, we are using a wide-range tagger of Spanish [Padro 98] and Sintagm Parser for processing the whole dictionary.*

Word sense:	aguardiente_1_1			
Hyponym-of:	bebida_1_3			
Definition:	**bebida** alcohólica que por destilación se obtiene del vino (*alcoholic drink obtainedby distillation from wine*).			
SinPar:	[SN:	[n:bebida,		
		SA:	[a:	alcohólico]],
	SW:	[p0r:	que],	
	SP:	[r0p:	por,	
		SV:	[x:se,	
		v0v:	obtener],	
		SN:	[n:	destilación]],
	SP:	[r0a:	del,	
		SN:	[n:	vino]]].

Word sense:	ojén_1_1			
Hyponym-of:	aguardiente_1_1			
Definition:	**aguardiente** dulce, anisado (*sweet, anise flavoured liquor*).			
SinPar:	[SN:	[n:	aguardiente,	
		SA:	[a:	dulce,
			a:	anisado]]].

A more in-depth conceptual analysis of these analysed definitions can be performed using the semantic knowledge previously acquired (i.e., the taxonomies) and exploiting the intrinsic characteristics of the underlying definitions (i.e., the semantic domain of the taxonomies being analysed).

3.6.2 Mapping the semantic knowledge onto the LKB

Once the semantic acquisition process is finished, the taxonomic and other semantic information implicitly underlying the dictionary definitions must be translated into a formal representation language allowing consistent management of the lexical data acquired. This mapping process between the information extracted and the LKB is described in this section.

3.6.2.1 The Conversion Rule System

The main aim of the Conversion Rules System (CRS) [Ageno et al. 92c], [Ageno et al. 92d] in the SEISD environment is to perform the conversion of the semantic information extracted from the partially analysed dictionary senses to lexical entries constrained by the Type System of the LKB. That is, taking the analysed and validated taxonomy generated using the TaxBuild System, the CRS was designed in order to perform the translation from one structure to the other in the most declarative way. The lexicon produced by the CRS can then be loaded into the LKB system [Copestake 92a].

The mapping process requires knowledge from several heterogeneous sources of information including the results of analysed dictionary definitions, disambiguated taxonomic relations already extracted, the Type System defined in the LKB and bilingual dictionaries. However, rather than using a closed system with a fixed methodology, the CRS was developed using the PRE, allowing a variety of approaches and improvements.

3.6.2.2 Using the CRS to map lexical knowledge

Using the PRE, the CRS does not impose any fixed methodological strategy, although some metaknowledge must be provided to PRE. But whatever the methodology used, several decisions must be taken: the kind of control needed, the rulesets to be designed, the rules belonging to each ruleset, the relative priority assigned to each rule, and so on.

An initial set of two different modules for nouns and verbs was designed. For noun taxonomies a PRE module with three rulesets was implemented, while for verbs another PRE module was created in which only two rulesets were necessary.

Four different modes of interaction are provided by the system, involving increasing human intervention. Three of them are interactive while the other is performed in batch mode without any kind of interaction. Using the interactive modes, the conversion process derives lexicons (semi)automatically, asking the user for confirmation, modification or rejection of the information provided by the CRS (see Section 5.4).

Several strategies have been implemented for performing the conversion process, depending on the information available. Consider, for instance, the previous dictionary definition analysed by SinPar and placed as a node in the taxonomy of *bebida* (drink):

Word sense:	ojén_1_1				
Definition:	aguardiente dulce, anisado (*sweet, anise flavoured liquor*)				
SinPar:	[SN:	[n:	aguardiente,		
		PROPERTIES:		[a:	dulce,
				a:	anisado]]].

Once *bebida* has been assigned to the **c_art_subst** type (comestible-artifact-substance supertypes) in the Type System of the LKB, all the features of that type (local or inherited) are available. In addition, one of the translations of the adjective *dulce* using the bilingual dictionary is *sweet*, a subtype of taste, the constrained value of the feature *taste* of the *c_art_subst*, producing for this node in the taxonomy the following lexical entry:

ojén x_1_1
< lex-noun-sign rqs > < aguardiente_X_I_1< lex-noun-sign rqs >
< lex-sign sense-id : sense-id dictionary > = ("VOX")
< lex-sign sense-id : sense-id word > = ("ojén")
< lex-sign sense-id : sense-id homonym-no > = ("1")
< lex-sign sense-id : sense-id sense-no > = ("1")
< rqs : qual : taste > = sweet
< rqs : qual : smell > = flavoured.

But this is not the only information available for *ojén_1_1* once this lexical entry is placed in the LKB and every lexical entry is expanded using the LKB inheritance mechanisms. For example, following the hypernym chains and using the information extracted for the hypernyms of *ojén_1_1* made explicit during the taxonomic and semantic acquisition processes, the alcoholic property of *aguardiente* is inherited by *ojén_1_1* .

3.6.3 Multilingual lexical knowledge acquisition

An important issue for our system is multilinguality. In this section we present the way we can acquire automatically multilingual links from our lexical resources.

3.6.3.1 Tlinks

The initial assumption is that the basic units for defining lexical translation equivalence should be the lexical entries in the monolingual LKBs, which should, in general, correspond to word senses in the dictionary. Although in the simplest cases we can consider the lexical entries themselves as translation equivalent, in general, more complex cases occur corresponding to lexical gaps, differences in morphologic or lexical features, specificity, etc. (e.g., [Fernández 95], [Hirst 95] or [Soler 96]).

The tlink (Translation Link) mechanism [Copestake et al. 92] is general enough to allow the monolingual information to be augmented with translation specific information, in a variety of ways. We first describe the tlink mechanism in the LKB and then outline how some of these more complex equivalences can be represented.

We can define tlinks in terms of relations between Feature Structures (FSs). Lexical (or phrasal) transformations in both source and target languages[1] are a desirable capability, so we can state that a tlink is essentially a relation between two rules (of the sort already defined in the LKB) where the rule inputs have been instantiated by the representations of the word senses to be linked.

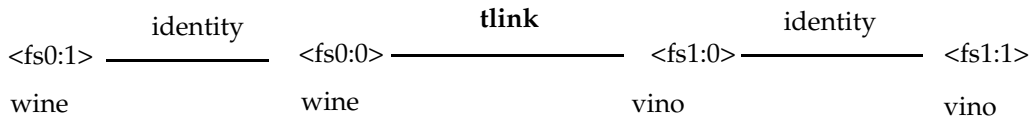


Figure 3.3, a tlink between "wine" and "vino".

As shown in Fig. 3.3, *wine* can be encoded as translation equivalent to *vino* by the identity rule.

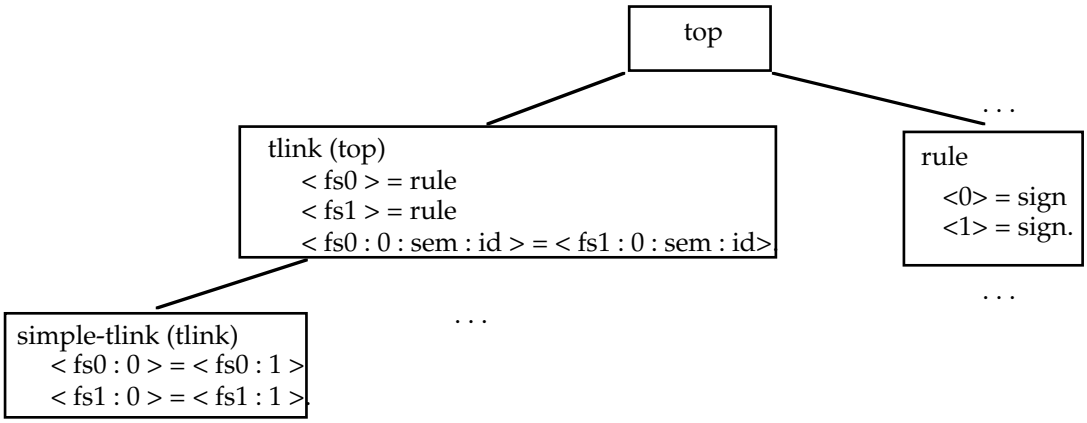


Figure 3.4, partial view of tlink type hierarchy.

Like other LKB objects, a tlink can be represented as a Feature Structure, as shown in Fig. 3.4. The Type System mechanism, in the LKB, allows further refinement and differentiation of tlink classes in several ways. A simple tlink is applicable when two lexical entries are straightforwardly translation equivalent, without any transformation. Thus, assuming that the LDOCE sense *absinth_l_0_1*, is translation equivalent to the DGILE *absenta_x_1_1*, we would have the following tlink:

simple_tlink
< fs0 : 1 > == absinth_l_0_1
< fs1 : 1 > == absenta_x_1_1.

The "syntactically sugared" version, which appears in tlink files, is:

absinth_l_0_1 / absenta_x_1_1 :
simple-tlink.

A partial tlink is applicable when we want to transfer the qualia structure from one sense to another, a phrasal tlink is necessary when we need to describe a single translation equivalence with a phrase, etc.

1 *In fact Tlinks are undirected relationships.*

3.6.3.2 TGE: Tlinks Generation Environment

Of course, tlinks can be established manually, but the multiplicity of cases occurring and the existence of several heterogeneous knowledge sources, such as bilingual dictionaries, monolingual LDBs and multilingual LKBs allows and motivates the mechanization of the process. To help perform this task we have developed an interactive[1] environment: the TGE [Ageno et al. 94], which is a module of SEISD.

The TGE has been implemented using the PRE. This approach has already been used in the CRS and was motivated mainly by the need to provide a flexible and open way of defining tlink formation mechanisms.

3.6.3.3 Using the TGE to generate tlinks

Like the CRS, the TGE may be considered a toolbox and, thus, it does not impose a single methodological strategy. Whatever the methodology followed, several decisions must be taken: different strategies and control mechanisms for tlink formation, several degrees of interaction with the user, different knowledge resources used, etc.

An initial set of modules has been designed according to the typology of tlinks partially depicted in Figure 5. It included four sorts of tlinks that showed different conceptual correspondences between the two languages. A more in-depth study of English/Spanish mismatches (i.e., [Soler 93], [Fernández 95]) could lead to an enrichment of the typology, and consequently, to the need to extend the extant modules.

To date, seven modules, each of them implemented as a ruleset, have been developed. Each of them generates one of the four kinds of tlinks. Each module follows a different strategy to guess a possible tlink, looking at the three accessible knowledge sources. Consider, for instance, the simple tlink ruleset:

• **Simple Tlink Module.** This is the case when there is a direct translation of the source entry in the bilingual dictionary. Consider the following example:

```
absenta_x_1_1 -----> absenta          LKB source entry
absenta -----> absinth                bilingual dictionary
absinth -----> absinth_l_0_1          LKB target entry
===>
absenta_x_1_1 / absinth_l_0_1 :
simple-tlink.
```

Absenta is translated in the bilingual dictionary by *absinth*, absinth_l_0_1 is a valid lexical entry of the target lexicon, and therefore a **simple-tlink** connecting the two entries is created.

The final process can be done in semi-automatic [Ageno et al. 94] or a fully automatic way [Rigau et al. 95] (see Chapter 6).

3.6.4 Semantic knowledge validation and exploitation

Once the information contained in the dictionary definitions has been represented as a lexicon in the MLKB, some testing processes should be performed on the lexicon acquired in order to improve the information extracted (e.g., detect possible errors or inconsistencies, extract more information, etc.), and then, to determine which changes to make in the next acquisition loop. The LKB guarantees the appropriateness of the lexicon against the Type System and provides some generative inference mechanisms (e.g., the inherence mechanism distributes the information from the top level lexical units to the most specific ones, lexical

[1] *TGE can also run without human interaction.*

rules produce new lexical entries from the preexisting ones, etc.) but no facilities are provided for performing complex consultations on the content of the lexical entries represented in the lexicon.

Of the two representational formalisms used for representing lexical entries, neither the LDB nor the LKB are able to aid the lexicographer in this validation process. The LDB provides, basically, database-like access to lexical information, while the LKB software manages a lexical knowledge representation based on typed Feature Structures (FSs) and defines valid operations on entries. For the purposes of both validation and exploitation of the information acquired, it would be useful to have a new system which had the function of both systems: LDB-like access to an LKB lexicon.

We developed the LDB/LKB merging system [Rigau et al. 94] taking into account a central guideline: LKB lexicons [Briscoe et al. 90], [Copestake, 92a] can be expressed, loaded and stored as in any other dictionary, in such a way that the LDB software [Carroll 90a] can be used without modifications or restrictions. The original LKB entries can be reconstructed from their LDB representation. This allows us to replace the LKB's lexical reading and access mechanism with the LDB functions, which gets round the current problem that reading in LKB lexicons is very slow, showing a considerable drop in performance when faced with real-size lexicons, and in the long term will allow for efficient access to indexed.

The central idea of loading lexicon files like other dictionaries source files into the LDB environment seems quite straightforward, but several problems (e.g., how to describe sources, what information index, how to access indexed information, how to query subsumed information, etc.) arise when it is approached in detail (see Section 4.6).

3.7 Conclusions

This chapter has been devoted firstly to the general methodology for creating an MLKB from monolingual and bilingual MRDs. That is, the main issues to be taken into consideration when designing the base methodology: the characteristics of the lexical resources used, the information to be extracted from them, how to carry out the process and how to represent and exploit the information extracted.

This chapter has also been devoted to a basic overview of **SEISD**, the software system that supports the methodology previously described and the functions that cover the main components of SEISD. Thus, the main aim of this Chapter was to provide a clear vision of the tasks that SEISD environment has to perform. That is, the extraction of semantic information implicitly located in DGILE (performed by **TaxBuild** and **SemBuild**), the mapping process of the information extracted to the LKB (covered by the **CRS**), the multilingual acquisition process (developed by the **TGE**) and the validation and exploitation of the lexical knowledge acquired (carried out by the **LDB/LKB System**). The task each module is in charge was described in detail. The problems related to each task are faced in Chapter 4 and the solutions we provide are described in Section 4.6 (validation and explotation of the knowledge acquired) Chapter 5 (lexical acquisition from MRDs from the monolingual point of view) and Chapter 6 (for the multilingual one).

Chapter 4

Main Issues of the Acquisition Process

4.1. Introduction

This chapter is devoted to the main problems related to the acquisition of lexical knowledge using SEISD (briefly described in the previous chapter) in order to acquire lexical knowledge from DGILE[1]. Although in Chapter 2 we provided a detailed state of the art on lexical acquisition, further and deeper studies must be carried out of the different tasks SEISD is in charge. This chapter focus on these studies. Section 2 explains different methodological approaches for classifying those concepts represented in a conventional MRD. Section 3 is devoted to several approaches for the Genus Sense Disambiguation (GSD) problem. Section 4 deals with the extraction of the main semantic relations from the dictionary definitions and their mapping onto the LKB. Section 5 deals with the multilingual enrichment of the LKB, and finally Section 6 is devoted to the main mechanisms used for the validation and exploitation of the multilingual LKB.

4.2. Definition of the main semantic subsets

This section deals with the methodological considerations for, firstly, the detection (and/or selection) of the main semantic subsets underlying MRD definitions and, secondly, for applying a descriptive approach to select the most representative dictionary senses leading to a full coverage of a semantic subset.

4.2.1 Predefined semantic primitives

DGILE (like most MRDs[2]) was not built according to a predefined classification of the concepts present in the universe, and lacks extensive semantic coding. Therefore, sets of

---

[1] *Although we used DGILE, most of our work can be applied to any other conventional MRD.*

[2] *Only a few dictionaries have been built following a fully prescriptive approach. LDOCE is partially annotated (86% of dictionary senses) with 16 basic and 17 composite semantic codes (or subject codes for verbs) and 124 major pragmatic codes (44% of dictionary senses), both of them organised as hierarchies ([Slator 88] imposed deeper structure onto the LDOCE pragmatic code hierarchy). The latest version of LDOCE -- LDOCE3-NLP -- is fully annotated. To date, only CIDE has been built using a complete classification system for words in terms of their meanings. Words in CIDE are grouped in a hierarchical system (creating semantic sets) according to shared and inherited semantic features. Furthermore, there are also several thesaurus such as the Roget's International Thesaurus, which classifies all words into 1,042 semantic categories, and for Spanish, DILEC (Casares), with 38 semantic classes organised hierarchically, and DILEV, with a three-level hierarchy of ontological concepts.*

related dictionary senses cannot be semantically classified in advance. However, following a purely bottom-up strategy, a natural set of taxonomic chains (that is, a natural classification of the concepts) represented in DGILE could be obtained. Thus, a straightforward way to obtain a LKB derived from the implicit taxonomy structure of the dictionary definitions can be done following a purely bottom up strategy with the following steps: 1) parsing each definition for obtaining the genus, 2) performing a genus disambiguation procedure, and 3) building a natural classification of the concepts as a concept taxonomy with several tops. Following this purely descriptive methodology, the semantic primitives of the LKB could be obtained by collecting those dictionary senses appearing at the top of the complete taxonomies derived from the dictionary. By characterizing each of these tops, the complete LKB could be produced. Using this approach the complete noun taxonomy was derived for DGILE (see [Rigau et al. 97][1] or section 5.3).

Roughly three different levels can be considered in the whole taxonomy: the top level, where the most general and ambiguous concepts would be described; the middle level, where specific concepts with clear taxonomic links would be placed; and the bottom level, where the most highly specific concepts of the dictionary would be described. However, several problems arise 1) due to the limitation of the genus sense disambiguation techniques applied (i.e, [Bruce et al. 92] report 80% accuracy using automatic techniques, while [Rigau et al. 97] report 83%) and 2) the source (i.e, circularity, errors, inconsistencies, omitted genus, etc.):

a) **Circularity**. Cycles frequently appear when linking dictionary senses by means of hypo/hypernym relations extracted via genus terms of the dictionary definitions, especially at high levels of the taxonomy. These cycles or tangled hierarchies, which were first detected and studied by [Amsler 81], can be seen as representing truly semantic primitives clustering the concepts present. That is, dictionary definitions necessarily end in circularity because lexicographers run out of words for defining other words. Those cycles linked to more general words defining closely related concepts make such circles arbitrary. A possible cause of these arbitrary circularities is the existence of near-synonyms [Vossen 94]. Consider, for example, the following definitions.

conjunto_1_4	**agregado** de varias cosas. (*set: aggregate of several things*).
conjunto_1_5	**totalidad** de una cosa, considerada sin atender a sus partes o detalles. (*set: totality of a thing, considered without attending to its parts or details*).
conjunto_1_6	**grupo** de personas que actúan bailando y cantando en espectáculos de variedades. (*set: group of persons who perform by dancing and singing in a variety show*).
agregado_1_1	**conjunto** de cosas homogéneas que forman un cuerpo. (*aggregate: set of homogeneous things that constitute a body).*
totalidad_1_3	**conjunto** (*totality: set*).
grupo_1_5	**conjunto** de elementos que se relacionan entre sí conforme a determinadas características. (*group: set of elements related to each other by means of particular characteristics*).

In this case, three fine-grained senses of *conjunto* (*set*) are defined by means of circular definitions.

An obvious way to avoid this problem is to collapse all the senses belonging to a cycle into a single node. Thus, in the course of constructing NounSense [Bruce et al. 92], all cycles inherent in the dictionary definitions were identified, analysed, removed from the network and substituted by semantic primitives.

---

[1] *This taxonomy contains 111,624 dictionary senses and has only 832 dictionary senses which are tops of the taxonomy (these top dictionary senses have no hypernyms), and 89,458 leaves (which have no hyponyms). That is, 21,334 definitions are placed between the top nodes and the leaves.*

b) **Errors and inconsistencies**. The existence of near-synonyms described above and the lack of any systematic control over dictionary-writing results in the loss of real meaning going across dictionary definitions. Consider, for instance, the following "correct" hypernym chain:

olla_1_1	**vasija** para cocer manjares, calentar agua, etc. (*pot: vessel for cooking food, heating water, etc.)*.
vasija_1_1	**receptáculo** para contener líquidos o cosas destinadas a la alimentación (*vessel: receptacle for containing liquids or things for nourishment)*.
receptáculo_1_1	**cavidad** en que se contiene o puede contenerse cualquier substancia. (*receptacle: cavity for containing any substance)*.
cavidad_1_1	**espacio** hueco de un cuerpo cualquiera. (*cavity: hollow space in any body)*.
espacio_1_1	**medio** homogéneo, isótropo, continuo e ilimitado en que situamos todos los cuerpos y todos los movimientos. (*space: homogeneous, isotropic, continuous and unlimited medium in which all the bodies and movements are located)*.

In this case, obviously, *olla* (*pot*) is neither a *cavidad* (*cavity*) nor *espacio* (*space*) but an instrument. That is, the genus term of *receptáculo_1_1* cannot be *cavidad*. This kind of problem usually appears in the higher parts of the hierarchies, where concepts become more general, abstract and ambiguous.

There are other inconsistencies between headword and genus term. For instance, it is possible that the genus term (or one of its possible senses) doesn't occur as headword in the MRD itself. This is the case of *tontería* (*silliness*) or *afiliado* (*affiliate*) which appear, respectively, 11 and 5 times as a genus term and are not defined within the dictionary itself.

c) **Definitions with omitted genus**. There are 2,362 noun definitions in DGILE (2%) in which no explicit genus term appears. Consider, for instance, the following examples:

comida_1_1	lo que se come. (*food: that which is eaten*).
denunciante_1_1	que hace una denuncia. (*denouncer: one who denounces*).

In the first case, the genus term appear as a pronoun, while in the second case, the genus term, person, is omitted (the relative clause *que* can be used for persons and things).

Furthermore, the top dictionary senses, following this strategy, do not usually represent the semantic subsets that the LKB needs to characterize in order to represent useful knowledge for NLP systems. In other words, there is a mismatch between the knowledge directly derived from an MRD and the knowledge needed by a LKB.

To illustrate the problem we are facing, let us suppose we plan to place the FOOD concepts in the LKB. Neither collecting the taxonomies derived from a top dictionary sense (or selecting a subset of the top dictionary senses of DGILE) closest to FOOD concepts (e.g., *substancia* -substance-), nor collecting those subtaxonomies starting from closely related senses (e.g., *bebida* -drinkable liquids- and *alimento* -food-) we are able to collect exactly the FOOD concepts present in the MRD. The first are too general (they would cover non-FOOD concepts) and the second are too specific (they would not cover all FOOD dictionary senses because FOODs are described in many ways).

In order to solve all these problems we propose to use a mixed methodology. That is, by attaching selected top concepts (and its derived taxonomies) to prescribed semantic primitives represented in the LKB. Thus, first, we prescribe a minimal ontology (represented by the semantic primitives of the LKB) capable of representing the whole lexicon derived from the MRD, and second, following a descriptive approach, we collect, for every semantic primitive placed in the LKB, its subtaxonomies. Finally, those subtaxonomies selected for a semantic primitive are attached to the corresponding LKB semantic category. Obviously, apart of the problems for defining the minimal ontology (see bellow), the difficult part of

this approach is to select from the MRD those representative concepts (and its taxonomies) for a given semantic primitive. We perform, in Section 5.2, a new approach for selecting those concepts (or top beginners) for a given LKB semantic primitive.

Furthermore, by defining mid-level semantic primitives that are neither too general nor specific, the construction of specific grammars for every semantic subset becomes easier (see Section 4.4). More general concepts (those that appear on the top level) include a great amount of hyponym concepts preventing the determination of an acceptable set of common properties while specific concepts (those that appear on the bottom level) do not define complete semantic taxonomies and are not useful for determining sharable properties [Castellón 93].

Several prescribed sets of semantic primitives have been created as ontological knowledge bases (e.g., Penman Upper Model [Bateman 90], CYC [Lenat & Guha 90], EDR [Uchida 90] or WordNet [Miller 90]). For instance, WordNet noun top unique beginners are 24 semantic categories (e.g., ACT, ANIMAL, ARTIFACT, ATTRIBUTE, etc.)[1]. However, identifying and characterizing an appropriate set of primitives from an MRD and determining the correct placement of those primitives in the hierarchy is a difficult task. Following a pure descriptive approach, [Bruce et al. 92] use as semantic primitives the LDOCE semantic codes (which forms a rather vague type hierarchy). These semantic primitives form the top level of the NounSense acyclic network.

As we said before, following a mixed approach, that is, prescribing a minimal set of semantic primitives and attaching to them the corresponding taxonomies derived directly from MRDs, the main dictionary senses representative of a semantic primitive must be selected. Thus, in our case, once the top level of the taxonomic structure has been prescribed (the Type System of the LKB) we must assign to each type (corresponding to mid-level concepts) all those main top dictionary senses which represent those types in the MRD. Then, those subtaxonomies derived from these main top beginners can be attached to the corresponding types of the LKB. Until now, the selection of those representative concepts of a semantic primitive has been performed by introspection.

For [Copestake 90], dictionary senses which are suitable starting points for building taxonomies can be identified because they occur with high frequency as class terms. This is not always true, as there is no direct mapping between the genus term frequency and the topology of the taxonomies underlying the MRD (that is, information regarding density, the number of different senses per level, etc.). Consider Table 4.1, with the 40 most frequent genus terms extracted from DGILE.

acción	4408	aparato	479	partidario	275	insecto	207
persona	4352	mujer	464	enfermedad	271	movimiento	204
efecto	3157	pez	451	arte	270	hierba	204
algo	2381	terreno	411	moneda	258	sitio	203
planta	1569	ave	395	fruto	257	mamífero	199
calidad	970	golpe	369	doctrina	245	máquina	196
instrumento	704	hombre	325	tela	241	figura	196
cosa	588	substancia	292	estado	224	sistema	195
árbol	555	animal	279	porción	219	vasija	191
lugar	517	arbusto	277	medida	217	unidad	191

Table 4.1, most frequent genus term in DGILE.

For instance, *acción* (*action, act*), *persona* (*person*) or *planta* (*plant*) etc. have flat, broad taxonomies while other less frequent genus terms such as *substancia* (*substance*) are more productive (taxonomies derived from it are deeper and have relatively more hyponym descendants).

[1] *Rather than a semantic domain classification this is a coarse-grained semantic classification useful only for the lexicographers during the creation process. (e.g., there are no verbs processes of cooking nor places to eat related in any way to the noun food classification) .*

Due to the fact that starting top points usually appear less frequently as a genus term than some possible hyponyms (e.g., *animal* appears after *pez* (*fish*) and *ave* (*bird*), and before *insecto* (*insect*) and *mamífero* (*mammal*)) the selection of the top concepts underlying the MRD must be done carefully.

Yet, obviously, genus terms with high frequency indicate an important concept with a great amount of instances, and possibly an important part of a taxonomy.

The selection of the main (in the sense that they facilitate the acquisition process) semantic primitives (and their representative dictionary senses) also depends on their main characteristics. Consider the following example:

cosa (*thing*) appears as a genus term 588 times.
objeto (*object*) appears as a genus term 86 times and *objeto* IS-A *cosa*.
instrumento (*instrument*) appears as a genus term 704 times and *instrumento* IS-A *objeto*.
aparato (*device*) appears as a genus term 479 times and *aparato* IS-A *instrumento* .

It is clear that an instrument has important characteristics (such as, for instance, PURPOSE information) that a general object does not have. Then, in this case, *instrumento* must be selected as a top beginner for the **made-by-humans-object** type of the LKB.

4.2.2 Semantic coverage

It is not clear how many different semantic primitives are necessary to build a complete LKB. [Yarowsky 92] uses the major 1,042 categories of *Roget's International Thesaurus* as approximations of conceptual classes. [Liddy & Paik 92] use the 124 major pragmatic fields (or subject areas) of LDOCE as semantic primitives, and [Bruce et al. 92] a revised version of LDOCE's 34 semantic categories. [Rigau 94] use as semantic primitives the 24 noun top unique beginners of the nominal part of WordNet. [Hearst & Schütze 95] describe a method for converting the hierarchical structure of WordNet into a flat system of 726 semantic categories. The number of semantic subsets necessary to represent the knowledge about the world in an LKB depends on the precision and degree of accuracy of the concepts we wish to represent. Moreover, [Richardson 97] do not use semantic primitives building a semanticaly labeled LKB of words.

The Type System represents a set of semantic primitives by means of compositional types using the LKB multiple inheritance mechanism. Thus, there are types such as **c_art_substance** that combine the **comestible**, **artifact** and **substance** primitive semantic types with their own particular features. The semantic class of human processed drinks and foods (extracted from the MRD via clear taxonomic links starting from those dictionary senses representing the concepts *drink* and *food*) can be classified under this prescribed semantic type. Once the main semantic primitives have been prescribed in order to represent the whole LKB Type System, the selection of the appropriate subtaxonomies representing each type must be performed. Likewise following a mixed methodology, for a given type [Castellón 93] selected by introspection[1] a small set of top dictionary senses as starting points for deriving the corresponding subtaxonomies.

Nevertheless, manually selecting a small set of top dictionary senses as the most general concept representative of the semantic class and following the hyponym chains of dictionary senses often fails to lead to full coverage of the semantic subset. Not all dictionary senses belonging to the same semantic class present in the MRD are discovered by following the hyponym chains from only one dictionary sense. This phenomena can be illustrated with the following dictionary definition:

queso_1_1 **masa** que se obtiene cuajando la leche, exprimiéndola para que deje suero y echándole sal para que se conserve. (*cheese: mass obtained by curdling milk,...*).

[1] *The nominal part of DGILE contains more than 14,000 different genus terms.*

This dictionary sense belongs clearly to the more general concept FOOD, but this dictionary sense does not appear following a top-down construction of the taxonomy FOOD from the dictionary sense *alimento* (food) [Castellón 93]. That is, rather than defining *queso* (*cheese*) by its utility (useful to be eaten by humans), it is described by its state of aggregation. Obviously, the concept underlying this dictionary sense should be attached to the **c_art_subst** type, like other foods. Furthermore, the taxonomy of *queso* (*cheese*) contains 30 different dictionary senses, and all of them have been omitted from the taxonomy for *food* extracted from DGILE (presented in [Castellón 93] containing 135 dictionary senses and extracted following the hyponym chains from the dictionary sense *alimento* (*food*)). This example clearly shows that selecting by introspection a limited set of dictionary senses as unique top beginners does not cover the totality of dictionary senses of an MRD belonging to a selected semantic primitive, because the top beginners can be described by several characteristics. Then, for each prescribed semantic class several top beginners could be attached. Section 5.2 describes a new descriptive approach for automatically detecting a large set of top dictionary senses for a given prescribed semantic primitive.

4.3. Genus disambiguation

This section is devoted to methodological considerations and techniques for genus sense resolution. After a short introduction, an overview of the main contributions to the word sense resolution and genus sense disambiguation are shown. At the end, we provide a summarization of lexical measures of relatedness.

4.3.1 Genus Term Selection vs. Genus Sense Disambiguation

A central problem that must be treated in depth consists of extracting taxonomies from the implicit knowledge that appears in dictionary definitions. This main problem can be divided into two different subproblems: first, the location of the genus term in the definitions, and second, as the genus term appears not as a sense but simply as a word, the next subproblem consist of selecting the correct sense (which usually appears in the same dictionary) for that genus term.

In order to select the correct semantic head or genus term for noun and verb definitions and discard those for which the top word is not the genus term, a specialized grammar has been developed. Frequently, the genus term for noun definitions is the first noun present in it and for verb definitions the first verb [Amsler 81]. Obviously, there are many cases where this simple rule for genus detection does not hold [Ageno et al. 91a], for instance, the linkers [Meijs 90] (also called key modifiers [Nakamura & Nagao 88] or disturbed heads [Bruce et al. 92]) between the headword, lemma or definiendum and the head of the dictionary sense. Consider for instance the next definition:

antebrazo_1_1 **parte** del **brazo** desde el codo hasta la muñeca. ([a] *forearm, part of the arm from the elbow to the wrist*).

In this case, **brazo** (arm) is the genus term of **antebrazo_1_1** definition and **parte** (part) is the linker. Table 4.2 shows the ten different kinds of relations between noun dictionary senses we gathered from DGILE.

set-of	(conjunto, grupo, serie, etc.)	1,912
mixture-of	(mezcla, etc.)	90
mass-of	(masa, etc.)	94
related-to	(relativo, etc.)	118
without-of	(falta, falto, ausencia, etc.)	298
with-of	(presencia, etc.)	36
part-of	(parte, pieza, pedazo, etc.)	2,175
member-of	(miembro, elemento, uno, etc.)	70
special-is-a	(tipo, especie, variedad, etc.)	1,145
is-a		86,755

Table 4.2, kinds of relations between noun headwords and genus.

Consider the following example, in which the genus term appears as the third noun in the definition (after *órbita* and *planeta*).

afelio_1_1 En la órbita de un planeta, el **punto** más alejado del Sol. (*aphelion: in the orbit of a planet, the most distant* ***point*** *from the Sun*).

For verb definitions, the selection of the genus term can be simplified by searching in the definitions for words ending with the suffixes **ar**, **er**, or **ir** (and in the pronominal cases adding also the suffix **se**). Consider the following example:

abr**ir**_1_9 Romp**er** o despeg**ar** (cartas o paquetes). (open: rip or unstick (envelopes or packets)).

Although specialized for obtaining the genus term, this grammar is domain independent and covers all noun and verb definitions with a precision that exceeds 98% (see Section 5.3).

Once the Genus Term Selection (GTS) has been performed, the second task, the Genus Sense Disambiguation (GSD), which can be considered as a special case of the more general Word Sense Disambiguation (WSD) problem, must be carried out. The following section deals firstly with the more general problem of WSD and secondly with the GSD problem.

4.3.2 Word Sense Disambiguation

Lexical Ambiguity Resolution or Word Sense Disambiguation (WSD)[1] is a long-standing problem in Computational Linguistics[2]. Much recent work in Lexical Ambiguity Resolution offers the prospect that a disambiguation system might be able to receive unrestricted text as input and tag each word with the most likely sense with fairly reasonable accuracy and efficiency. The most widely accepted approach is to attempt to use the context of the word to be disambiguated together with information about each of its word senses to solve this problem. Although most of the techniques for word sense resolution are presented as stand-alone, it is our belief, following the ideas of [McRoy 92], that fully fledged Lexical Ambiguity Resolution should combine several information sources and techniques.

---

[1] *Although this work is concerned to Genus Sense Disambiguation (GSD), it can be considered a particular case of Word Sense Disambiguation (WSD).*

[2] *One can find early references (e.g., [Kaplan 50] or [Yngve 55]) related to the lexical ambiguity, but unfortunately it still remains as an open problem looking at the corpora news list discussion started by Adam Kilgarriff (June 1995, and briefly described in [Ribas 95]) and from the recent conclusions at SIGLEX WorkShop: TagginText with Lexical Semantics: Why, What and How? (e.g., [Resnik & Yarowsky 97] "Word sense disambiguation (WSD) is perhaps the great open problem at the lexical level of natural language processing".)*

Several approaches have been proposed for attaching the correct sense (from a set of prescribed ones) of a word in context[1]. Some of them serve as models only for simple systems (e.g., connectionist methods [Cottrel & Small 89], Bayesian networks [Eizirik et al. 93]) while others can be fully tested in real-size texts (e.g., statistical methods [Yarowsky 92], [Miller et. al. 94], [Yarowsky 94], [Yarowsky 95], knowledge-based methods [Sussna 93], [Agirre & Rigau 95] and [Agirre & Rigau 96a], or mixed methods [Richardson et al. 94], [Resnik 95], [Jiang & Conrath 97]). WSD performance is reaching a high standard, although usually only small sets of words with clear sense distinctions are selected for disambiguation (e.g., [Yarowsky 95] using an unsupervised method reports a success rate of 96% when disambiguating twelve words with two clear sense distinctions each).

In order to evaluate the performance of a word-sense identification technique, [Gale et al. 93] suggest that the appropriate basis for comparison would be a system that assumes that each word is being used in its most frequently occurring sense. They review the literature on how well word-disambiguation programs perform; as a lower bound, they estimate that the most frequent senses of polysemous words would be correct 75% of the time and they propose that any sense-identification system that does not give the correct sense of polysemous words more than 75% of the time would not be worth serious consideration. Tagging both syntactically and semantically the open-class words of the Brown Corpus, [Miller et al. 94] propose benchmarks (guessing, most frequent and co-occurrence) for systems of automatic sense identification. In this case, using WordNet, the most frequent heuristic yields the correct sense for polysemous words 58% of the time. This difference between the two baselines is due to the different polysemous degree and different coverage of words tested[2].

Recent WSD approaches may be roughly classified by several features: (1), by the kind of knowledge they use to perform the disambiguation process; (2), by the training procedure they use (in a supervised method); (3), by the way they combine the information provided by different sources; (4), by the context (if any) they use.

Statistical model methods

Since [Brown et al. 91a], who used a supervised training model extracted from the Canadian Hansard bilingual corpus, many approaches to WSD using stochastic models have been proposed.

Some approaches ignore context, providing a simple and ready-to-use procedure for performing the sense disambiguation process. The results of such a method would be useful as a lower bound for approaches that use more refined knowledge [Gale et al. 92b]. The results seem to indicate that the most frequent heuristic works at 75% accuracy. Using SemCor (a part of the Brown corpus semantically tagged with WordNet synsets), [Miller et al. 94] report an average accuracy of 67% (including monosemous). Using WordNet 1.5 (the first version to contain the synsets explicitly ordered by frequency), [Peh & Ng 97] report a success rate of 63% for polysemous nouns (74% for all nouns).

Another family tries to select the sense that seems to be most appropriate for the overall context of the ambiguous word. These works make use of the strong discourse relation that exists between the senses of words and the general topic(s) of text. [Gale at al. 92a] defend the "one sense per discourse hypothesis: if a polysemous word appears two or more times in a well written discourse, it is extremely likely that it will share the same sense". They found that the tendency to share sense in the same discourse is extremely strong (98%). The context size has been measured using words and characters. [Gale et al. 92a], [Yarowsky 92] and [Gale et al. 93] use a 100-word window size of context surrounding the ambiguous word in order to

---

[1] *The most simple approaches are those that ignore context and select for each ambiguous word its most frequent sense (e.g., [Gale et al. 92a], [Miller et al. 94]).*

[2] *[Leacock et al. 95] report over 90% correct answers for a supervised experiment with two-sense distinctions for the word line. With the addition of a third sense, the classifiers yield a sharp degradation, obtaining a mean of 76% correct answers. Moreover, selecting a particular degree of polysemy some senses are harder to resolve than others.*

choose the sense with the highest weight provided by the context[1]. [Shütze 92b] report the best window size, using 1000 characters.

WSD techniques using supervised training models in a general context have been widely studied. [Brown et al. 91a], [Gale et al. 92a] and [Gale et al. 93] (following the idea that "two languages are better than one" [Dagan et al. 91]) train a language model on bilingual corpora taking bilingual differences in translation as different words senses. [Leacock et al. 95] compare the performance of three different classifiers: Bayesian, neural networks and content vectors. They demonstrate that each of the techniques is able to distinguish six senses of the word line with an accuracy greater than 70% using a training set of 200 examples. [Yarowsky 94] exploits both local syntactic patterns and more distant collocational evidence by identifiying the single best disambiguating evidence in the target context using desision lists. Unsupervised learning methods in a general context have been also tested. [Yarowsky 92] collects cooccurrence data from *Grolier's Encyclopaedia* for every semantic category represented in *Roget's Thesaurus* in order to obtain the salient words for each category. [Karov & Edelman 96] describe a circular converging process between word similarity and context similarity measures, combining three MRDs and the *Wall Street Journal* corpus for disambiguating 500 examples of four polysemous words with a success rate of 92%.

Some works such as [Schütze 92b] perform a disambiguation process with minimal human intervention. That is, rather than a completely unsupervised process (without no human intervention and without any hand-tagging work), [Shütze 92b] uses a post-hoc alignment of clusters to word senses. The clusters are generated automatically from cooccurrence data extracted from corpora using a content vector representation. Because the cluster partition does not necessarily correspond to word senses, he manually assigns each to a word sense (inspecting 10-20 sentences per cluster).

Other approaches have considered that local context (introducing such information as word order collocations) may be a good predictor of the appropriate sense. These approaches could be summarized as the "one sense per collocation hypothesis" [Yarowsky 93]. He demonstrates empirically that for certain definitions of collocation (adjacent words of a certain POS, in any direction) a polysemous word exhibits essentially one sense per collocation. The different approaches using local context differ in several ways: the features of the local context they take into account; the way information provided by the features is combined; the function used to calculate the weight that each feature provides for the different senses; and finally, the training method (if any) they use. The features of the local context used by the different approaches range widely: morphological forms, word adjacency (or within a small number of positions) in left/right direction, POS of near words, deeper syntactic relations, etc.

Applying supervised learning models on local context, while [Yarowsky 93] and [Yarowsky 94] consider the optimum feature in performing the disambiguation task without defining the relation between different features, [Bruce & Wiebe 94] decompose the probabilistic model that would result from taking all features as interdependent. [Ng & Lee 96] present a supervised WSD method (Lexas) which integrates several knowledge, including part of speech, morphological form, local colocations and verb-object syntactic relation. Rather than characterizing each sense of a word, a set of occurrences (examples) of the senses are provided. That is, the method compares the context of the word to be disambiguated (sintetized as a vector of several features) with every tagged context (examples sintetized also as a vector of features). The method selects the example (and the asociated sense tag) with minimal distance to the context. They report also an evaluation on a large data set (192,800 occurrences of 121 nouns -mean of 7.8 senses- and 70 verbs -mean of 12.0 senses-) using the Brown Corpus (54% accuracy on polysemous words) and the Wall Street Journal (69% accuracy). [Dagan et al. 97] compare four similarity-based estimation methods agains back-off and maximum-likelihood estimation methods.

[1] *Context is modelled as a bag of words ignoring important linguistic information such as word order and collocations.*

Using mainly the property described in [Yarowsky 93] and [Yarowsky 94], [Yarowsky 95] describes an iterative bootstrapping procedure over the features learned during the unsupervised training process.

Several supervised learning models performed the training process using a manually sense tagged corpus (e.g., [Bruce & Wiebe 94], [Leacock et al. 95], [Ng & Lee 96]). Some works train the data exploiting the bilingual knowledge placed in the parallel texts (e.g., [Brown et al. 91a], [Gale et al. 92a] and [Gale et al. 93]). Some other works propose the use of artificial sense ambiguities like pseudo-words (e.g., [Shütze 92b], [Gale et al. 92b], [Dagan et al. 97]) and other the application of the methodology to other similar tasks like the accent restoration (e.g., [Yarowsky 94]).

The results provided in most papers cited above seem to be very good: over 90% accuracy or even higher. However, the sense distinctions introduced (which are not true word sense distinctions, such as words translated differently across languages, different word accentuation, OCR ambiguities, homophones, artificial ambiguities such as pseudo-words [Dagan et al. 97]) do not seem to be fine-grained enough.

Knowledge-based methods

Since [Lesk 86] many researchers have used MRDs as a structured source of lexical knowledge for the WSD problem. These approaches mainly seek to avoid the need for a large amount of training materials required in supervised methods. WSD techniques using preexisting structured lexical knowledge resources differ in (1) the lexical resource used (monolingual and/or bilingual MRDs, thesauri, Lexical Knowledge Base, etc.); (2), the information contained in these resources that the method exploits; (3) the property used to relate words and senses.

[Lesk 86] proposes a method for guessing the correct word sense in context by counting word overlaps between dictionary definitions of the various senses. [Veronis & Ide 90] propose a similar method but uses a spreading activation network (see [Hirst 88] and [Hayes 77]), constructed from *Collins Dictionary of the English Language*. [Sutcliffe & Slater 94] compare the two methods mentioned above using the Merrian-Webster. For a POS tagged test set of 100 sentences in *Animal Farm* by George Orwell, they found for the Lesk method 31% of ambiguous senses to be correctly tagged (40% overall) and for the Ide and Veronis method 68% (and 72% overall). [Cowie et al. 92] uses the simulated annealing technique for overcoming the combinatorial explosion of the Lesk method using LDOCE. For a non-POS tagged test set of 50 sentences they report a success rate of 47% at a sense level, and 72% at the homograph level (both results are for all types of words). Again using the simulated annealing technique and LDOCE but on a POS tagged test set of 209 words from the Wall Street Corpus, [Wilks & Stevenson 97] found, for word tokens which had more than one homograph, 57% correctly sense assigned and 86% assigned correctly to the homograph.

[Guthrie et al. 91] propose the use of the subject semantic codes of LDOCE to partition the dictionary and collect neighbours (like salient words [Yarowsky 92]) for WSD in a Lesk style. [Wilks et al. 93] use the cooccurrence data extracted from LDOCE to construct word-context vectors and thus, word-sense vectors. They perform a large set of experiments testing relatedness functions between words and vector similarity functions, disambiguating 197 non-POS disambiguated ocurrences of the word bank (13 senses) reporting a success rate of 45% at sense level, and 85% at the homograph level. [Harley & Glennon 97], using an ad-hoc weighting mechanism for the different sources of lexical knowledge present in the completely coded *Cambridge International Dictionary of English* (CIDE), report disambiguating 4,000 hand-tagged words, an overall accuracy (for all types of words) of 73% at a sense level (an average of 19 senses per word) and 78% accuracy at a homograph level (an average of 5 senses per word).

Other approaches measure the relatedness between words, taking as a reference a structured semantic net. Thus, [Sussna 93] employs a complex notion of conceptual distance between network nodes in order to improve precision during document indexing, at a sense level reporting 55.8% precision for polysemous nouns (71% overall). [Rigau et al. 97] uses the conceptual distance measure described in [Agirre et al. 94] as one of the methods for the GSD

problem in DGILE and LPPL. This technique reaches 49% precision in DGILE for polysemous nouns (57% overall). [Rigau et al. 98] uses also this technique during the first labeling process of DGILE. Conceptual density, a more complex semantic measure between words, is defined in [Agirre & Rigau 95] and used in [Agirre & Rigau 96a] as a proposal for WSD using the Brown corpus. This approach achieves a precision of 43% at a sense level for polysemous nouns (64.5% overall) and 53.9% at a file level for polysemous nouns (71.2% overall at a file level)[1].

Mixing statistical and knowledge-based methods

Lately, some researchers have carried out different experiments combining different sources of lexical knowledge (structured and non-structured) for WSD, and thus different techniques for exploiting the knowledge contained. These techniques differ in: (1) the kind of lexical resources used in the different steps of the method proposed (corpora, monolingual MRDs, bilingual MRDs, thesauri, LKBs); (2) the particular characteristics of these resources that are exploited during the disambiguation task; and (3) the measures used to compare similarities between lexical units.

[Yarowsky 92] uses *Roget's Thesaurus* to partition *Grolier's Encyclopaedia* and collect the salient words for each category. In this case, rather than a sense level the WSD task is carried out at a more coarse-grained Roget's category level (words are divided into 1,042 semantic categories). He report an average of 92% correctly disambiguating 12 polysemous words. In a similar approach, [Liddy & Paik 92] use LDOCE subject semantic codes and the *Wall Street Journal* corpus to compute a subject-code correlation matrix. For 166 POS tagged sentences they report a success rate of 89% assigning the correct subject code (words are divided into 122 semantic categories). [Karov & Edelman 96] describe a system which learns from a corpus a set of typical usages for each of the senses of the polysemous word listed in an MRD. They propose a pair of feedback iterative similarity formulae between words and sentences. [Yarowsky 95] proposes the use of MRDs to collect seed words in the first step of his cycling procedure which collect local features for WSD using corpora.

[Ribas 95] applies class-based selectional local restrictions, taken from WordNet and trained on unsupervised corpora, as introduced by the verb to disambiguate nouns that occur as heads of complements of the verb. [Resnik 93], [Richardson et al. 94], [Resnik 95], [Jiang & Conrath 97] present a method for automatic sense disambiguation based on the information content measure gathered from corpora also using WordNet. The similarity between two classes is approximated by the information content of the first class in the noun hierarchy that subsumes both classes. The information content of a class is approximated by estimating the probability of occurrence of the class in a large corpus.

Combining several heterogeneous techniques and independent lexical resources [Rigau et al. 97] report an accuracy of 83% at a sense level (an average of 5.75 senses per noun) and 79% for polysemous nouns (an average of 6.65 senses per noun). Our unsupervised approach uses implicit information contained in MRDs to construct content vector representations and test different techniques and similarity measures for assigning the correct hypernym genus sense. We also use a bilingual MRD to assign semantic categories from WordNet to word senses and perform, in a similar way to [Yarowsky 92], an unsupervised training process to collect salient words for each semantic category (see [Rigau et al. 98]). This work also uses the notion of conceptual distance between network nodes taking as a reference WordNet1.5.

Due to the fact that supervised learning methods for WSD require a large amount of sense tagged corpora we believe that unsupervised (or minimally supervised) methods using structured lexical resources currently offer the best approach for dealing with a real WSD task. For instance, [Ng 97] estimate 16 man-years to construct a sense-tagged corpus for using Lexas, a supervised training model method based on examples (the accuracy of the method ranges from 58.7% on Brown Corpus to 75.2% Wall Street Journal).

The results obtained by the various approaches described above seem to be extremely promising. However, when comparing results reported on these papers, it seems difficult to

[1] *The file level degree of polysemy is greater than the homograph level.*

extract any definitive conclusion on the real performance of each approach. There are many uncontrolled variables (measures of evaluation, training material, granularity of sense distinctions, range of POS categories of the words the method disambiguates, languages used, etc.) and no common test set useful for comparing the results of all methods (SemCor is not useful enough for supervised approaches or for languages other than English). Despite the results described above, semantic tagging of raw texts still remains an open problem [Resnik & Yarowsky 97][1].

4.3.3 Genus Sense Disambiguation

Lexical ambiguity pervades language in texts, including dictionary definitions themselves. The words used in dictionary definitions of words, and their senses, are themselves lexically ambiguous.

A taxonomy of words [Nakamura & Nagao 88] can be done without any costly disambiguation process simply by linking the headword of a dictionary sense to its genus term by an ISA relation (or hypo/hypernym link). Taxonomies constructed in this way could also be useful for NLP systems [Dolan et al. 93]. While a system whose tasks include word sense tagging must be able to take an input text, determine the concept that each word or phase denotes, and identify the role relations that link these concepts, for genus sense resolution it is only necessary to attach the genus term to its correct sense, and there are only a few possible relations between the headword and the genus term.

Although the GSD problem may seen easier than the WSD one, the fact that the most frequently used words in general domain texts (such as dictionaries) are the most ambiguous [Zipf 45] does not facilitate genus sense resolution[2]. While the average of senses per noun headword in DGILE is 1.73, the average of senses per noun genus term is 2.75. To illustrate this, Table 4.3 shows the degree of polysemy of the most frequent noun genus terms in DGILE. From left to right: word, number of noun senses and frequency as genus term.

acción	12	4408	aparato	11	479	partidario	6	275
persona	8	4352	mujer	3	464	enfermedad	3	271
efecto	8	3157	pez	3	451	arte	6	270
algo	1	2381	terreno	5	411	moneda	3	258
planta	11	1569	ave	3	395	fruto	6	257
calidad	5	970	golpe	21	369	doctrina	7	245
instrumento	4	704	hombre	14	325	tela	15	241
cosa	3	588	substancia	10	292	estado	18	224
árbol	10	555	animal	3	279	porción	5	219
lugar	8	517	arbusto	1	277	medida	8	217

Table 4.3, polysemous degree of the most frequent noun genus in DGILE.

Furthermore, dictionary senses for a word frequently differ in subtle distinctions producing a large set of very closely related dictionary senses [Jacobs 91]. There are numerous reasons why a dictionary might split an entry into multiple senses, only some of which have to do with meaning [Gale et al. 93]. Dictionaries senses with the same meaning may split an entry when there are differences in:

---

[1] *Senseval, a WSD evaluation framework under the auspices of ACL SIGLEX has just started.*

[2] *However, the GSD problem can be restricted, as the headword and the genus term have to be the same part of speech. That is, when constructing the FOOD taxonomy only nouns must be considered.*

- part of speech (nouns vs. adjectives, etc.).
- syntactic features (person, number, gender, etc.).
- valency structures (transitive vs. intransitive verbs, etc.).

Unfortunately, lexicographers do not always agree on how to split a dictionary entry into senses. For example, [Atkins and Levin 88] show the difficulties in manually merging two short entries (whistle and whistler) from two monolingual English collegiate-style dictionaries.

Moreover, some sense distinctions are larger than others. Meaning is probably best thought of as a continuous quantity, with an infinite number of "shades" between any two points. That is, frequently there is no a single solution (only one dictionary sense) for the genus term of a hyponym dictionary sense. Obviously, this is not the case of the example suggested by [Lesk 86] about the ambiguous use of the word *insular* in Melville's *Moby Dick*, where in "*your insular city of Manhattan*"it means both "surrounded by water" and "narrow-minded". In addition, the hypothesis of "one sense per discourse" suggested by [Gale et al. 92a] cannot be useful for the GD problem because the dictionary senses are not long pieces of text[1].

Although a large number of dictionaries have been exploited as lexical resources, the most widely used monolingual MRD for NLP is LDOCE which was designed for learners of English. It is clear that different dictionaries do not contain the same explicit information. This information placed in LDOCE allows an easier extraction of other implicit information (e.g., taxonomies) [Bruce & Guthrie 92]. Does this mean that only highly structured dictionaries like LDOCE are suitable for exploitation as lexical resources for NLP systems? In this thesis we used a completely different dictionary. The *Diccionario General Ilustrado de la Lengua Española* (DGILE) is substantially poorer in coded information than LDOCE[2]. These dictionaries are very different in number of headwords, degree of polysemy, size and length of definitions (cf. Table 4.4[3]). While DGILE is a good example of a large dictionary (it aims to cover the whole of the Spanish vocabulary for Spanish readers), LDOCE was designed for learners of English (it aims to cover a part of English for non-English readers). This thesis attempts to demonstrate that by combining appropriate methodologies, we can construct complete taxonomies from any conventional dictionary in any language.

	DGILE		LDOCE	
	total	nouns	total	nouns
Headwords	93,484	53,799	35,956	23,251
Senses	168,779	93,275	76,059	42,129
Total number of words	1,227,380	903,163	731,466	480,999
Average length of definition	7.26	9.68	8.62	11.42
Senses per headword	1.8	1.73	2.12	1.82
Senses per genus		2.75		3.72
Senses per genus (polysemous only)		3.74		5.08
Real polysemy		5.96		8.47
Real polysemy (polysemous only)		6.63		9.23

Table 4.4, comparison of LDOCE and DGILE.

Due to the lack of explicit semantic information in DGILE the only source of information for attaching a dictionary sense to the correct hypernym dictionary sense (and thus,

---

1 *More than 11% of noun dictionary senses have no differentia. These definitions have only one synonym word of the headword. Furthermore, these definitions have no context for the disambiguation process.*

2 *In LDOCE, dictionary senses are explicitly ordered by frequency, definitions were written using a controlled vocabulary of 2,000 words, 86% of dictionary senses have semantic codes and 44% of dictionary senses have pragmatic codes.*

3 *Real polysemy considers senses per genus, but taking the real frequency of each genus.*

constructing disambiguated taxonomies) is the definitions themselves. Thus, with DGILE the GSD problem can be seen to be closer[1] to the more general WSD problem than with LDOCE. This is because no direct use of the semantic codes can be made between a raw text and the LDOCE coded dictionary senses[2].

[Copestake 90] describes an approach to extract word sense taxonomies from LDOCE in a (semi)automatic top-down fashion. She uses four different heuristics (subject codes, style codes, sense ordering, word overlapping between definitions) which exploit intrinsic characteristics of LDOCE. The program was tested by building two substantial taxonomies (1,700 noun senses, 7% of the total) in two hours of interaction to derive them both.

Selecting the correct sense genus term for LDOCE, [Bruce et al. 92] report a success rate of 80% (really 90%, hand-coding only ten words). This impressive rate is achieved using several intrinsic characteristics of LDOCE: dictionary senses are ordered by frequency, 86% of dictionary senses have semantic codes and 44% of dictionary senses have pragmatic codes.

Using only the implicit lexical knowledge present in a Spanish MRD, we report (see Section 5.3) an accuracy of 83% at a sense level (an average of 5.75 senses per noun) and 79% for polysemous nouns (an average of 6.65 senses per noun). The method we propose combines several sources of lexical information appliying several unsupervised methods.

4.3.4 Measures of semantic relatedness

Dictionary definitions (written for human readers) do no more than provide a starting point to initiate the understanding of the meaning of the concept underlying these definitions. That is, a human user must still search out a vast amount of common sense knowledge, perform a complex reasoning, and decide on a satisfactory meaning for the definition.

How can an automatic procedure (without having manually coded semantic information and without knowing anything about the world) select the correct hypernym dictionary sense (and then disambiguate the genus term) from a set of possible candidates (ranging from 1 to 39) considering only the definitions themselves (including the hyponym definition)?

A large number of different studies have been performed in order to establish measures of relatedness between words. From a) those that collect cooccurrence evidence from corpora (e.g., [Church & Hanks 90]) or MRDs (seen as corpora) (e.g., [Wilks et al. 93]) to b) those that gather evidence from preexisting lexical knowledge resources (such as WordNet) (e.g., [Agirre & Rigau 96]) or c) others that combine both approaches (e.g., [Resnik 97], [Rigau et al. 97]).

Cooccurrence-based measures

Several measures of relatedness between words based on cooccurrence in a text have been described; Mutual Information, t-test, etc. [Church et al. 91], the cosine function [Schütze 92a], conditional probability (and variations), intersection over union, dependency extraction and standard deviation [Wilks et al. 93]. Some of them collect the cooccurrence data from corpora while others collect them from structured lexical resources such as MRDs. [Yarowsky 92] uses a Mutual Information-like measure for computing the saliency weight for each word in each corpus partition. [Grefenstette 92a], [Grefenstette 92b], [Grefenstette 93] propose a similarity between words based on the Jaccard measure on syntagmatic representations of term-attribute pairs extracted from corpora using Sextant.

Using cooccurrence data obtained from MRDs, the scope of relatedness between words is constrained by each definition. Thus, using an MRD, words cooccur if they appear in the same sense definition. While [Niwa & Nitta 94] and [Kasahara et al. 95] only consider the relation between headword and words in the definition, [Wilks et al. 93] and [Rigau et al. 97] collect

---

[1] *Although in the GSD problem no part-of-speech selection is needed, nor (in most cases) selection of the correct lemma.*

[2] *This is why for the GSD problem using LDOCE [Bruce et al. 92] reports a success rate of 80%, yet (using different techniques) for the WSD problem, likewise using LDOCE, [Wilks & Stevenson 97] only reports a success rate of 57%.*

the cooccurrence data from the whole dictionary. [Niwa & Nita 94] compare cooccurrence vectors extracted from the *Wall Street Journal* corpus (20 million words) and distance vectors extracted from *Collins English Dictionary* (1.6 million words) using as a similarity measure between two words the inner product of its normalized vectors. Although collected data from corpora provides better results for the WSD problem, the different size of the two sources (the corpus is more than 10 times larger) means that the quality of the information gathered from the two sources cannot be compared. However, the emerging relations between words using these measures are not always semantic ones, strong relations appears between collocations, compounds, lexical-syntactic patterns, etc. [Karov & Edelman 96] describe a circular converging process between word similarity and context similarity measures.

Other works, rather than collecting cooccurrence data between words aim to collect statistical evidence between words and semantic classes. [Guthrie et al. 91] propose the intersection over union function as a subject-dependant cooccurrence measure collecting data from each of the LDOCE subject semantic codes. In a similar approach, [Yarowsky 92] proposes a Mutual Information-like measure between words and the semantic classes extracted from *Roget's Thesaurus* and [Rigau et al. 97] and [Rigau et al. 98] from WordNet semantic files. Using a different approach, rather than collecting cooccurrence between words and semantic classes, [Liddy & Paik 92] construct an LDOCE subject-code correlation matrix using the Pearson product correlation processing the WSJ corpus.

Knowledge-based measures

In this case, the relatedness among words is inferred from preexisting structured lexical resources.

Less attention has been paid lately to measures of relatedness based on semantic structured hierarchical nets. In this case, taking a structured hierarchical net as a reference, conceptual distance attempts to provide a basis for determining closeness in meaning between words. The conceptual distance between two concepts is defined in [Rada et al. 89] as the length of the shortest path that connects the concepts in a hierarchical semantic net. Besides applying conceptual distance in a medical bibliographic retrieval system and merging several semantic nets, they demonstrate that their measure of conceptual distance is a metric. In a similar approach, [Sussna 93] employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. [Wu & Palmer 94] also propose a conceptual similarity measure for resolving the lexical selection problem in Machine Translation. Extending these ideas, [Agirre et al. 94] describes a new **Conceptual Distance** formula for automatic spelling correction, and [Rigau 94], using this conceptual distance formula, presents a methodology to enrich dictionary senses with WordNet semantic tags. The same measure is used in [Rigau et al. 95] for linking taxonomies extracted from DGILE and LDOCE, in [Rigau et al. 97] as one of the methods for the GSD problem in DGILE, in [Rigau et al. 98] during the first labeling process, and in [Atserias et al. 97] as one of the methods for attaching Spanish words to WordNet synsets.

Conceptual distance provides a basis for determining closeness in meaning among words, taking a structured hierarchical net as a reference. The conceptual distance between two concepts is essentially the length of the shortest path that connects the concepts in a hierarchical net. To compute the distance between any two words (w_1,w_2) all the corresponding concepts (or senses) are sought, and the minimum of the sum of the inverse depth in the path between each possible combination of c_{1i} and c_{2j} is returned.

(4.1)
$$dist(w_1,w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_i} \in w_2}} \sum_{c_k \in path(c_{1_i},c_{2_i})} \frac{1}{depth(c_k)}$$

That is, the conceptual distance between two concepts depends on the length of the shortest path that connects them and the specificity of the concepts in the path. That is to say, the lower the concepts are in a hierarchy, the closer they seem to be.

Conceptual Density, a more complex semantic measure between words, is defined in [Agirre & Rigau 95] and used in [Agirre & Rigau 96a] as a proposal for WSD of the Brown corpus. Basically, Conceptual density compares areas of subhierarchies and should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.
- the measure should be independent of the number of concepts involved.

To illustrate how Conceptual Density can aid to disambiguate a word, in figure 4.1 the word W has four senses and several context words. Each sense of the words belongs to a subhierarchy of WordNet. The dots in the subhierarchies represent the senses of either the word to be disambiguated (W) or the words in the context. Conceptual Density will yield the highest density for the subhierarchy containing more senses of those, relative to the total amount of senses in the subhierarchy. The sense of W contained in the subhierarchy with highest Conceptual Density will be chosen as the sense disambiguating W in the given context. In figure 4.1, sense2 would be chosen.

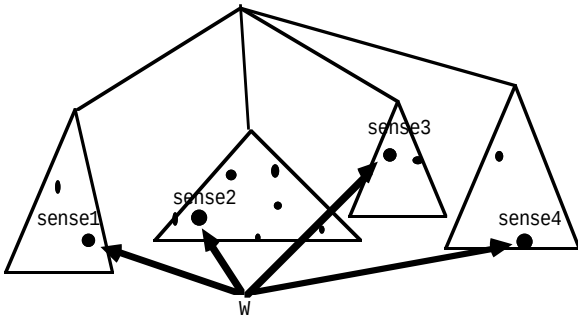


Figure 4.1: senses of a word in WordNet .

Given a concept *c*, at the top of a subhierarchy, and given *nhyp* (mean number of hyponyms per node), the Conceptual Density for *c* when its subhierarchy contains a number *m* (marks) of senses of the words to disambiguate is given by the formula 4.2.

(4.2)
$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{\alpha}}}{descendants_c}$$

Formula 4.2 shows a parameter α which tries to smooth the exponential *i*, as *m* ranges between 1 and the total number of senses in WordNet. Several values were tried for the parameter, and it was found that the best performance was attained consistently when the parameter was near 0.20 (see [Agirre & Rigau 95]).

Combining cooccurrence-based and knowledge-based measures

Using cooccurrence measures on data gathered from corpora (or MRDs), the relations that emerge between words do not always reflect semantic relations. Furthermore, measures computed on semantic nets like WordNet (which do not contain inter-category semantic relations) provide only a few semantic relations (class/subclass, meronym relations, entailment, etc.). That is, it seems that by combining and exploiting both approaches together a more powerful method would be obtained.

Thus, others have proposed a combination of both cooccurrence-based and knowledge-based measures. [Resnik 93] and [Resnik 95] capture semantic similarity (closely related to

conceptual distance) by means of the information content of the concepts in a hierarchical net, combining semantic classes taken from WordNet with cooccurrence data extracted from corpora. [Richardson et al. 94] and [Jiang & Conrath 97] also combine WordNet and informational measures taken from corpora.

Using MRDs, [Kozima & Furugori 93] define a measure of similarity between words which depends on the spreading activation of a semantic network derived from LDOCE and a word significance of the words actively collected from corpora. [Fukumoto & Suzuki 96] collect cooccurrence data extracted from corpora to construct Mutual Information vectors for characterizing dictionary definitions. In [Rigau et al. 97] eight different heuristics using different cooccurrence-based and knowledge-based measures are combined, their individual performances being added together produce better results than stand-alone. Some of these heuristics use cooccurrence-based measures gathered from monolingual and bilingual MRDs used as corpora, while others are knowledge-based ones taking WordNet as a semantic reference (also using bilingual dictionaries to apply the knowledge-based measures to Spanish and French as opposed to English). [Richardson 97] applies several statistical measures (e.g., inverse document frequency, mutual information, etc.) in order to weight the labeled semantic relations extracted from several MRDs.

4.4 Semantic knowledge acquisition from the differentia

4.4.1 Parsing dictionary definitions

Following our methodology (see Chapter 3), once a disambiguated taxonomy is created and all dictionary senses included are connected by hypernym links (except the top ones, which are connected to the Type System) and hyponym links (except the terminal dictionary senses), a further semantic enrichment process can be performed. Thus, instead of limiting the lexical acquisition to taxonomies, some approaches perform an in-depth analysis of the dictionary definitions, taking advantage of the sublanguage used by lexicographers to define dictionary senses.

Therefore, knowledge appearing in the differentia [Calzolari 91] of the definition has to be extracted and assigned to the appropriate semantic roles in the LKB. This task involves a more in-depth analysis of these definitions.

A number of research groups are currently developing parsing systems capable of analysing natural language text robustly and accurately. Such systems, varying in the depth of analysis from lexical parsing or tagging (identifying syntactic features for individual words only) [Karlsson et al. 95], through shallow or phrasal parsing (finding phrases, e.g., NPs, or forming a hierarchical syntactic structure but not exploiting subcategorization) [Abney 91], to full parsers (which deal with unbounded dependencies, etc., and are able to recover predicate-argument structure) [Briscoe & Carroll 95].

Some early works on acquiring lexical knowledge from definitions perform a simple string pattern matching approach (i.e., [Chodorow et al. 85], [Markowitz et al. 86], [Nakamura & Nagao 88]). Other researchers (i.e., [Alshawi 89], [Artola 93], [Castellón 93]) have preferred partial rather than full parsing of the definitions (i.e., [Fox et al. 88], [Jensen & Binot 88], [Dolan et al. 93], [Vanderwende 95]). If the analysis is performed correctly[1], a wide-range parsing tool (capable of parsing a dictionary definition completely) could extract more and better information than a partial one. Nevertheless, none of these researchers provide accurate information either on the coverage (which analyses are performed out of those that

[1] *[Briscoe & Carroll 95] report a 51% correct analysis, ranking in the top three for sentences of less than 20 words. However, what a "correct" analysis means is a topic of hard discussion not addressed in the current work.*

are possible) or the accuracy (analyses performed correctly) of the analyses carried out on the dictionaries they are working on[1].

[Fox et al. 88] analysed a sample of W7 and CDEL using the LSP parsing tool [Sager 81] with an average success of 69%.

[Alshawi 89] define a hierarchy of partial patterns. The parsing procedure starts from the top of the hierarchy of patterns to the bottom. If a partial pattern performs a match, then the procedure attempts to continue the analysis with its direct descendants, which describe more specific patterns. This recursive analysis ends when no further specific pattern is provided in the grammar or when the matching procedure fails. In the first case, the procedure returns the last successful match. Analysing LDOCE and using a grammar of 90 patterns, [Alshawi 89] reports 77% correct identification of the genus, 61% of cases obtaining additional information, 88% of which were considered correct.

[Artola 93] and [Castellón 93] also use the analysis technique proposed by [Alshawi 89]. [Artola 93] reports out of 58% of complete analysis of noun definitions[2] (80% for verbs and 69% for adjectives) using a hierarchy of 65 patterns (49 for verbs and 45 for adjectives). For [Castellón 93] grammar depends on the domain analysed. For substance definitions she uses 36 patterns; 36 for tools; 31 for persons and 38 for places.

[Dolan et al. 93], [Vanderwende 95] and [Richardson 97] used the Microsoft English Grammar (MEG). MEG consists of a set of augmented phrase structure grammar rules which analyse the definitions using a bottom-up chart parsing engine.

Rather than a single shot process, [Vanderwende 95] proposes a cycling methodology improving the analysis with the knowledge acquired during the previous cycle.

4.4.2 Placing the semantic knowledge in the LKB

Taking advantage of the defining formulae, which are "significant recurring phrases" [Markowitz et al. 86], some early works perform a string pattern matching approach (i.e., [Chodorow et al. 85], [Markowitz et al. 86]) to build the LKB directly.

Others prefer structural patterns that match the syntactic analysis (i.e., [Jensen & Binot 87], [Alshawi 89], [Ravin 90], [Klavans et al. 90], [Artola 93], [Castellón 93], [Dolan et al. 93]). While some approaches only consider a one-to-one relation between the defining formulae and the type of lexical information it identifies (i.e., [Jensen & Binot 87], [Artola 93]), later studies (i.e., [Ravin 90], [Klavans et al. 90], [Vanderwende 95]) have shown that some defining formulae can convey several types of semantic information. In that sense, [Castellón 93] preferred to process subsets of closely related dictionary senses (i.e., taxonomies) in order to extract different semantic information from different semantic domains. Thus, instead of a single grammar for analysing all dictionary definitions it seems more feasible to build several (one for each semantic primitive) domain-specific grammars allowing in-depth semantic acquisition from the differentia.

[Alshawi 89] defines, for each syntactic pattern defined in the hierarchy, a semantic rule which makes it possible to specify the semantic structure to be produced. [Castellón 93] collects, from a total number of 2,433 definitions analysed, 5,730 syntactic and semantic relations[3] (including hypernymy), that is, 2.35 relations per definition.

[Dolan et al. 93] report a 78% overall accuracy extracting semantic relations. The hypernymy relation (about the half the relations) was judged to be accurate in 87% of the cases. From a total number of 45,000 definitions the acquisition procedure extracts 94,000 semantic relations (including hypernymy). That is, 2 semantic relations per definition.

---

[1] *They evaluate the results (if it exist) by providing the total amount of semantic relationships collected.*

[2] *LPPL has 3.82 mean number of words per noun definition.*

[3] *Several syntactic relationships collected (i.e., prepositional phrases) were not semantically interpreted.*

4.5 Multilingual mapping of lexical units

As for monolingual lexicons, three major classes of techniques for multilingual lexical acquisition have been developed: machine-aided manual construcion, extraction from corpora and extraction from MRDs. Each of these techniques have advantages and disandvantages. Various software tools have been developed which make manual lexicon construction and maintenence quicker and easier. This approach has been taken in large-scale machine translation systems, see for example [Niremburg & Raskin 88] and a description of METAL in [Hutchins & Somers 92]. Manual construction is the most reliable technique, but is a hard time-consuming approach to encode the multilingual lexicon.

Automatic or (semi)automatic extraction of information from bilingual corpora is very promising, but a vast amount of data is needed for reliable entries to be constructed on any but the most common words. Unless the corpus closely matches the intended text type of the NLP system, relevant senses of words are likely to be missing. While for unstructured multilingual lexical knowledge resources a great deal of work has been carried out, less attention has been devoted to structured resources like bilingual or multilingual dictionaries. An important point as regards the use of lexical resources is availability. Multilingual organizations (such as the Canadian parliament with the Canadian Hansard bilingual corpus, the United Nations or the European Union) have provided the research community with multilingual corpora produced by them.

Bilingual corpora can be used for many purposes [Dagan et al. 91], among others, to acquire new word-to-word lexical correspondances [Smadja 92]. Lexicon compilation methods mainly attempt to extract pairs of words or compounds that are translations of each other from previously sentence-aligned parallel texts (e.g., [Eijk 93], [Kumano & Hirakawa 94] or [Utsuro et al. 94]). Bilingual corpora alignment can be performed at character, word or sentence level (e.g., [Brown et al. 91b], [Gale & Church 91], [Church 93] or [Kupiek 93]). Furthermore, [Fung 95] proposes an algorithm for bilingual lexicon acquisition that bootstraps off the corpus alignment process.

However, valuable contributions have also been made using bilingual MRDs. [Rizk 89] studies the problem of ambiguous sense references in an English/French dictionary. [Tanaka & Umemura 94] use two intermediate Japanese/English and French/English bilingual dictionaries to construct a Japanese/French bilingual dictionary automatically.

Furthermore, some approaches have been proposed for linking semantic structures in different languages. Thus, [Ageno et al. 94] use a Spanish/English bilingual dictionary to link Spanish and English taxonomies extracted from DGILE and LDOCE (semi)automatically, attaching 95% of the Spanish entries having only 31.5% bilingual correspondence (due to the different sizes of the bilingual and monolingual dictionaries). In a similar approach, [Rigau et al. 95] propose an automatic approach for linking Spanish taxonomies extracted from DGILE to WordNet synsets.

Others have proposed to link words from one language to semantic structures in another. Thus, [Knight & Luk 94] use the Collins Spanish/English bilingual dictionary to link Spanish words to WordNet using overlapping techniques, but they do not provide accuracy figures for the 50,000 proposed mappings. [Okumura & Hovy 94] describe semi-automatic methods for associating a Japanese lexicon with an English ontology using a bilingual dictionary. They report an accuracy for nouns of over 55%, 42% for verbs and 48% for adjectives. In a similar approach, [Rigau & Agirre 95] propose several complementary techniques for attaching directly Spanish and French words extracted from the bilingual dictionaries to WordNet synsets. In this work, a total accuracy of 85% is reported using conceptual density techniques for 47% of French entries, and 78% for 91% of Spanish entries exploiting some inherent properties of the lexical resources used. Extending these ideas, [Atserias et al. 97] combine several monolingual and bilingual resources and several techniques to map Spanish words from a bilingual dictionary to WordNet in order to build a semantic net with a parallel structure. In this work, a total accuracy of 85% is reported for 75% of Spanish entries. In that sense this approach is similar to that presented by [Artale et al. 97] with the multilingual lexical matrix. Furthermore, [Farreres et al. 98] propose to aid this process using the taxonomic structure collected from a monolingual Spanish MRD.

Finally, some attempts have been made to combine structured and non-structured lexical resources in order to build bilingual lexicons. Thus, [Klavans & Tzoukermann 96] use the Collins Robert bilingual French/English MRD in conjunction with the Hansard corpus for the creation of a bilingual LDB.

4.6 Validation of the Lexical Knowledge Base

This Section descrives the LDB/LKB merging system, a system we developed to provide the LDB query mechanism to the LKB. Thus, using this new software on the lexical entries loaded into the MLKB the user can validate and evaluate the knowledge acquired in the previous steps of the methodology.

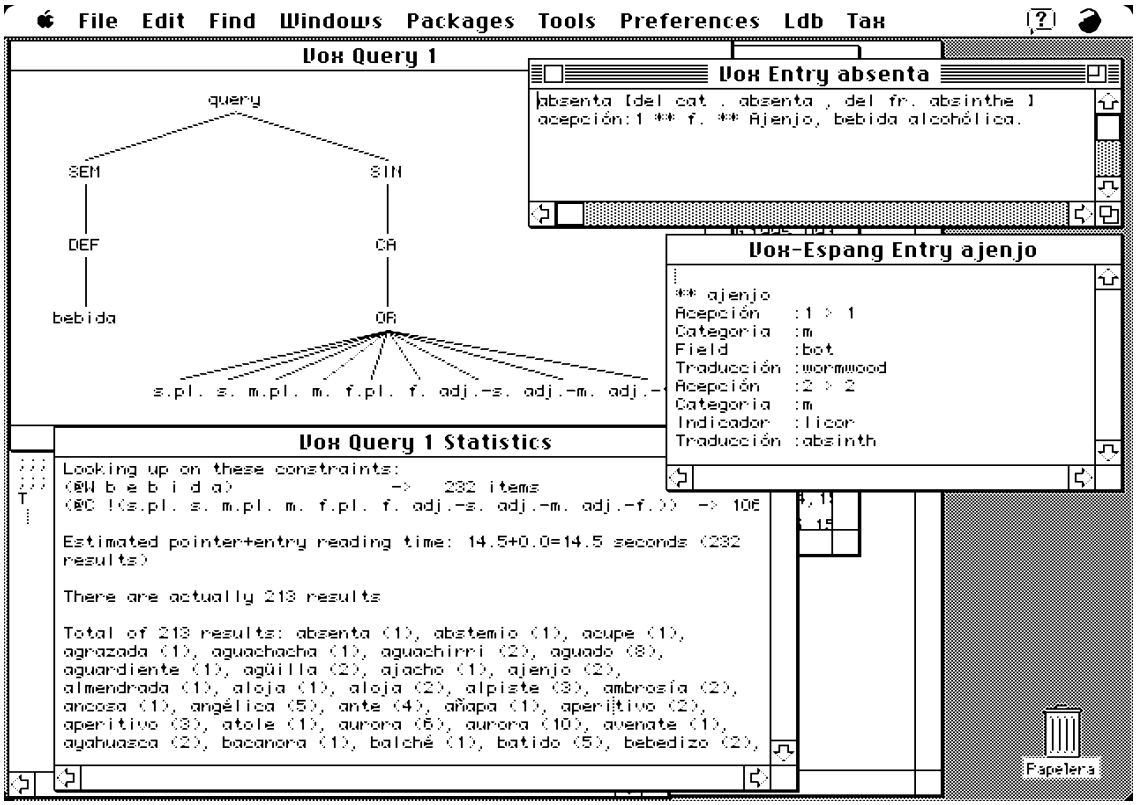
4.6.1 Querying the Lexical Knowledge Base

Once the information contained in the dictionary definitions has been represented as a lexicon in the LKB, some testing processes should be performed on the resulting lexicon in order to improve the information acquired (e.g., to detect possible errors or inconsistencies, derive more information, etc.) to determine which changes to make in the next acquisition loop. For this purpose the two basic representational formalisms used in Acquilex to represent lexical entries (the LDB and the LKB) provided no sufficient capabilities. The LDB provides basically database-like access to lexical information, while the LKB software manages a lexical knowledge representation based on typed Feature Structures (FSs) and defines valid operations on entries. Thus, the LKB system does not provide facilities for performing complex consultations on the content of the lexical entries represented in the lexicon. The LKB only guarantees the appropriateness of the lexicon against the Type System and provides some generative inference mechanisms (e.g., the inheritance mechanism distributes the information from the top level lexical units to the most specific ones, lexical rules produce new lexical entries from the preexisting ones, etc.). For the purposes of both the validation and the exploitation of the information acquired, it would be useful to have a new system which had the function of both systems: LDB-like access to an LKB lexicon.

4.6.2 The LDB

Within Acquilex, the Lexical Data Base (LDB) that implements the two level dictionary access model [Boguraev et al. 91] was implemented to gives flexible access to MRDs. The LDB is endowed with a graphic interface which provides a user-friendly environment for query formation and information retrieval. It allows several dictionaries to be loaded and queried in parallel. Thus, the LDB provides facilities for loading and indexing multiple MRDs, displaying its lexical entries and querying them on any part of its content. A brief description of the LDB can be found in [Carroll 92].

Screen 4.1 shows a typical LDB session with a tree-like query to DGILE dictionary and two different dictionary entries. The query asks for those DGILE noun dictionary senses with the word bebida in its definition (Vox Query 1 window). From the total number of senses in DGILE only 213 hold the query constraints (Vox Query 1 Statistics window). Among these, the user has selected (clicking on it) the first one, *absenta* (absinth). The LDB system has displayed the entry *absenta* (Vox Entry absenta window). Then, the user has asked for the entry *ajenjo* from the Spanish/English bilingual dictionary (Vox-Espang Entry ajenjo window).



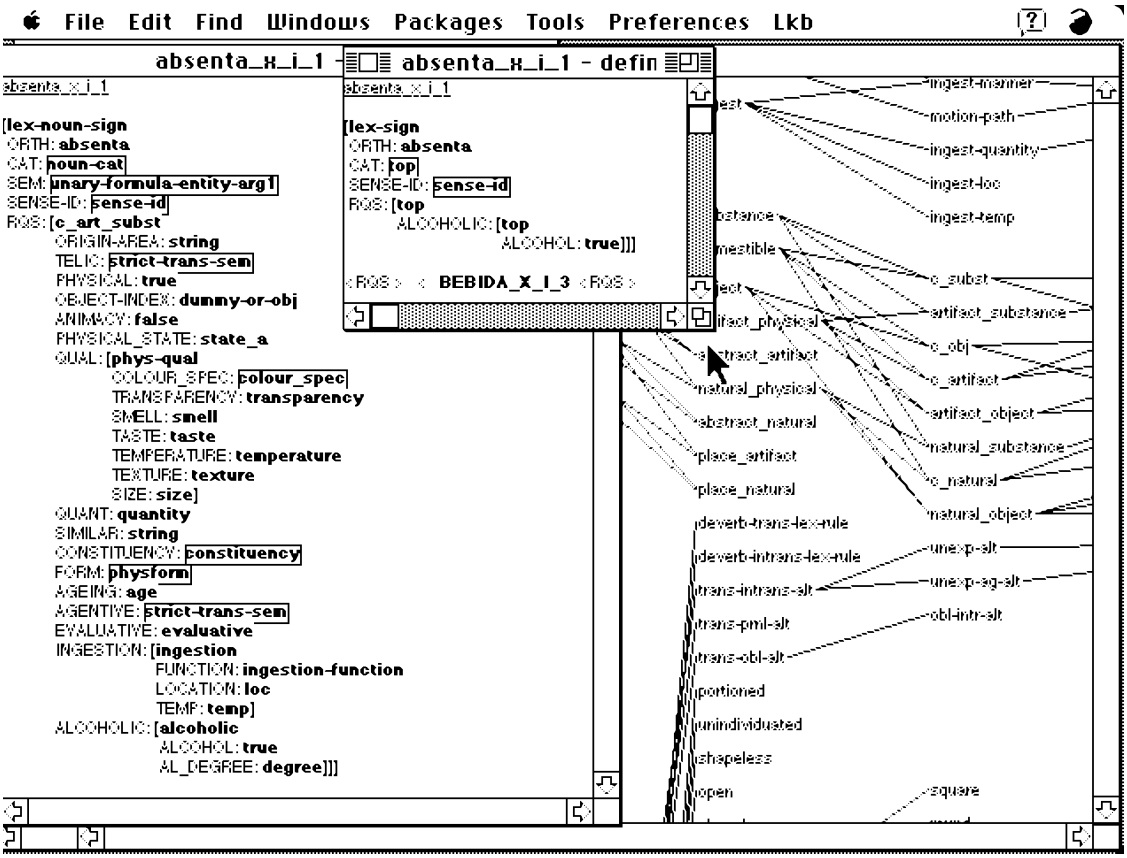
Screen 4.1, an LDB session.

4.6.3 The LKB

As we said previously in Section 3.5.2, the operations that the LKB supports are (default) inheritance, (default) unification and lexical rule application. The two main components of the LKB are the Type System and the Lexicon. The Type System represented as a type hierarchy defines a partial order (noted $\subseteq$,, "is more specific than") on the types and establishes consistency conditions. Because the type hierarchy is a partial order it has properties of reflexivity, transitivity and anti-symmetry (from which it follows that the type hierarchy cannot contains cycles). The Type System supports only non-default inheritance, that is monotonic, multiple and orthogonal while the typed feature structure system has been extended with multiple default inheritance mechanism in order to represent the lexical entries in an appropriate way.

Thus, the LKB provides facilities for creating type systems, loading lexicons and displaying fully expanded feature structures, type checking, and so forth. The Type System which has been developed for use in Acquilex project is fairly large (about 500 types and 200 features) and currently nearly 1000 lexical entries for Spanish nouns and verbs containing syntactic and semantic information have been stored within it. A brief description of the LKB System can be found in [Copestake 92a] and a complete one in [Copestake 92b].

Screen 4.2 shows in the right side a partial view of the Type System and in the left one the expanded and defined lexical entry for *absenta* (absinth). The expanded lexical entry for *absenta* contains information inherited from the type hierarchy or its hypernym lexical entries.



Screen 4.2, an LKB session.

4.6.4 The LDB/LKB integration

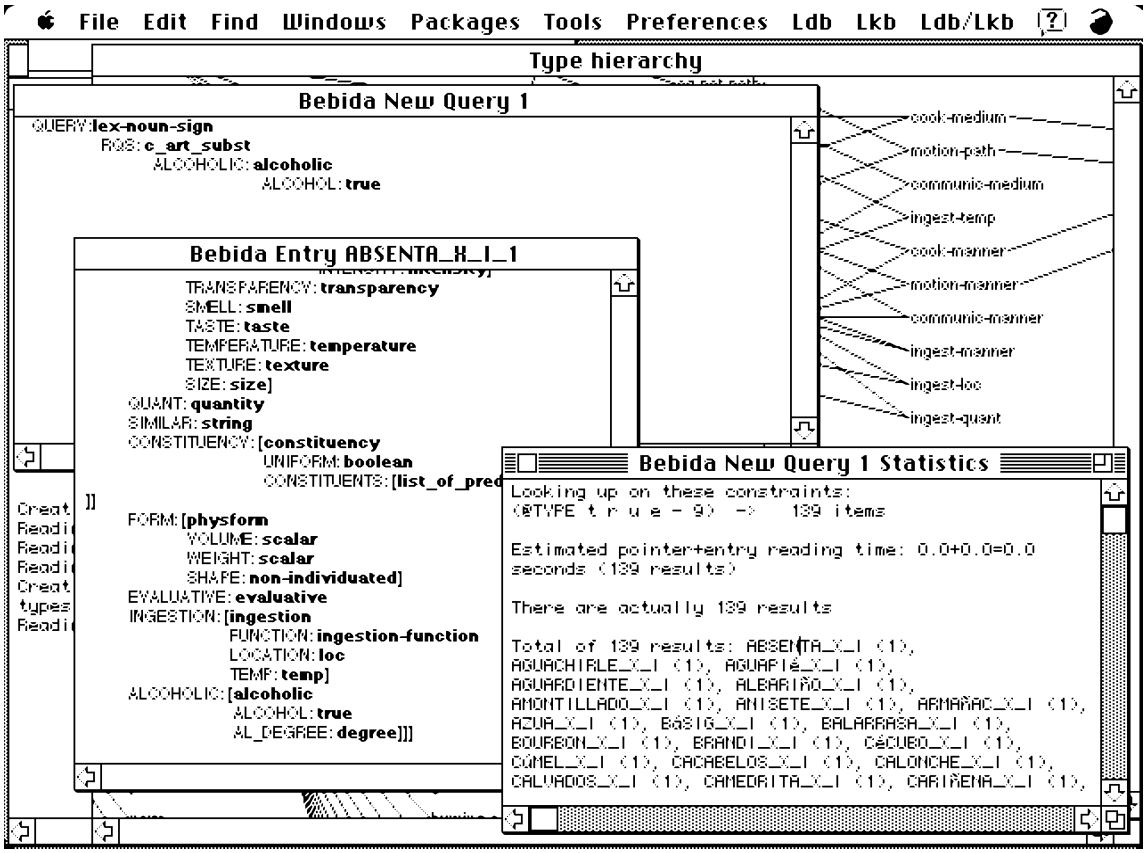
The LDB/LKB merging system [Rigau et al. 94] has been developed taking into account a central guideline: LKB lexicons [Briscoe et al. 90], [Copestake, 92a] can be expressed, loaded and stored as in any other dictionary, in such a way that the LDB software [Carroll 90a] can be used without modifications or restrictions. The original LKB entries can be reconstructed from their LDB representation. This allows us to replace the LKB's lexical reading and access mechanism with the LDB functions, which gets round the current problem that reading in LKB lexicons is very slow, showing a considerable drop in performance when faced with real-size lexicons, and in the long term will allow for efficient access to indexed.

The central idea of loading lexicon files like other dictionary source files into the LDB environment seems quite straightforward, but several problems (e.g., how to describe sources, what information index or how to access indexed information) arise when it is approached in detail.

A new graphic query interface allowing an LKB-like user-friendly interaction with the LDB/LKB system has been implemented. Thus, the query construction with the manipulation of FS chunks within the Type System (clicking on it), the lexical entries and the query windows have been carried out.

As stated above, the LDB/LKB system not only has the functions of both the LDB and the LKB software but also offers new features, (e.g., partial queries, subsumption queries, etc.). Thus, the LDB/LKB system performs the construction of the LDB lexicons and indexes from the lexical entries loaded in the LKB in a fully automatic way, consulting the Type System and the lexicon.

A new query mechanism built over the existing one allows the LDB/LKB system to perform the least possible number of queries on the indexes and the LDB lexicons, consulting the Type System and the identifier structure.



Screen 4.3, an LDB/LKB session.

Screen 4.3 shows a typical LDB/LKB session with a feature/value-like query to DGILE lexicon and a LKB lexical entry. The query asks for those lexical entries loaded in the LKB marked as alcoholic (Bebida New Query 1 window). From the total number of lexical entries loaded into the LKB only 130 hold the query constraints. Among these, the user has selected (clicking on it) the first one, *absenta* (absinth). The LDB/LKB system has displayed the entry *absenta* (Bebida Entry absenta window).

4.7. Conclusions

This chapter has described the main issues relating to the acquisition of lexical knowledge using SEISD. The first section explains different approaches for classifying the concepts present in an MRD. Section 2 deals with the methodological considerations for the selection of the main semantic subsets described in an MRD and its most representative dictionary senses. Section 3 is devoted to several approaches to the Genus Sense Disambiguation (GSD) problem. Section 4 deals with the extraction of the main semantic relations from the dictionary definitions and their mapping onto the LKB, and Section 5 with the multilingual enrichment of the LKB. Finally, Section 6 explains the LDB/LKB merging system, the main mechanism we developed and used for the validation and exploitation of the multilingual LKB.

Chapter 5

Monolingual Lexical Knowledge Acquisition

5.1. Introduction

This chapter covers the main experiments and results in the acquisition of lexical knowledge by using SEISD on the monolingual dictionary *Diccionario General Ilustrado de la Lengua Española* (DGILE). As we said in Section 1.4, our methodology for acquiring lexical knowledge from conventional dictionaries is divided into six partial steps. In Chapter 4 we performed an extended overview of the critical issues related to each step, and now in this Chapter we describe the main experiments we carried out using SEISD acquiring lexical knowledge from DGILE. Now, we set in again the steps of the methodology with the sections covered by this Chapter.

After this introduction, Section 2 deals with the first step of the methodology (justified in Section 4.2), that is, the automatic selection of the main semantic primitives present in DGILE, and for each of these, the selection of its most representative dictionary senses. Section 3 is devoted to the second step of the methodology: the automatic acquisition of taxonomies from DGILE, that is, the acquisition of lexical knowledge from the genus terms (which is performed by TaxBuild, see Sections 3.6.1.3 and 4.3). Finally, Section 4 deals with the third and fourth steps of the methodology, that is, the acquisition of lexical knowledge from the differentia (step three, the analysis, performed by SemBuild and step four, the conversion to the LKB, by CRS; see Sections 3.6.1.4, 3.6.2 and 4.4). Step five of the methodology (acquisition of multilingual knowledge, performed mainly by TGE) is presented in Chapter 6 and step six (validation of the knowledge acquired, performed by the LDB/LKB system) was presented in Section 4.4.

Although we applied the whole methodology on a particular MRD, DGILE is a conventional MRD general enough (that is, we do not use any particular characteristic of the MRD) to expect similar performance with other dictionaries. Furthermore, the automatic acquisition of taxonomies described in Section 5.3 was performed also from the French LPPL (a smaller MRD) with comparable results (see [Rigau et al. 97] for further details).

5.2. Main semantic subsets in DGILE

This section presents a novel methodology to select the semantic primitives implicitly defined in a conventional dictionary, and second, the studies carried out for discovering the main top dictionary senses representative of a given semantic primitive.

5.2.1 Predefined semantic primitives in DGILE

As we said in Section 4.2, following a purely descriptive methodology, the predefined semantic primitives of a dictionary can be obtained by collecting those dictionary senses appearing at the top of the complete taxonomy derived from the dictionary. For DGILE, we have derived the complete noun taxonomy following the automatic method described in Section 5.3 (also described in [Rigau et al. 97]). This taxonomy contains 111,624 dictionary

senses and has only 832 dictionary senses which are tops of the taxonomy (these top dictionary senses have no hypernyms), and 89,458 leaves (which have no hyponyms). That is, 21,334 definitions are placed between the top nodes and the leaves. The average number of direct hyponyms per node is 5.01. Table 5.1 shows the top ten dictionary senses with most descendants.

11,148	ejecución_1_1	execution	5,503	efecto_1_2	quality
11,064	entidad_1_1	entity	3,529	animal_1_2	animal
8,707	persona_1_1	person	1,851	línea_1_5	line
8,569	resultado_1_1	result	1,554	efecto_1_1	effect
5,509	calidad_1_1	quality	1,584	modo_1_1	manner

Table 5.1, top ten dictionary senses and number of descendants.

Then, following a purely descriptive approach, each of the 832 partial taxonomies represent a superclass or semantic primitive in DGILE. By characterizing each one, the complete LKB could be produced. However, with this approach many problems arise (see Section 4.2), most of them produced by the dictionary itself (circularity, errors and inconsistencies, definitions with omitted genus, etc.), others because no perfect GSD process can be performed on conventional dictionaries ([Bruce et al. 92] report a success rate of 80% and [Rigau et al. 97] report 83% accuracy), and finally, others by the fact that a top dictionary sense does not represent (or have the characteristics) of those superclasses we wish to characterize in the LKB in order to represent useful knowledge for a NLP system[1]. To illustrate the problem we are facing, let us suppose we plan to place the FOOD concepts in the LKB. Neither collecting the taxonomies derived from a top dictionary sense (or selecting a subset of the top dictionary senses of DGILE) closest to FOOD concepts (e.g., *substancia* -substance-), nor collecting those subtaxonomies starting from closely related senses (e.g., *bebida* -drinkable liquids- and *alimento* -food-) we are able to collect exactly the FOOD concepts present in the MRD. The first are too general (they would cover non-FOOD concepts) and the second are too specific (they would not cover all FOOD dictionary senses because FOODs are described in many ways).

In other words, there is a mismatch between the knowledge directly derived from an MRD and the knowledge needed by the LKB. As we said previously (see Sections 3.2.4 and 4.2), to overcome this problem we adopted a mixed methodology. First, we prescribed a minimal ontology (represented by the Type System of the LKB) capable of representing the whole lexicon derived from the MRD, and second, following a descriptive approach, we collect, for each semantic primitive placed in the Type System, its subtaxonomies. Thus, those subtaxonomies selected for a semantic primitive are attached to a type of the LKB.

The following sections show how to fit the taxonomies collected from an MRD using a descriptive approach into a prescribed set of semantic primitives. To illustrate the process, instead of the Type System we used as semantic primitives the 24 lexicographer's files (or semantic files) into which the 60,557 noun synsets (87,641 nouns) of WordNet 1.5 are classified[2]. Thus, we are considering the 24 semantic tags of WordNet as the main LKB types to which all dictionary senses must be attached. In order to overcome the language gap we also use a bilingual Spanish/English dictionary. The following section shows the method we used to classify all nominal DGILE senses to respect the 24 WordNet semantic files (or semantic tags), and Section 5.2.3 explains the selection of the main top dictionary senses for a

[1] *Recall the example in Section 4.2.1 where INSTRUMENTS have some features -- for instance, PURPOSE -- that the general things lack.*

[2] *One could use other semantic classifications, such as Roget's Thesaurus [Yarowsky 92], the LDOCE semantic or pracmatic codes [Slator 91] or even better, a Spanish semantic classification such as the "Diccionario Ideológico de la Lengua Española J. Casares" (DILEC). Really, when using this methodology a minimal set of informed seeds are needed. These seeds can be collected from MRDs, thesauri or even by introspection. (see [Yarowsky 95]).*

given semantic primitive (in particular, the method is applied to FOOD, file 13 of WordNet).

5.2.2 Attaching DGILE dictionary senses to semantic primitives

In order to classify all nominal DGILE senses to respect WordNet semantic files, we used a similar approach to that suggested by [Yarowsky 92], that is, to enrich dictionary definitions using an on-line thesaurus (in this case, the 24 WordNet lexicographer's files rather than Roget's 1042 categories). While Yarowsky uses *Grolier's Encyclopaedia* to collect the salient words for each category we use the dictionary itself. However, rather than collect evidence from a blurred corpus (words belonging to a Roget's category are used as seeds to collect a subcorpus for that category; that is, a window context produced by a seed can be placed in several subcorpora), we collect evidence from dictionary senses labelled by a conceptual distance method (that is, a definition is assigned to a unique semantic file).

This task is divided into three fully automatic consecutive subtasks. First, we tag a subset (due to the difference in size between the monolingual and the bilingual dictionaries) of DGILE dictionary senses by means of a process that uses the conceptual distance formula; second, we collect salient words for each tag; and third, we enrich each DGILE dictionary sense with a semantic tag collecting evidence from the salient words previously computed.

5.2.2.1 Attach WordNet synsets to DGILE headwords

For all DGILE definitions, the conceptual distance (see Section 4.3.4) between headword and genus has been computed using WordNet 1.5 as a semantic net. However, not all headwords and genus terms have English translations in the bilingual dictionary we used (HBil, see Section 3.2.1). Thus, we obtained results only for those definitions having English translations for both headword and genus. By computing the conceptual distance between two words (w_1,w_2) we are also selecting those concepts (c_{1i},c_{2j}) which represent them and seem to be closer with respect to the semantic net used. Conceptual distance is computed using formula (5.1).

(5.1)
$$dist(w_1,w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_i} \in w_2}} \sum_{c_k \in path(c_{1_i},c_{2_i})} \frac{1}{depth(c_k)}$$

We derived from DGILE a lexicon of 92,693 noun definitions, selecting the genus term for each (using a specialized grammar for detecting the noun genus term and disturbed heads with a success rate of 97.7%). Table 5.2 summarizes all data.

a	Noun definitions	93,394	
b	Noun definitions with genus	92,693	
c	Genus terms	14,131	
d	Genus terms with bilingual translation	7,610	54% of c)
e	Genus terms with WordNet translation	7,319	52% of c)
f	Headwords	53,455	
g	Headwords with bilingual translations	11,407	21% of f)
h	Headwords with WordNet translations	10,667	20% of f)
i	Definitions with bilingual translations	30,446	33% of b)
j	Definitions with WordNet translations	28,995	31% of b)

Table 5.2, data of first attachment using conceptual distance.

To illustrate this process, consider the following example:

abadía_1_2 **Iglesia** o **monasterio** regido por un abad o abadesa. (abbey, *a church or a monastery ruled by an abbot or an abbess*).

where the possible translations (in our bilingual dictionary) for *abadía* are *abbacy* (monosemous) and *abbey* (three possible synsets) and for *iglesia*, *church* (four synsets). That is, the algorithm has to decide between 16 possible combinations (because the two words we are disambiguating have four synsets each). The following table shows the translations found in the bilingual dictionary with their associated WordNet1.5 synsets.

synset	File	English	*Spanish*	Gloss
02038520	artifact	abbey	*abadía*	a monastery ruled by an abbot
02038601	artifact	abbey	*abadía*	convent ruled by an abbess
02038681	artifact	abbey	*abadía*	a church associated with a monastery or convent
05403491	place	abbacy	*abadía*	the jurisdiction or office of an abbot
00572109	act	church	*iglesia*	a service conducted in a church
02291138	artifact	church	*iglesia*	for public (especially Christian) worship
05168576	group	church	*iglesia*	institution to express belief in a divine power
05203171	group	church	*iglesia*	clergymen collectively

Figure 5.1 shows a partial view of the WN1.5 hypernym hierarchy. In bold are a pair of possible translations (using the bilingual dictionary) for *abadía* and *iglesia*. The other hypernym chains involved do not appear because they belongs to complete different taxonomies and are not cross-realted (act, place and group).

Thus, the system computes the Conceptual Distance between all these possible combinations of senses selecting for *abadía* the English translation *abbey* (synset 2038681) and for *iglesia church* (synset 2291138) because using the Conceptual Distance formula they appear closer (one is a direct hyponym of the other). However, in this case, none of the artifact senses for *abadía* would be discarded. The algorithm, in this case, has been too strict.

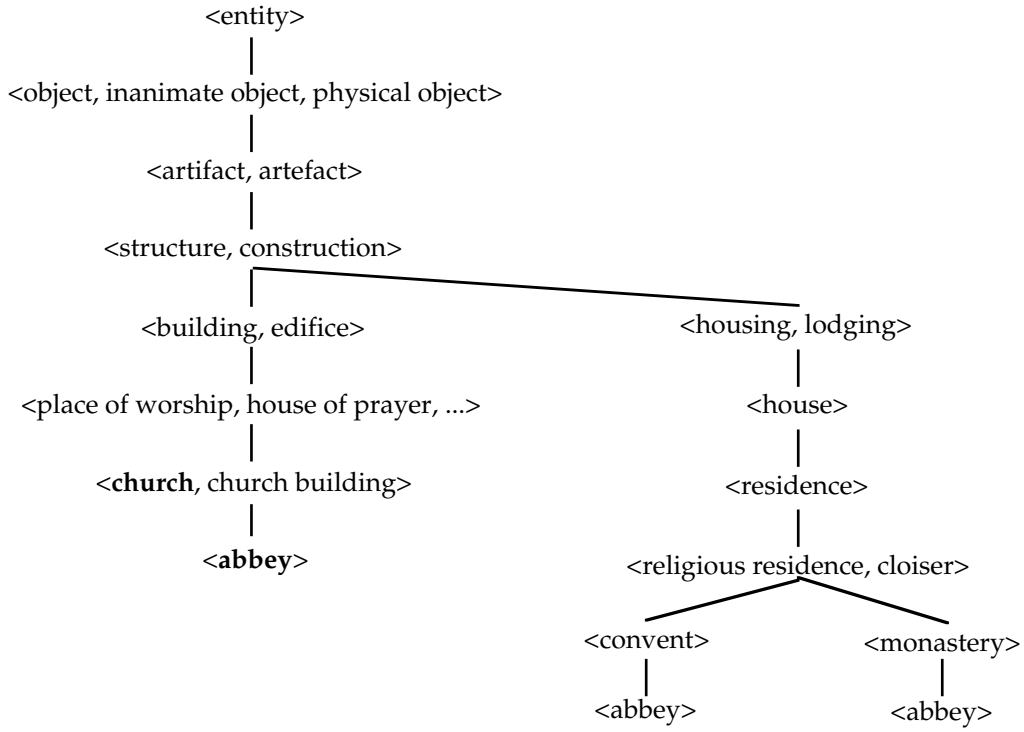


Figure 5.1, partial view of WordNet hierarchy.

As the bilingual dictionary is not disambiguated to respect WordNet synsets (every Spanish word has been assigned to all possible connections to WordNet synsets via the bilingual translations), the degree of polysemy has increased from 1.22 (WordNet 1.5) to 5.02, and obviously, many of these connections are not correct. This is one of the reasons why after processing the whole dictionary we obtained only an accuracy of 61% at a sense (synset) level (that is, correct synsets attached to Spanish headwords and genus terms) and 64% at a file level (that is, correct lexicographer's file assigned to DGILE dictionary senses)[1]. We processed 32,208[2] dictionary definitions, obtaining 29,205 with a synset assigned to the genus (for the rest we did not obtain a bilingual-WordNet relation between the headword and the genus).

In this way, we obtained a former version of 29,205 dictionary definitions tagged with a semantic tag (which corresponds to a WordNet lexicographer's file) with an accuracy of 64%. That is, a corpus (collection of dictionary senses) classified in 24 partitions (each one corresponding to a semantic tag). Table 5.3 compares the distribution of these DGILE dictionary senses through WordNet semantic tags with respect to WordNet 1.5.

Semantic file	#DGILE senses	#WordNet synsets
03 tops	77 (0.2%)	35 (0.0%)
04 act	3,138 (10.7%)	4895 (8.0%)
05 animal	712 (2.4%)	7,112 (11.7%)
06 artifact	6,915 (23.7%)	9,101 (15.0%)
07attribute	2,078 (7.1%)	2,526 (4.2%)
08 body	621 (2.1%)	1,370 (2.3%)
09 cognition	1,556 (5.3%)	2,007 (3.3%)
10 communication	4,076 (13.9%)	4,115 (6.8%)
11 event	541 (1.8%)	752 (1.2%)
12 feeling	306 (1.0%)	397 (0.6%)
13 food	749 (2.5%)	2,290 (3.8%)
14 group	661 (2.2%)	1,661 (2.7%)
15 place	416 (1.4%)	1,755 (2.9%)
16 motive	15 (0.0%)	28 (0.0%)
17 object	437 (1.5%)	839 (1.4%)
18 person	3,279 (11.2%)	5,563 (9.1%)
19 phenomenon	147 (0.5%)	452 (0.7%)
20 plant	581 (2.0%)	7,971 (13.2%)
21 possession	287 (1.0%)	829 (1.4%)
22 process	211 (0.7%)	445 (0.7%)
23 quantity	344 (1.2%)	1,050 (1.7%)
24 relation	102 (0.3%)	343 (0.6%)
25 shape	165 (0.6%)	284 (0.4%)
26 state	805 (2.7%)	1,870 (3.0%)
27 substance	642 (2.2%)	2,068 (3.4%)
28 time	344 (1.2%)	799 (1.3%)
Total	29,205	60,557

Table 5.3, first attachment of DGILE noun senses to respect WordNet semantic files.

[1] *Although the evaluation of file assignment would seem to be easier than synset assignment ([Ng & Lee 96] report only a 57% agreement manually tagging SemCor), this task was also very difficult because in WordNet many similar meanings are placed in different files.*

[2] *The difference with 30,446 is accounted for by repeated headword and genus for an entry.*

The largest differences appear with the classes ANIMAL and PLANT, which correspond to large taxonomic scientific classifications occurring in WN1.5 but which do not usually appear in a bilingual dictionary.

5.2.2.2 Collect the salient words for every semantic primitive.

Once we have obtained the first DGILE version with semantically labelled definitions, we can collect the salient words (that is, those representative words for a particular category) using a Mutual Information-like formula. Intuitively, a salient word[1] is one that appears significantly more often in the context of a semantic category than at other points in the whole corpus, and hence is a better than average indicator for that semantic category. This idea can be formalized with a Mutual Information-like estimate [Church & Hanks 90], salience:

(5.2)
$$S(w, SC) = \log_2 \frac{\Pr(w|SC)}{\Pr(w)}$$

Where w means word and SC semantic class. The words selected are those most important to the semantic category, where importance is defined as the product of salience and local frequency. That is to say, important words should be distinctive and frequent. This importance is measured by means of formula 5.3, Association Ratio.

(5.3)
$$AR(w, SC) = \Pr(w|SC) \log_2 \frac{\Pr(w|SC)}{\Pr(w)}$$

We performed the training process considering only the content word forms from dictionary definitions[2] and we discarded those salient words with a negative score. Thus, we derived a lexicon of 23,418 salient words (one word can be a salient word for many semantic categories, see Table 5.4). Obviously, a larger bilingual dictionary (with more specific vocabulary) would produce more context per category and a larger and more accurate set of salient words per semantic class.

---

[1] *Instead of performing all the experiments on word lemmas, this study has been carried out using word forms because word forms rather than lemmas are representative of typical usages of the sublanguage used in dictionaries.*

[2] *After discarding functional words.*

Semantic file	#DGILE senses	#Content words	#Salient words
03 tops	77 (0.2%)	540	-
04 act	3138 (10.7%)	16,963	2,593
05 animal	712 (2.4%)	6,191	849
06 artifact	6915 (23.7%)	45,988	4,515
07attribute	2078 (7.1%)	11,069	1,571
08 body	621 (2.1%)	4,285	665
09 cognition	1556 (5.3%)	9,699	1,362
10 communication	4076 (13.9%)	24,633	3,301
11 event	541 (1.8%)	3,071	477
12 feeling	306 (1.0%)	1,623	263
13 food	749 (2.5%)	4,679	717
14 group	661 (2.2%)	4,338	647
15 place	416 (1.4%)	2,587	402
16 motive	15 (0.0%)	87	9
17 object	437 (1.5%)	2,733	412
18 person	3279 (11.2%)	19,273	2,304
19 phenomenon	147 (0.5%)	784	114
20 plant	581 (2.0%)	4,965	700
21 possession	287 (1.0%)	1,712	278
22 process	211 (0.7%)	987	177
23 quantity	344 (1.2%)	2,179	317
24 relation	102 (0.3%)	600	76
25 shape	165 (0.6%)	1,040	172
26 state	805 (2.7%)	4,469	712
27 substance	642 (2.2%)	5,002	734
28 time	344 (1.2%)	2,172	321
Total	32,208	181,669	23,418

Table 5.4, salient words per context.

5.2.2.3 Enrich DGILE definitions with WordNet semantic primitives.

Using the salient words per category (or semantic class) gathered in the previous step we labelled the DGILE dictionary definitions again. When any of the salient words appear in a definition, there is evidence that the word belongs to the category indicated. If several of these words appear, the evidence is compounded. We add together their weights, over all words in context, and determine the category for which the sum is greatest, using (5.4).

(5.4) $$W(SC) = \sum_{w \in definition} AR(w, SC)$$

Thus, we obtained a second semantically labelled version of DGILE (see Table 5.5). This version has 86,759 labelled definitions (covering more than 93% of all noun definitions) with an accuracy rate of 80% (we have gained, since the previous labelled version, 62% coverage and 16% accuracy). Consider for instance the labeled DGILE lexical entry *biberón* (baby bottle). In this case, the first sense is labeled as ARTIFACT and the second as FOOD.

biberón_1_1	**ARTIFACT**	4.8399	Frasco de cristal ...	glass frask ...
biberón_1_2	**FOOD**	7.4443	Leche que contiene este frasco ...	milk contained in that frask ...

Semantic file	#DGILE senses (1)	#DGILE senses (2)	#WordNet synsets
03 tops	77 (0.2%)	-	35 (0.0%)
04 act	3138 (10.7%)	4,188 (4.8%)	4895 (8.0%)
05 animal	712 (2.4%)	4,544 (5.2%)	7,112 (11.7%)
06 artifact	6915 (23.7%)	12,958 (14.9%)	9,101 (15.0%)
07attribute	2078 (7.1%)	4,146 (4.8%)	2,526 (4.2%)
08 body	621 (2.1%)	3,208 (3.6%)	1,370 (2.3%)
09 cognition	1556 (5.3%)	3,672 (4.2%)	2,007 (3.3%)
10 communication	4076 (13.9%)	6,012 (6.9%)	4,115 (6.8%)
11 event	541 (1.8%)	1,544 (1.7%)	752 (1.2%)
12 feeling	306 (1.0%)	1,016 (1.2%)	397 (0.6%)
13 food	749 (2.5%)	2,614 (3.0%)	2,290 (3.8%)
14 group	661 (2.2%)	3,074 (3.5%)	1,661 (2.7%)
15 place	416 (1.4%)	2,073 (2.4%)	1,755 (2.9%)
16 motive	15 (0.0%)	22 (0.0%)	28 (0.0%)
17 object	437 (1.5%)	1,645 (1.9%)	839 (1.4%)
18 person	3279 (11.2%)	13,901 (16.0%)	5,563 (9.1%)
19 phenomenon	147 (0.5%)	425 (0.4%)	452 (0.7%)
20 plant	581 (2.0%)	4,234 (4.9%)	7,971 (13.2%)
21 possession	287 (1.0%)	1,033 (1.2%)	829 (1.4%)
22 process	211 (0.7%)	6948 (8.0%)	445 (0.7%)
23 quantity	344 (1.2%)	1,502 (1.7%)	1,050 (1.7%)
24 relation	102 (0.3%)	288 (0.3%)	343 (0.6%)
25 shape	165 (0.6%)	677 (0.8%)	284 (0.4%)
26 state	805 (2.7%)	1,973 (2.3%)	1,870 (3.0%)
27 substance	642 (2.2%)	3,518 (4.0%)	2,068 (3.4%)
28 time	344 (1.2%)	1,544 (1.8%)	799 (1.3%)
Total	32,208	82,759	60,557

Table 5.5, comparison of the two labelling process with to respect WN1.5 semantic tags.

The biggest differences appear (apart from the classes ANIMAL and PLANT) in the classes ACT and PROCESS. This is because during the earlier automatic labelling many dictionary definitions with genus *acción* (act or action) or *efecto* (effect) were classified erroneously as ACT or PROCESS.

These results are difficult to compare. While [Yarowsky 92] extracted the concordances of 100 surrounding words for each occurrence of each word belonging to the Roget's category from the 10-million-word *Grolier's Encyclopaedia*, we are using a smaller context window (the noun dictionary definition has 9.68 words on average) and a microcorpus (181,669 words). By training salient words from a labelled dictionary (only 64% correct) rather than a raw corpus we expected to obtain less noise. Although Yarowsky's work does not report performance labelling dictionary senses, he reports 92% accuracy tagging selected words in a corpus.

Although we used in this work the 24 lexicographer's files of WordNet as semantic primitives, a more fine-grained classification could be made[1]. For example, all FOOD synsets are classified under **<food, nutrient>** synset in file 13. However, FOOD concepts are themselves classified into 11 subclasses.

[1] *In this way, a smaller context (and worse salient words per category) would be obtained. To overcome this problem a larger bilingual dicionary should be used.*

<food, nutrient> -- (any substance that can be metabolized by an organism ...)
=> **<yolk>** -- (nutritive material of an ovum ...)
=> **<gastronomy>**-- (the art or science of food)
=> **<comestible, edible, eatable, ...>** -- (any substance that can be used as food)
=> **<fare>** -- (the food and drink that are regularly consumed)
=> **<foodstuff>** -- (a substance that can be used or prepared for use as food)
=> **<nutriment, nourishment, sustenance, ...>** -- (a source of nourishment)
=> **<cookery, cooking, cuisine, ...>** -- (the practice or manner of preparing food ...)
=> **<commissariat, provisions, food stock, food supply, ...>** -- (a stock of foods)
=> **<feed, provender>** -- (food for domestic livestock)
=> **<miraculous food, manna>**
=> **<beverage, drink, potable>** -- (any liquid suitable for drinking)

Thus, if the LKB we are planning to build needs to represent **<beverage, drink, potable>** separately from the concepts **<comestible, edible, eatable, ...>** a finer set of semantic primitives should be chosen, for instance, considering each direct hyponym of a synset belonging to a semantic file also as a new semantic primitive or even selecting for each semantic file the level of abstraction we need.

A further experiment could be to iterate the process by collecting from the second labelled dictionary (a bigger corpus) a new set of salient words and reestimating again the semantic tags for all dictionary senses (a similar approach is used in [Riloff & Shepherd 97]).

5.2.3 Selecting the main top beginners for a semantic primitive

This section is devoted to the location of the main top dictionary sense taxonomies for a given semantic primitive in order to correctly attach all these taxonomies to the correct type in the LKB.

In order to illustrate this process we will locate the main top beginners for the FOOD dictionary senses. However, we must consider that many of these top beginners are structured. That is, some of them belong to taxonomies derived from other ones, and then cannot be directly placed within the FOOD type. This is the case of *vino* (*wine*), which is a *zumo* (*juice*). Both are top beginners for FOOD and one is a hyponym of the other.

First, we collect all genus terms from the whole set of DGILE dictionary senses labelled in the previous section with the FOOD tag (2,614 senses), producing a lexicon of 958 different genus terms (only 309, 32%, appear more than once in the FOOD subset of dictionary senses).

As the automatic dictionary sense labelling is not free of errors (around 80% accuracy)[1] we can discard some senses by using filtering criteria.

• **Filter 1** (F1) removes all FOOD genus terms not assigned to the FOOD semantic file during the mapping process between the bilingual dictionary and WordNet.

• **Filter 2** (F2) selects only those genus terms which appear more times as genus terms in the FOOD category. That is, those genus terms which appear more frequently in dictionary definitions belonging to other semantic tags are discarded.

• **Filter 3** (F3) discards those genus terms which appear with a low frequency as genus terms in the FOOD semantic category. That is, infrequent genus terms (given a certain threshold) are removed. Thus, F3>1 means that the filtering criteria have discarded those genus terms appearing in the FOOD subset of dictionary definitions less than twice.

---

[1] *Most of them are not really errors. For instance, all fishes must be ANIMALs, but some of them are edible (that is, FOODs). That is, all fishes labelled as FOOD have been considered mistakes. This is solved in WordNet by placing two different synsets in the noun hierarchy (one as ANIMAL and the other as FOOD), but in DGILE only one dictionary sense appears, representing both.*

Table 5.6 shows the first 36 top beginners for FOOD. That is, those FOOD genus terms which appears 10 or more times. Bold face is used for those genus terms removed from the final list of FOOD genus terms because they appear more times as genus terms in other semantic files. Thus, *pez* (*fish*) is an ANIMAL, *cosa* (*thing*) is an ARTIFACT, *pequeño* (*child*) is an error made during the genus selection, *líquido* (*liquid*) is a SUBSTANCE and *vasija* (*vessel*) is an ARTIFACT. From left to right: number of times the genus appears in the FOOD subset of dictionary senses, Spanish noun and a possible English translation.

90	bebida	drink	28	alimento	food	15	harina	flour
86	vino	wine	27	uva	grape	14	zumo	juice
78	**pez**	fish	25	trigo	wheat	13	sopa	soup
56	comida	food	25	queso	cheese	13	**líquido**	liquid
55	carne	meat	24	guiso	stew	13	grano	grain
48	pasta	(many)	24	**cosa**	thing	12	azúcar	sugar
40	pan	bread	22	**pequeño**	child	11	**vasija**	vessel
39	plato	dish	22	pastel	pie	11	panecillo	small bread
33	guisado	casserole	21	bollo	bun	11	olla	pot
32	salsa	souce	20	manjar	delicacy	11	leche	milk
31	licor	liquor	18	torta	cake	11	embutido	sausage
31	dulce	sweet	16	aguardiente	liquor	10	refresco	refreshment

Table 5.6, main top beginners for FOOD.

Table 5.7 provides the results applying different filtering criteria. In order to select strictly the LKB **c_art_subst** (rather than WordNet FOOD) genus we should consider a more fine-grained labelling.

FOOD	#Genus Terms	Accuracy	#Definitions	Accuracy
LABEL2	958		2,614	
LABEL2+F3>1	309	62%	1,961	77%
LABEL2+F1	409		1,831	
LABEL2+F1+F3>1	203	78%	1,625	86%
LABEL2+F2	439		1,741	
LABEL2+F2+F3>1	187	82%	1,489	**92%**
LABEL2+F1+F2	247		1,482	
LABEL2+F1+F2+F3>1	154	88%	1,389	**95%**

Table 5.7, accuracy of genus terms and definitions with different filtering criteria.

Table 5.8 shows the performance of the second labelling with respect to filter 3 (genus frequency) varying the threshold. From left to right, filter, number of genus terms selected, accuracy, number of definitions and their respective accuracy.

LABEL2	#Genus Terms	Accuracy	#Definitions	Accuracy
LABEL2+F3>9	32	89%	908	88%
LABEL2+F3>8	37	**90%**	953	88%
LABEL2+F3>7	39	88%	969	87%
LABEL2+F3>6	45	88%	1,011	87%
LABEL2+F3>5	51	87%	1,047	82%
LABEL2+F3>4	62	85%	1,102	86%
LABEL2+F3>3	73	78%	1,146	84%
LABEL2+F3>2	99	69%	1,224	80%
LABEL2+F3>1	151	62%	1,328	77%

Table 5.8, performance of second labelling criteria with respect to filter 3.

Table 5.9 shows the performance of the second labelling using filter 1 (bilingual category mismatch) with respect to filter 3 (genus frequency).

LABEL2+F1	#Genus Terms	Accuracy	#Definitions	Accuracy
LABEL2+F1+F3>9	31	**94%**	895	**90%**
LABEL2+F1+F3>8	35	**95%**	931	**90%**
LABEL2+F1+F3>7	37	**91%**	947	89%
LABEL2+F1+F3>6	43	**92%**	989	**90%**
LABEL2+F1+F3>5	49	**92%**	1,025	**90%**
LABEL2+F1+F3>4	55	**91%**	1,055	**90%**
LABEL2+F1+F3>3	64	85%	1,091	88%
LABEL2+F1+F3>2	85	82%	1,152	87%
LABEL2+F1+F3>1	125	78%	1,234	86%

Table 5.9, performance of second labelling criteria with respect to filter 1 varying filter 3.

Table 5.10 shows the performance of the second labelling using filter 2 (genus frequent category mismatch) with respect to filter 3 (genus frequency).

LABEL2+F2	#Genus Terms	Accuracy	#Definitions	Accuracy
LABEL2+F2+F3>9	31	**100%**	893	**100%**
LABEL2+F2+F3>8	35	**100%**	929	**100%**
LABEL2+F2+F3>7	37	**95%**	945	**98%**
LABEL2+F2+F3>6	41	**94%**	973	**98%**
LABEL2+F2+F3>5	47	**92%**	1,009	**97%**
LABEL2+F2+F3>4	56	**91%**	1,054	**96%**
LABEL2+F2+F3>3	65	87%	1,090	**95%**
LABEL2+F2+F3>2	82	83%	1,141	**93%**
LABEL2+F2+F3>1	123	82%	1,223	**92%**

Table 5.10, performance of second labelling criteria with respect to filter 2 varying filter 3.

The above tables show that at the same level of genus frequency, filter 2 (removing genus terms which are more frequent in other semantic categories) is more accurate that filter 3 (removing all genus terms the translation of which cannot be FOOD). For instance, no error appears when selecting those genus terms which appear 10 or more times (F3) in the second labelling process and are more frequent in that category than in any other (F2).

Table 5.11 shows the coverage of correct genus terms selected by criteria F1 and F2 to respect criteria F3. Thus, for genus terms appearing 10 or more times, by using either of the two criteria we are collecting 97% of the correct ones. That is, in both cases the criteria discards less than 3% of correct genus terms.

LABEL2+F3 vs. F1	Coverage	LABEL2+F3 vs. F2	Coverage
LABEL2+F3>9 vs. F1	**97%**	LABEL2+F3>9 vs. F2	**97%**
LABEL2+F3>8 vs. F1	**95%**	LABEL2+F3>8 vs. F2	**95%**
LABEL2+F3>7 vs. F1	**95%**	LABEL2+F3>7 vs. F2	**95%**
LABEL2+F3>6 vs. F1	**96%**	LABEL2+F3>6 vs. F2	**91%**
LABEL2+F3>5 vs. F1	**96%**	LABEL2+F3>5 vs. F2	**92%**
LABEL2+F3>4 vs. F1	89%	LABEL2+F3>4 vs. F2	**90%**
LABEL2+F3>3 vs. F1	90%	LABEL2+F3>3 vs. F2	89%
LABEL2+F3>2 vs. F1	86%	LABEL2+F3>2 vs. F2	83%
LABEL2+F3>1 vs. F1	83%	LABEL2+F3>1 vs. F2	81%

Table 5.11, coverage of second labelling criteria with respect to filter 1 and 2 varying filter 3.

5.2.4 Conclusions

As we have seen in Chapter 3 and Section 4.2, our approach for building LKBs from structured lexical resources is mainly descriptive (the main source of knowledge is MRDs), but a minimal prescribed structure is provided (the LKB Type System). This approach differs from previous ones because of the mixed methodology applied (e.g., the complete descriptive approach of [Bruce et al. 92]). This mixed approach was also followed by [Castellón 93], but rather than selecting the main top beginners for a given semantic category manually, we have presented a complete fully automatic methodology for selecting them. The methodology we propose combines lexical knowledge acquired with minimal supervision[1] from structured lexical knowledge resources, and proceeds roughly as follows:

We labelled automatically the whole noun dictionary twice. The first time computing the conceptual distance between headword and genus of the noun definitions. Assigning WordNet synsets to Spanish headwords, the program classified 31% of the DGILE definitions into 24 different semantic classes (corresponding to the 24 lexicographer's files) with 64% accuracy. The second time, following the method proposed by [Yarowsky 92], we used this preliminary classification to partition DGILE into 24 subcorpora. We used this classification to acquire the salient words for each semantic class the subcorpus was representing. Using these salient words we labelled DGILE again, classifying 93% of all noun definitions with an overall accuracy of 80%. Finally, for each semantic category, after a filtering process we gathered all its representative genus terms. All the genus terms gathered for a semantic category are the main top beginners for the semantic primitive we were looking for.

To bridge the language gap between WordNet and DGILE we used the Esencial Spanish/English bilingual dictionary. Better results could be expected performing the whole process using an English dictionary such as LDOCE.

5.3. Semantic knowledge acquisition from the genus terms in DGILE

This section focuses on the approaches we have considered for the (semi)automatic and automatic construction of taxonomies from DGILE.

In this section we will first present an environment for the (semi)automatic construction of taxonomies. This interactive system allows the user to select the correct hypernym (dictionary sense of the genus) manually (when no automatic one is provided). This environment was used by [Castellón 93] and [Taulé 95] and is described in detail in [Ageno et al. 91b] and [Verdejo et al. 91].

5.3.1 (Semi)automatic construction of taxonomies

The (semi)automatic use of TaxBuild uses a top-down fashion approach for constructing taxonomies. Once the user has selected a top dictionary sense as a root for the semantic hierarchy (rather than by instrospection, the selection of these top dictionary senses can be done using the methodology detailed in Section 4.2), all the dictionary senses containing the top word are retrieved using the LDB indexes on the dictionary definition field. Some of these have the top word as a genus term of the dictionary senses (this means that are candidate hyponyms) while others have the top word in the *differentiae*. In order to select the correct semantic head or genus term for noun and verb definitions and discard those which the top word is not the genus term, we used the specialized grammar designed for this porpuse (see Section 3.6.1.2). Now, the user has to perform the Genus Sense Disambiguation task. That is,

[1] *We have to decide which WordNet synsets (as semantic classes) represent the LKB types (concept classes). In the example we have presented the synsets selected were those within the semantic file 13 (FOOD).*

the user must decide if the candidate hyponym belongs to the taxonomy being constructed. If the user decides that the candidate hyponym does not belong to the taxonomy, the same candidate hyponym is tested. If it belongs a new cicle is started with this candidate hyponym as a top dictionary sense.

The TaxBuild user interface aids the user in all these tasks. However, due to the manual sense disambiguation task of these approach only a few taxonomic fragments has been derived from DGILE. For nouns, we derived lexicons for *substancia* (*substance*, including *food*), *persona* (*person*), *lugar* (*place*) and *instrumento* (*instrument*). Using TaxBuild we constructed (semi)automatically (that is, in a supervised mode, see [Castellón 93]) the complete disambiguated noun taxonomies, taking these words as a starting point containing 3,210 dictionary senses (382 belonging to the FOOD domain).

5.3.2 Automatic construction of taxonomies

This section presents a method of combining a set of unsupervised algorithms in a way they can accurately disambiguate the genus terms of a conventional dictionary without any special encoding. Although most of the techniques for word sense resolution have been presented as stand-alone, it is our belief that full-fledged lexical ambiguity resolution should combine several information sources and techniques. Thus, we present a set of techniques (than we call heuristics) which have been applied in a combined way to disambiguate the genus terms of DGILE, enabling us to construct complete taxonomies for Spanish. Tested accuracy is around 80% overall and 95% for two-way ambiguous genus terms, showing that taxonomy building is not limited to structured dictionaries such as LDOCE.

This work tries to proof that using an appropriate method to combine several heuristics (each one using different information sources in different ways) we can disambiguate the genus terms, and thus construct complete taxonomies from any conventional dictionary in any language with reasonable precision.

5.3.2.1 Test Sampling

In order to test the performance of each heuristic and their combination, we selected a test set at random with 391 noun senses. From this sample, we consider only those with a correctly selected genus (more than 97%). This test set was disambiguated by hand allowing for each genus multiple correct senses (1.37 correct hypernym senses on average, ranging from 1 to 8). Table 5.12 shows the data for the test set:

	DGILE
Test Sampling	391
Correct Genus Selected	382 (98%)
Monosemous	61 (16%)
Senses per genus	5.75
idem (polysemous only)	6.65
Correct senses per genus	1.38
idem (polysemous only)	1.51

Table 5.12. Test set.

5.3.2.2 Measures for testing

As not all the heuristics can always be applied, in order to evaluate the performance of the whole process we provide the following figures: coverage (that is, the percentage of answers), precision (that is, the percentage of actual answers that where correct) and recall (that is, the percentage of possible answers that where correct). That is, precision is calculated from coverage and recall from all test.

5.3.2.3 Derived lexical resources used by the heuristics

Some of the heuristics use elaborated lexical knowledge derived from the dictionary itself or by combining lexical knowledge derived from several structured lexical knowledge sources. This is the case of heuristics from 5 to 8. Heuristic 5 (simple concordance, see Section 5.3.2.4.5) and 6 (coocurrence vectors section 5.3.2.4.6) use coocurrence data gathered from the whole dictionary definitions. Heuristic 7 uses salient word vectors obtained by the process described in section 5.2.2. Heuristic 8 uses both the HBil dictionary and WordNet (see Section 3.2.1).

5.3.2.3.1 Cooccurrence data

Following [Wilks et al. 93] two words cooccur if they appear in the same definition (word order in definitions are not taken into account). For instance, for DGILE, a lexicon of 300,062 cooccurrence pairs among 40,193 word forms was derived (stop words were not taken into account). Table 5.13 shows the first eleven words out of the 360 which cooccur with *vino* (wine) ordered by Association Ratio. Association Ratio between two words (see formula 5.5) can be defined as the product of the Mutual Information [Church & Hanks 90] and the probability of occurring both words in the same definition:

(5.5) $$AR(w_1, w_2) = \Pr(w_1, w_2) MI(w_1, w_2) = \Pr(w_1, w_2) \log_2 \frac{\Pr(w_1, w_2)}{\Pr(w_1)\Pr(w_2)}$$

In table 5.13, from left to right, Association Ratio and number of occurrences.

The lexicon (or machine-tractable dictionary, MTD) thus produced from the dictionary is used by heuristics 5 and 6 (see bellow).

11.1655	15	tinto (red)
10.0162	23	beber (to drink)
9.6627	14	mosto (must)
8.6633	9	jerez (sherry)
8.1051	9	cubas (cask, barrel)
8.0551	16	licor (liquor)
7.2127	17	bebida (drink)
6.9338	12	uva (grape)
6.8436	9	trago (drink, swig)
6.6221	12	sabor (taste)
6.4506	15	pan (bread)

Table 5.13. Example of coocurrences for vino (wine).

5.3.2.3.2 Multilingual data

Heuristics 7 need external knowledge, not present in the dictionaries themselves. This knowledge is composed of semantic field tags and hierarchical structures, and both were extracted from WordNet (see 5.2.2). Attaching WordNet synsets to DGILE words we are also attaching to the Spanish dictionary senses its semantic files. Using this semantic files we can classify 29,205 DGILE noun definitions in 24 partitions (each one corresponding to a semantic category). Using a Mutual information-like formula, we can collect the Spanish salient words for each category. Intuitively, a salient word apperars more often in the context of a semantic category than at at other points in the dictionary. Thus, we derived a lexicon of 23,418 salient word forms. Table 5.14 shows thew first ten Spanish salient words (ordered by salience) for FOOD category. From Left to right, association ratio, salient word, number of occurrences in FOOD context and finally, number of ocurrences in the whole dictionary.

AR	Word	# in FOOD	# in DGILE
7.5984	bebida (drink)	40	58
6.6362	carne (meat)	44	104
6.5707	dulce (sweet)	29	50
6.3316	azúcar (sugar)	29	56
6.2459	comida (food)	32	70
6.0055	harina (flour)	21	35
5.9917	fruto (fruit)	49	163
5.8268	leche (milk)	23	46
5.5468	salsa (souce)	13	16
5.1547	zumo (juice)	12	17

Table 5.14, ten salient words for FOOD.

As one word can be a salient word for many semantic categories, for each word we obtain a 24 dimensional weighted vector. Thus, table 5.15 shows the salient categories for *iglesia* (church) and *iglesias* (churches). AR is obtained from formula 5.3.

	Semantic File	AR
iglesia (church)	04 act	0.7371
	11 event	0.4229
	14 group	2.5772
	18 person	0.3923
	21 possession	0.3445
	28 time	2.8352
iglesias (churches)	06 artifact	0.8455
	18 person	0.7387
	28 time	1.3365

Table 5.15, semantic categories for salient words iglesia and iglesias.

Heuristic 8 also needs external knowledge. In this case, this knowledge is provided by means of the bilingual dictionary. Firstly, each Spanish word has looked up in the bilingual dictionary, and its English translation are found. For each translation WordNet yielded its senses, in the form of WordNet concepts (synsets). The pair made of the original word and each of the concepts linked to it, was included in a file, thus producing a MTD with links between Spanish words and WordNet concepts. Obviously some of this links are not correct, as the translation in the bilingual dictionary would not mean all of the senses in WordNet.

For instance when accesing the semantic fields for *vino* we get a unique translation, wine, which has two senses in WordNet <wine,vino> as a beverage, and <wine, wine-coloured> as a kind of colour. In this example two links would be produced (*vin*, <wine,vino> and (*vin*, <wine, wine-coloured>). This link allow as to get two possible semantic fields for *vino* (noun.food, file 13 and noun.attribute, file 7) and the structure of the hierarchy in WordNet for each of the concepts (see the figures of this mapping in table 6.3).

5.3.2.4 Heuristics for Genus Sense Disambiguation

As the methods described in this work have been developed for being applied in a combined way, each one must be seen as a container of some part of the knowledge (or heuristic) needed to disambiguate the correct hypernym sense. Not all the heuristics are suitable to be applied to all the definitions. Each heuristic assigns each candidate hypernym sense a weight, i.e. a real number value ranging from 0 to 1 (after a scaling process, where maximum score is scaled to 1). The heuristics applied range from the simplest (e.g. heuristic 1, 2, 3 and 4) to the more informed ones (e.g. heuristics 5, 6, 7 and 8), and use information present in the entries under study (e.g. heuristics 1, 2, 3 and 4) or extracted from the whole dictionary

as a unique lexical knowledge resource (e.g. heuristics 5 and 6) or combining lexical knowledge from several heterogeneous lexical resources (e.g. heuristic 7 and 8).

5.3.2.4.1 Heuristic 1: Monosemous genus term

This heuristic is applied when the genus term is monosemous. As there is only one hypernym sense candidate, the hyponym sense is attached to it. Only 12% of noun dictionary senses have monosemous genus terms in DGILE.

5.3.2.4.2 Heuristic 2: Entry Sense ordering

This heuristic assumes that senses are ordered in an entry by frequency of usage. That is, the most used and important senses are placed in the entry before less frequent or less important ones. This heuristic provides the maximum score to the first sense of the hypernym candidates and decreasing scores to the other.

5.3.2.4.3 Heuristic 3: Explicit Semantic domain

This heuristic assigns the maximum score to the hypernym sense which has the same semantic domain tag as the hyponym. This heuristic is of limited application: Less than 10% of the definitions are marked in DGILE with one of the 96 different semantic domain tags (e.g. med. for medicine or der. for law).

5.3.2.4.4 Heuristic 4: Word Matching

Following [Lesk 86], this heuristic trusts that related concepts will be expressed using the same content words. Given two definitions, an hyponym and one candidate hypernym, this heuristic computes the total amount of content words shared (including headwords). Because of the morphological productivity of Spanish, we have considered different variants of this heuristic. Finally, this heuristic yields better results when matching the first four characters of words.

5.3.2.4.5 Heuristic 5: Simple Concordance

This heuristic uses cooccurrence data collected from the whole dictionary (see section 5.3.2.3.1). Thus, given a hyponym definition (O) and a set of candidate hypernym definitions, this method selects the candidate hypernym (E) which returns the maximum score given by formula (5.6):

(5.6)
$$SC(O,E) = \sum_{i \in O \wedge j \in E} cw(w_i, w_j)$$

The cooccurrence weight (*cw*) between two words can be given by Cooccurrence Frequency, Mutual Information [Church & Hanks 90] or Association Ratio [Resnik 92]. We tested them using different context window sizes. Best results were obtained using the Association Ratio and a window size of 7 words.

5.3.2.4.6 Heuristic 6: Cooccurrence Vectors

This heuristic is based on the method presented in [Wilks et al. 93] which also uses cooccurrence data collected from the whole dictionary (c.f. section 4.1). Given a hyponym definition (O) and a set of candidate hypernym definitions, this method selects the candidate hypernym (E) which returns the maximum score following formula (5.7)

(5.7) $$CV(O,E) = sym(v_O, v_E)$$

The similarity (*sym*) between two definitions can be measured by the dot product, the cosine function or the Euclidean distance between two vectors (v_O and v_e) which represent the contexts of the words presented in the respective definitions following formula (5.8).

(5.8) $$V_{Def} = \sum_{w_i \in Def} ci(w_i)$$

The vector for a definition (*VDef*) is computed adding the cooccurrence information vectors of the words in the definition (*ci(wi)*). The cooccurrence information vector for a word is collected from the whole dictionary using Cooccurrence Frequency, Mutual Information or Association Ratio. The best combination was the dot product, Association Ratio, and window size 7.

5.3.2.4.7 Heuristic 7: Semantic Vectors

Because DGILE is poorly semantically coded we decided to enrich the dictionary assigning automatically a semantic tag to each dictionary sense (see Section 5.2 for more details). Instead of assigning only one tag we can attach to each dictionary sense a vector with weights for each of the 24 possible semantic tags we considered (which correspond to the 24 lexicographers or semantic files of WordNet [Miller 90]). In this case, given an hyponym (O) and a set of possible hypernyms we select the candidate hypernym (E) which yields maximum similarity among semantic vectors:

(5.9) $$SV(O,E) = sym(v_O, v_E)$$

where *sym* can be the dot product, cosine or Euclidean Distance, as above. Each dictionary sense has been semantically tagged with a vector of semantic weights following formula (5.10).

(5.10) $$V_{Def} = \sum_{w_i \in Def} sw(w_i)$$

The salient word vector (sw) for a word contains a saliency weight [Yarowsky 92] for each of the 25 semantic tags. Again, the best method was the dot product, Association Ratio, and window size 7.

5.3.2.4.8 Heuristic 8: Conceptual Distance

Conceptual distance provides a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is essentially the length of the shortest path that connects the concepts in a hierarchical semantic net. In order to apply conceptual distance, WordNet was chosen as the hierarchical knowledge base, and bilingual dictionaries were used to link Spanish and French words to the English concepts.

Given a hyponym definition (O) and a set of candidate hypernym definitions, this heuristic chooses the hypernym definition (E) which is closest according to formula 5.11 (see also Section 5.2.2.1):

(5.11) $$CD(O,E) = dist(headword_O, genus_E)$$

That is, Conceptual Distance is measured between the headword of the hyponym definition and the genus of the candidate hypernym definitions using formula (5.12), c.f. [Agirre et al. 94]. To compute the distance between two words (*w1*, *w2*), all the corresponding concepts in WordNet (*c1*, *c2*) are searched via a bilingual dictionary, and the minimum of the summatory for each concept in the path between each possible combination of *c1* and *c2* is returned, as shown below:

(5.12)
$$dist(w_1, w_2) = \min_{\substack{c_{1_i} \in w_1 \\ c_{2_i} \in w_2}} \sum_{c_k \in path(c_{1_i}, c_{2_i})} \frac{1}{depth(c_k)}$$

Formulas (5.11) and (5.12) proved the most suitable of several other possibilities for this task, including those which included full definitions in (5.11) or those using other Conceptual Distance formulas [Agirre & Rigau 96].

5.3.2.4.9 Combining Results

The way we have chosen to combine all the heuristics in one simple decision is simple. The weigths each heuristic assigns to the rivaling senses of one genus are normalized to the interval between 1 (best weight) and 0. Formula 5.13 shows the normalized value a given heuristic will give to sense E of the genus, according to the weight assigned to the heuristic to sense E and the maximum weight of all the senses of the genus E_i.

(5.13)
$$vote(O, E) = \frac{weight(O, E)}{\max_{E_i}(weight(O, E_i))}$$

The values thus collected from each heuristic, are added up for each competing sense. The order in which the heuristics are applied has no relevance.

Table 5.14 summarizes the results for polysemous senses and gives also the overall results (which include monosemous genus). In general, the results obtained for each heuristic seem to be poor, but always over the random choice baseline (also shown in table 5.14). The best heuristic is the sense ordering heuristic (2). But, scaling each heuristic and adding the resulting weights (Sum) we obtained an improvement over sense ordering (heuristic 2) of 9% (from 70% to 79%) maintaining a coverage of 100%. Overall (c.f. table 5.16), the sum is able to correctly disambiguate 83% of the genus (8% improvement over sense ordering). Note that we are adding the results of eight different heuristics with eight different performances, improving the individual performance of each one (recall increase 8%).

Polysemous	random	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	**Sum**
recall	30%	-	70%	1%	44%	57%	60%	57%	47%	79%
precision	30%	-	**70%**	100%	72%	57%	60%	58%	49%	**79%**
coverage	100%	-	100%	1%	61%	100%	100%	99%	95%	100%
Overall										
recall	41%	16%	75%	2%	41%	59%	63%	59%	48%	83%
precision	41%	100%	75%	100%	79%	65%	66%	63%	57%	**83%**
coverage	100%	16%	100%	2%	56%	95%	97%	94%	89%	100%

Table 5.16. Overall results.

In order to test the contribution of each heuristic to the total knowledge, we tested the sum of all the heuristics, eliminating one of them each time. The results are provided in table 5.17.

	Sum	-(1)	-(2)	-(3)	-(4)	-(5)	-(6)	-(7)	-(8)
recall	83%	79%	72%	81%	81%	81%	81%	81%	77%
precision	**83%**	79%	**72%**	82%	81%	81%	81%	81%	77%
coverage	100%	100%	100%	98%	100%	100%	100%	100%	100%

Table 5.17. Knowledge provided by each heuristic.

[Gale et al. 93] estimate that any sense-identification system that does not give the correct sense of polysemous words more than 75% of the time would not be worth serious consideration. While heuristic 8 has the worst performance (see table 5.16, precision 57%), it has the second larger contribution (see table 5.17, precision decreases from 83% to 77%). That is, even those heuristics with poor performance can contribute with knowledge that other heuristics do not provide.

The results show that the combination of heuristics is useful, even if the performance of some of the heuristics is low. The combination performs better than isolated heuristics, and allows to disambiguate all the genus of the test set with a success rate of 83%.

Selecting the correct sense for LDOCE genus terms, (Bruce et al. 92) report a success rate of 80% (90% after hand coding of ten genus). This impressive rate is achieved using the intrinsic characteristics of LDOCE. Furthermore, using only the implicit information contained into the dictionary definitions of LDOCE [Cowie et al. 92] report a success rate of 47% at a sense level. [Wilks et al. 93] reports a success rate of 45% disambiguating the word bank (thirteen senses LDOCE) using a similar technique than to heuristic 6. In our case, combining informed heuristics and without explicit semantic tags, the success rates are 83% overall, and 95% for two-way ambiguous genus. Furthermore, 93% of times the real solution is between the first and second proposed solution.

The results show that the construction of taxonomies using lexical resources is not limited to highly-structured dictionaries as LDOCE, but can be applied to two very different dictionaries. In fact, this method has been also tested with a complete different dictionary: the French dictionary *Le Plus Petit Larousse* (LPPL), with similar results (see [Rigau et al. 97]). Nevertheless, quality and size of the lexical knowledge resources are important. As the results for LPPL show, small dictionaries with short definitions can not profit take from raw corpus techniques (heuristics 5, 6), and consequently the overall precision is lower.

We have also shown that summing is a useful way to combine knowledge from several unsupervised WSD methods, allowing to raise the performance of each one in isolation (coverage and/or precision). While it may appear that more intelligent ways to ensamble different heuristics should do better, the experience in the forecasting literature has been that simple, unweighted voting is very robust [Dietterich 97]. Furthermore, even those heuristics with apparently poor results (e.g. see DGILE heuristic 8 in table 5.17) provides knowledge to the final result not provided by the rest of heuristics. Thus, adding new heuristics with different methodologies and different knowledge (e.g. from corpora) better results can be expected. Although we used these techniques for GSD problem we expect similar results (or even better taken the "one sense per discourse" property [Gale et al. 92] and lexical knowledge acquired from Corpora) for the WSD problem.

5.3.2.5 Building automatically large scale taxonomies from DGILE

Our proposal for building taxonomies from MRDs follows the next procedure. Once the main top beginners (relevant genus terms) of a semantic category are selected (see Section 5.2) and every dictionary definition has been disambiguated (this section), we collect all those pairs labelled with the semantic category (using LABEL2) we are working on having one of the genus terms selected (and disambiguated). Using these pairs we finally build up the complete taxonomy for a given semantic category. That is, in order to build the complete taxonomy for a semantic primitive we fit the lower senses using the second labelled lexicon and the genus sense selected using the technique described in this section.

Table 5.18 summarizes the sizes of the FOOD taxonomies acquired from DGILE with respect to filtering criteria and the results manually obtained by [Castellón 93][1].

FOOD	[Castellón 93]	LABEL2+F2+F3>9	LABEL2+F2+F3>4
Genus terms	2	33	68
Dictionary senses	392	952	1,242
Levels	6	5	6
Senses in level 1	2	18	48
Senses in level 2	67	490	604
Senses in level 3	88	379	452
Senses in level 4	67	44	65
Senses in level 5	87	21	60
Senses in level 6	6	0	13

Table 5.18, comparison of FOOD taxonomies.

Using the first set of criteria (LABEL2+F2+F3>9), we acquire a FOOD taxonomy with 952 senses (more than two times larger than if it is done manually). Using the second one (LABEL2+F2+F3>4), we obtain another taxonomy with 1,242 (more than three times larger). While using the first set of criteria, the 33 genus terms selected produce a taxonomic structure with only 18 top beginners, the second set, with 68 possible genus terms, produces another taxonomy with 48 top beginners. However, both final taxonomic structures produce more flat taxonomies than if the task is done manually. This is because we are restricting the inner taxonomic genus terms to those selected by the criteria (33 and 68 respectively). Consider the following taxonomic chain, obtained using SEISD in a semiautomatic way by [Castellón 93]:

bebida_1_3 <- **líquido**_1_6 <- zumo_1_1 <- vino_1_1 <- rueda_1_1

As *líquido* (*liquid*) was not selected as a possible genus (by the criteria described above), the taxonomic chain for that sense is:

zumo_1_1 <- vino_1_1 <- rueda_1_1

Thus, a few arrangements (18 or 48 depending on the criteria selected) must be done at the top level of the automatic taxonomies. Studying the main top beginners we can easily discover an internal structure between them. For instance, placing all *zumo* (*juice*) senses as hyponym of *bebida* (*drink*) (see the complete taxonomy for wines in the appendix).

As these taxonomies can be constructed using a filtering process on the genus terms, different sizes of taxonomies can be produced depending on the degree of accuracy we apply. For instance, with accuracy near 100% (with filter LABEL2+F2+F3>9) on genus terms selected we produce a noun taxonomy of 35,099 definitions. If we reduce the level of accuracy to 96% (with filter LABEL2+F2+F3>4), we obtain a taxonomy structure of 40,754 senses.

The results show that the construction of taxonomies using lexical resources is not limited to highly structured dictionaries. Applying appropriate techniques, monolingual dictionaries such as DGILE could be useful resources for automatically building accurate substantial pieces of an LKB.

[1] *We used the results reported by [Castellón 93] as a baseline during the acquisition process because her work was done using the earlier (semi)automatic version of SEISD and the same Spanish dictionary.*

5.4. Semantic knowledge acquisition from the differentiae in DGILE

5.4.1 Analysing Definitions.

SegWord [Sanfilippo 90] and FPar [Carroll 90b] were improved and tailored to analyze Spanish definitions (see Section 3.6.1.1). Neither of them is a complete analyser for Spanish. SegWord covers nouns, adjectives, adverbs, closed class words and some forms of verbs and FPar grammars have been designed to cover subsets of definitions semantically related. Obviously, using a wide range parsing tool for Spanish, rather than a partial one, better results can be obtained.

Currently, we are using a broad range morphological analyser of Spanish [Acebo et al. 94] (which also provides the possible Spanish lemmas for a given word form) and a tagger of Spanish [Padró 98] and a shallow DCG grammar[1] [Pereira & Warren 80] for parsing completely all dictionary definitions. Perhaps an in-depth grammar/parser of Spanish could lead to better results, but building such a tool is beyond the scope of this research, and given the kind of material to be parsed (because of the sublanguage used in dictionaries) and the acquisition goals, partial coverage does not seem to be a serious limitation. Thus, rather than analyse small parts of the dictionary definitions (i.e., [Alshawi 89], [Artola 93], [Castellón 93]) we propose (when no full parse can be performed with high accuracy) complete analysis of the dictionary definition using a shallow parser which provides a fully analysed set of chunks for an input definition.

A typical chunk consists of a single content word (or head) surrounded by a constellation of function words. A simple context-free grammar is quite adequate to describe the structure of the chunks. The idea is to factor the parse into pieces of structure that can be reliably recovered with a small amount of syntactic information, as opposed to those pieces of structure that require much larger quantities of information. By reducing the definition to chunks, there are fewer units whose associations must be considered. Resolving attachments generally requires information about lexical association between heads, hence it is postponed. But rather than select the smaller chunks from multiple options, the longest match must be selected.

Consider the example of a drink dictionary sense:

acapulco_1_1 ISA **cóctel** hecho con tequila, ron y zumo de piña. (acapulco, cocktail made with tequila, rum and pineapple juice).

After the tagging process we obtain for acapulco_1_1:

acapulco_1_1 ISA	cóctel	cóctel	N
	hecho	hacer	U0V
	con		R0P
	tequila	tequila	N
	,		Z0C
	ron	ron	N
	y		C0C
	zumo	zumo	N
	de		R0D
	piña	piña	N
	.		Z0P

[1] *Implemented in Prolog with 134 grammar rules.*

where N stands for noun, U0V for verb, R0P and R0D for prepositions and Z0C and Z0P for semicolon and dot respectively.

Thus, we are able to analyse the main sintagms of the definition:

```
acapulco_1_1 ISA  sn:[n:cóctel]
                  sv:[u0v:hacer]
                  sp:[r0p:con,sn:[n:tequila,n:ron,n:zumo,sp:[r0d:de,sn:[n:piña]]]]
                  sw:[z0p: .]
```

where SN stands for nominal syntagm, SV for verbal and SP for prepositional. All words no grouped in syntagms are left and labeled as alone: SW. In that sense, this parser is able to analyse completely all dictionary definitions.

Furthermore, selecting those relevant patterns restricted by the corpus which correspond to a semantic category (from LABEL2 lexicon, see section 5.2) we can perform an in depth semantic analysis of the definitions. For instance, the next pattern

```
hecho con SN      made with SN
```

in FOOD domain usually is a cue for detecting constituents of the headword defined.

```
acapulco_1_1 ISA  sn:[n:cóctel]
                  made_with():[n:tequila,n:ron,n:zumo,sp:[r0d:de,sn:[n:piña]]]
                  sw:[z0p: .]
```

In order to compare the performance of the current shallow parsing process, Table 5.19 summarizes the total number of semantic relations detected[1].

FOOD	[Castellón 93]	LABEL2+F2+F3>9	LABEL2+F2+F3>4
definitions	392	952	1,242
properties	137	717	825
pp-mod	197	1,118	1,310
goal	15	17	19
composed-by	44	82	96
simil	2	19	23
purpose	5	18	25
color	2	10	12
temp	5	14	14
origin	0	18	20
total	407	2,013	2,344
total syntagms[2]	883	2,760	3,270

Table 5.19, comparison of total number of semantic relations detected.

That is, while for some patterns the performance does not seem to improve (e.g., goal), on average we are doubling the total number of patterns acquired.

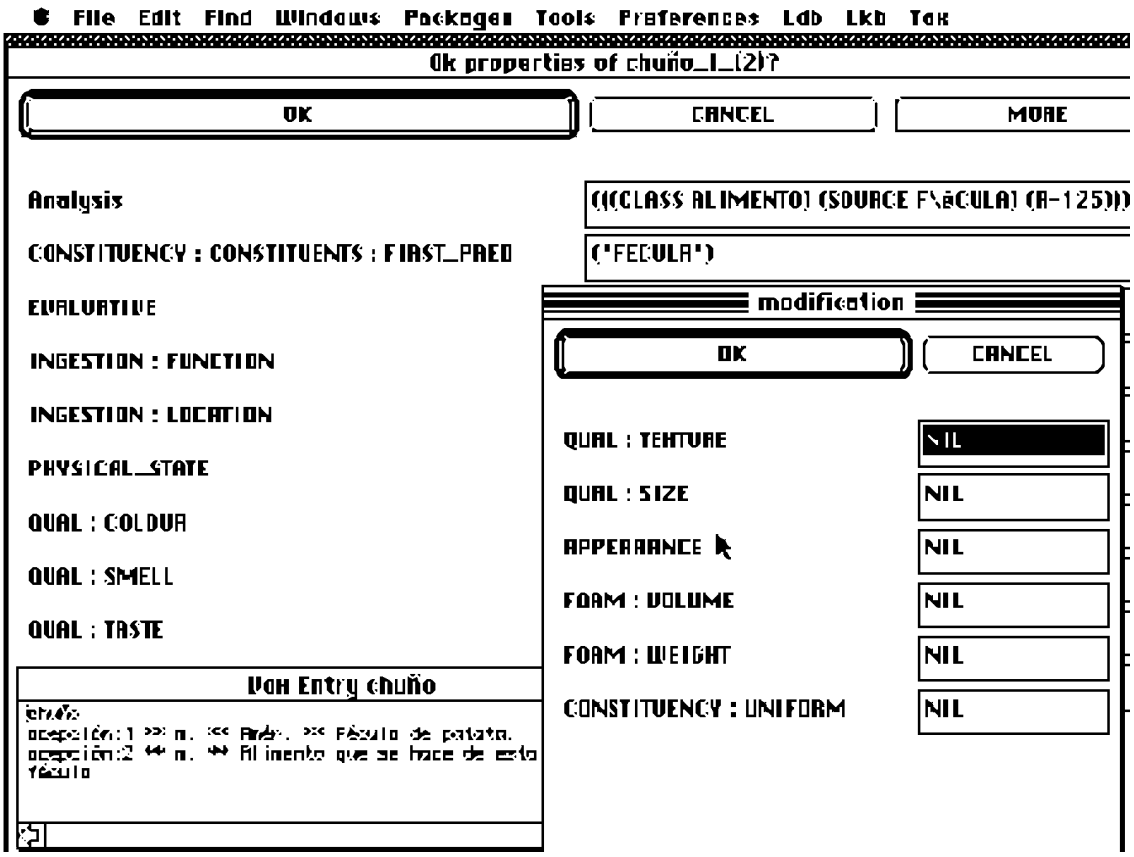
5.4.2 Placing definitions into the LKB.

Once we have analysed each definition, using the CRS (the Convertion Ruleset System, see [Ageno et al. 92c]) this knowledge can be placed within the LKB. That is, each definition is converted to a lexical entry that can be loaded into the LKB. We implemented this system using PRE language to provide flexibility and declarativity to the process. Thus, the system

[1] *Patterns containing more than one word are counted once.*

[2] *Syntagms with a unique word were not counted.*

allows the user to perform the mapping process (or convertion) in a (semi)automatic (the user can check or change data during the process) or fully automatic way. Screen 5.1 shows a (semi)automatic session when converting the lexical entry for *fécula* startch.



Screen 5.1, a CRS session.

Consider again the analysed definition considered previously. As the only semantic tag shared (using the non-disambiguated mapping between WordNet and the bilingual or those taxonomies derived from DGILE) by the three constituents *tequila*, *ron*, and *zumo* is FOOD, we can automatically derive that an *acapulco* is a cocktail made with the following INGREDIENTS: tequila, rum and pineapple juice.

```
acapulco_1_1 ISA   sn:[n:cóctel]
                   made_with(FOOD):[n:tequila,n:ron,n:zumo,sp:[r0d:de,sn:[n:piña]]]
                   sw:[z0p: .]
```

Furthermore, computing the conceptual distance among the translations of these four words (no traslation of acapulco is provided in the bilingual dictionary) a fully complete lexical disambiguation could be performed. In that way, we are planning to use the Top Ontology derived from EuroWordNet project that connects those Base Concepts (about 700 nominal synsets of WordNet1.5, see [Rodríguez et al. in Press] for further details) in order to perform a new mapping process between the analysed definitions and the types of the LKB. That is, gathering those concept patterns that coocurs in dictionary definitions in order to fulfil the LKB.

To date, when applying the CRS (that is, the mapping process from the analysed definitions to the LKB) only one type of the LKB was under consideration. It is our believe that larger and more accurate knowledge could be acquired if more types (or concept patterns) were considered. That is, instead of considering only a unique type (that can be seen as a star that connects concepts) we are planning to consider the complete constelation (that is, the complete LKB).

5.5. Conclusions

This chapter has been devoted to the feasibility of the automatic and productive acquisition of lexical knowledge from monolingual MRDs in order to construct a highly structured LKB.

Instead of following a purely descriptive approach we have adopted a minimal prescribed set of semantic categories (types in the LKB). Section 5.2 and 5.3 focused on the acquisition of lexical knowledge from the genus terms of the dictionary. Section 5.2 presented a novel methodology to classify by a double labelling process all the dictionary definitions with one of the semantic categories provided. Section 5.3 presented a robust lexical knowledge technique for disambiguating all noun genus terms (coverage 100%) with an overall precision of 83%. Then, after a filtering process, we gathered those relevant genus terms for a given semantic category which enables us to built up its taxonomies. Thus, we have presented a complete fully automatic methodology for acquiring large-scale taxonomies from non-semantically coded MRDs.

Section 5.4 was centered on lexical knowledge acquisition from the differentia. We presented a novel methodology which uses a wide-range morphological tagger with a combination of domain-oriented shallow parser to acquire larger and more accurate knowledge from dictionary definitions. Using the semantic knowledge previously acquired, these analysis can be processed in order to build a large and richly structured LKB.

Chapter 6

Multilingual Lexical Knowledge Acquisition

6.1 Introduction

The purpose of this chapter is to present the work carried out for the (semi)automatic and automatic construction of the multilingual facet of the LKB (see Section 4.5). In this Chapter, we present several methodologies for extracting lexical translation equivalencies from conventional monolingual and bilingual MRDs. We also describe a series of experiments on the construction of a MLKB for English and Spanish and present TGE, the software module system we developed within SEISD to perform this task (see Section 3.6.3).

While Section 2 presents the complete framework and resources used by TGE for linking lexical entries across languages, Section 3, 4 and 5 show the main experiments we carried out and the results obtained linking Spanish lexical units to English ones. Section 3 reports the main techniques and results doing such process using TGE in a (semi)automatic approach attaching Spanish sense taxonomies (derived from DGILE) to English ones (derived from LDOCE). Using the same framework, that is TGE, Section 4 presents a fully automatic approach (by means of the Conceptual Distance formula) to link Spanish sense taxonomies (derived from DGILE) to WordNet. Finally, Section 5, presents another novel approach to build automatically from bilingual dictionaries a MLKB for Spanish using the predefined structure of WordNet.

6.2 Multilingual Lexical Knowledge Acquisition

6.2.1 Introduction

This Section presents the complete framework and resources for linking lexical entries derived from MRDs across languages. Of course, these links can be established manually (assuming a severe time-consuming drawback), but the multiplicity of cases occurring, the existence of several heterogeneous knowledge sources (such as bilingual dictionaries, monolingual LDBs, monolingual taxonomies and multilingual LKBs) motivates the mechanization of the whole process. In order to perform this task, we designed and built a SEISD module, the Tlink Generation Environment (**TGE**) [Ageno et al. 94]. As this environment does not impose a single methodological strategy, we have been able to implement different approaches. While the first, described in Section 3, performs the bilingual connexion using a (semi)automatic approach, the second, described in Section 4, builds the MLKB using a fully automatic approach. Both methods take profit from the taxonomy structure of both languages involved, which have been derived previously from monolingual MRDs. Although it seems to be generally agreed that bilingual MRDs alone are insufficient for constructing MLKBs, in Section 5 we presents the main experiments we have performed attaching directly Spanish words derived from a bilingual dictionary to a predefined LKB (in this case, WordNet). In that way, we are able to build a paralel in structure MLKB without using knowledge derived from monolingual MRDs.

While in the approaches presented in Section 6.3 and 6.4, the basic units for defining lexical translation equivalence are the lexical entries in the monolingual LKBs, which should, in general, correspond to word senses in the dictionary, in Section 6.5 the source basic units are words and the target ones, senses (in this case, synsets of WordNet).

6.2.2 Translation Tlinks

Although in the simplest cases we can consider the lexical entries represented in the LKB themselves as translation equivalent, in general, more complex cases occur corresponding to lexical gaps, differences in morphologic or lexical features, specificity, etc. These complex relations can be expressed by means of the **tlink** (for translation link) mechanism of the LKB [Copestake 92b] (see also Section 3.6.3). We represent the relationships between words senses in terms of tlinks. The tlink mechanism is general enough to allow the monolingual information to be augmented with translation specific information, in a variety of ways.

LKB formalism uses a typed feature structure (FS) system for representing lexical knowledge. We can, so, define tlinks in terms of relations between FSs. Lexical (or phrasal) transformations in both source and target languages are a desirable capability so that we can state that a **tlink** is essentially a relationship between two rules (of the sort already defined in the LKB) where the rule inputs have been instantiated by the representations of the word senses to be linked.

As any other LKB object, a tlink can be represented as a feature structure. The type system mechanism, in LKB, allows further refinement and differentiation of tlink classes in several ways as is shown in figure 6.2.

A **simple-tlink** is applicable whenever two lexical entries which denote single place predicates (nouns, etc.) are straightforwardly translation equivalent, without any previous transformation. The example presented in figure 6.1 belongs to this class. As shown in Fig. 6.1, *furniture* can be encoded as translation equivalent to the plural *muebles* by specifying that the named rule *plural* has to be applied to the base sense in Spanish.

A **partial tlink** is applicable when we want to transfer the qualia structure from one sense to another. An example of this class is the Spanish entry *rioja*. There is no direct correspondence between this word and any English one because of the absence of such entry in the bilingual dictionary. We can however link the genus term, *vino*, to the corresponding English term *wine*, transferring to the later all the qualia structure from the former (and specially the *origin_area = Rioja*). In this way *rioja* can be roughly translated to English as a *wine with origin_area = Rioja*.

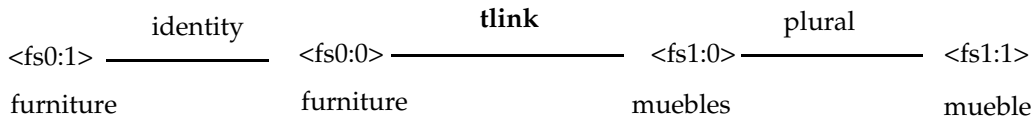


Figure 6.1. A tlink between "furniture" and "muebles".

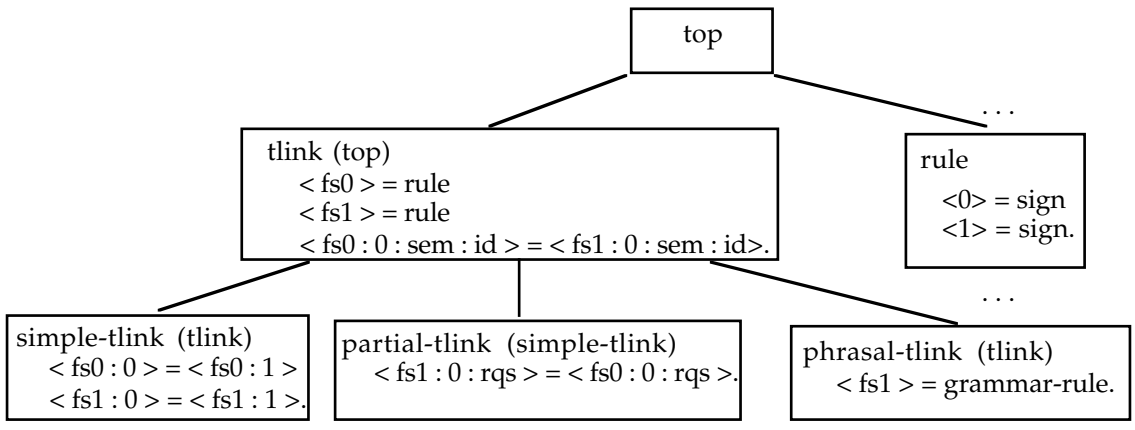


Figure 6.2, partial view of tlink type hierarchy.

Finally, the **phrasal tlink** is necessary when we need to describe a single translation equivalence with a phrase. *Ahumado*, for instance, must be linked to *smoked food*.

6.2.3 Multilingual Lexical Resources

Several MTDs have also been obtained from the two sides of the bilingual dictionary used in this thesis. The lispified version of both directions were loaded into the LDB [Hastings et al. 94]. Briefly, the Spanish/English dictionary **EEI** contains 16,463 entries and 28,002 translation fields, while the English/Spanish **EIE** contains 15,352 entries with 27,033 translation fields. Obviously, as this bilingual dictionary is smaller than the monolingual one no complete connections could be produced. See table 6.3 to complete this fugures.

6.2.4 Linking lexical entries across Languages

As we said before, **TGE** is a tool designed for supporting a tlink extraction methodology. The core of the methodology is the use of a bilingual dictionary as a main knowledge source. Depending on the characteristics of the dictionary entry (or on its absence) different kinds of tlinks with different degree of fitness can be produced. An important consideration is that in spite of using a bilingual dictionary as knowledge source what we are linking are not words but lexical entries placed in the LKB (that is, all the information we gathered from a dictionary definition) and owning not only orthographic information but also lexical information, basically the qualia structure, both local and inherited (because lexical entries are structured in taxonomic structures).

The way of organising the extraction process is by means of the performance of a set of extraction modules, each one corresponding to a different kind of tlink, implemented as rulesets in a Production Rules Environment (PRE, see [Ageno et al. 94]).

The TGE mapping program creates tlinks accessing three knowledge sources, namely, the bilingual dictionary, the source LKB lexicon (including its taxonomic relations) and the target LKB lexicon.

Whatever the approach used several decisions must be taken: The kind of control we need, the rulesets to be designed, the rules belonging to each ruleset, the relative priority assigned to each rule, and so on.

In the experiments reported in this Chapter, the mapping process repeats the application of TGE rulesets for each entry in the list of lexical entries the user wants to translate. The process performed over each entry consists in applying specialised modules in strict order. Depending on the execution mode selected by the user, the resulting tlinks can correspond either to the first successful module or to a selection of results from all of them.

An initial set of modules has been designed according to the typology of tlinks presented in section 6.2.2. It included four classes of tlinks that showed distinct conceptual correspondences between both languages. A more in-depth study of English-Spanish mismatches (see [Soler-93]) could lead to an enrichment of the typology, and consequently, to a need for extending the extant modules.

Up to now, six modules has been developed each of them implemented as a ruleset. Each of them generates one of the four kinds of tlinks. Each module follows a different strategy to guess a possible tlink, looking at the three accessible knowledge sources. The modules built up to now in TGE are described below:

- **Simple Tlink Module**

This is the case when there is a direct translation of the source entry in the bilingual dictionary.

Algorithm: The source orthography is looked up in the bilingual. If any translation corresponds to lexical entries in the target lexicon, a SIMPLE-TLINK is generated.

Example:

```
absenta -----> absinth                          bilingual dictionary
absinth -----> ABSINTH_L_0_1                    LKB entry
===>
ABSENTA_X_I_1 / ABSINTH_L_0_1 :
SIMPLE-TLINK.
```

"absenta" is translated in the bilingual by "absinth", ABSINTH_L_0_1 is a valid lexical entry of the target lexicon, and therefore a SIMPLE-TLINK connecting both entries is created.

• Compound Tlink Module

This is the case when the corresponding entry in the target lexicon is a composed one, being the target lexical entry made up of the concatenation of the two English words that appear in the bilingual entry.

Algorithm: the source orthography is looked up in the bilingual dictionary. We concatenate with an underline those translations composed by two words and then we test if an existing lexical entry in the target lexicon exist. Then a SIMPLE-TLINK connecting the source entry and the target entry (the composed one) is produced.

Example:

```
pelel    -----> pale ale                        Bilingual
pale_ale -----> PALE_ALE_L_0_0                  LKB entry
===>
PELEL_X_I_1 / PALE_ALE_L_0_0 :
SIMPLE-TLINK.
```

"pelel" is translated by "pale ale", these two words don't exist in the english lexicon but their concatenation does. The corresponding SIMPLE-TLINK is produced.

• Phrasal VerbTlink Module

This is the case of source words translated by the bilingual by means of phrasal words reduced to a head form with literal modifiers (particles, prepositions, ..). The tlink produced must connect the source entry and the target phrasal word. We have included some of these modifier signs like psorts in a special lexicon.

Algorithm: if the translation includes any English phrase composed with a verbal sign and a following particle, existing the first in the target verbal lexicon, and the second in the particle lexicon, then a PHRASAL-VERB-TLINK is produced.

Example:

```
aupar   -----> lift up                          Bilingual dictionary
lift    -----> LIFT_L_1_1 (verbal_sign)         LKB entry
up -----> UP_1 (particle)                       LKB entry
===>
AUPAR_X_I_1 / LIFT_L_1_1+UP_1 :
PHRASAL-VERB-TLINK.
```

"aupar" appears in the bilingual as "lift up". The first word is a verbal sign that appears in the verbal lexicon as LIFT_L_1_1. The second word has a corresponding entry in the particles lexicon: UP_1. A PHRASAL-VERB-TLINK is finally proposed.

• General Phrasal Tlink Module

This is the case when the translation appearing in the bilingual is composed of more than one word. Normally these explanations are made up as definitions, being composed of a genus and some modifiers. A tlink connecting the source entry and the genus appearing in the definition must be created.

Algorithm: if the translation definition includes any word that exists in the target lexicon by its own, it is considered the genus and consequently a PHRASAL-TLINK is proposed.

Example:

```
amontillado -----> pale dry sherry                Bilingual dictionary
pale  -----> nil
dry  -----> nil
sherry  -----> SHERRY_L_0_0                       LKB entry
===>
AMONTILLADO_X_I_1 / SHERRY_L_0_0 :
PHRASAL-TLINK.
```

The translation of "amontillado" is "pale dry sherry", "pale" and "dry" aren't members of the target lexicon, but "sherry" is. Therefore, a PHRASAL-TLINK between both single entries is produced.

• Parent Tlink Module

This is the case of very specific terms in the source lexicon, they are not treated in the bilingual dictionary, but their hypernyms in the taxonomy have a clear translation that can generate a partial tlink.

Algorithm: if the source entry has an ancestor in the loaded lexicons, it is looked up in the bilingual in order to get its translations. If it exists in the target lexicon, a PARTIAL-TLINK is produced, connecting the original source entry and the parent translation.

Example:

```
agasajo  -----> refresco                          Taxonomic Parent
refresco -----> drink                             Bilingual dictionary
drink    -----> DRINK_L_2_1                       LKB entry
===>
AGASAJO_X_I_3 / DRINK_L_2_1 :
PARTIAL-TLINK.
```

"agasajo" has "refresco" as its hypernym. "refresco", is looked up in the bilingual, resulting the translation "drink", which exists in the target lexicon. Therefore, a PARTIAL-TLINK connecting the source entry and the parent translation entry is produced.

• Grandparent Tlink Module

This is a very similar case to the previous one. In this case, the source word's grandparent is used to produce the partial tlink.

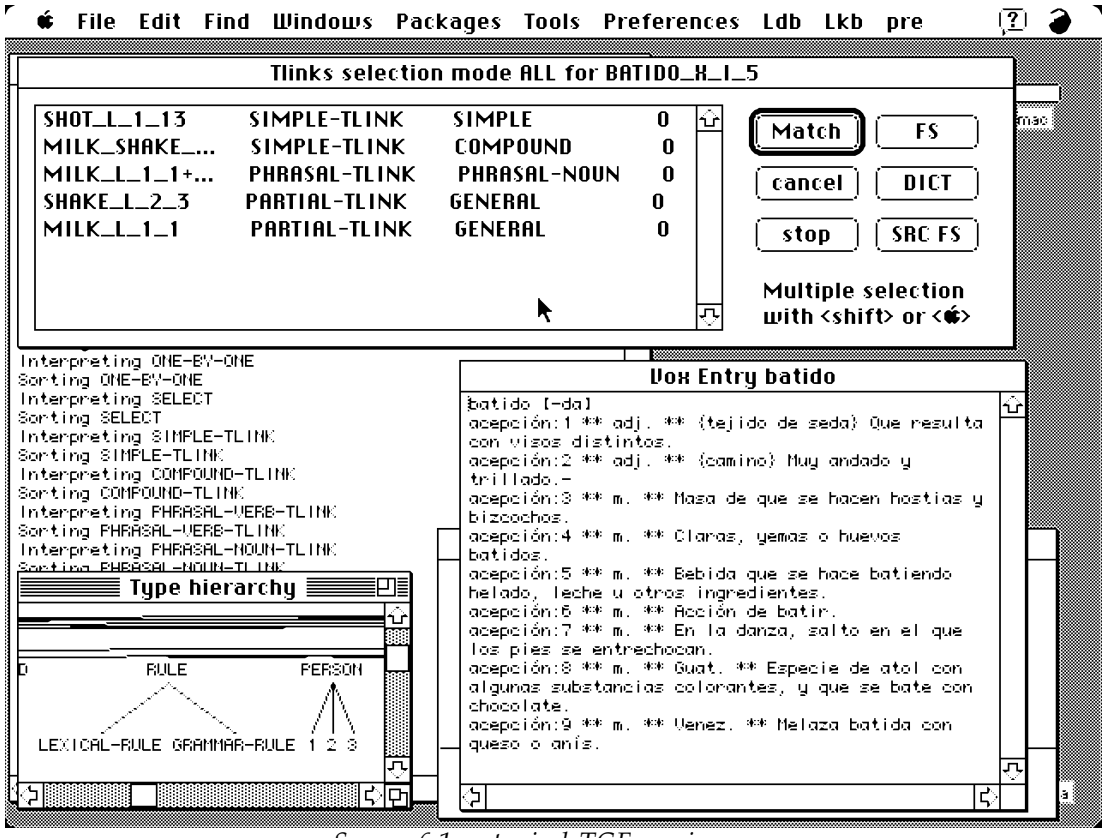
Algorithm: This module does the same than the previous one but with another layer in the taxonomy. That is to say, the grandparent is looked up in the bilingual and the corresponding target word is searched in the target lexicon. If the search succeeds a PARTIAL-TLINK between the source entry and the grandparent's translation is produced.

Example:

caridad -----> agasajo	taxonomic parent
agasajo -----> refresco	taxonomic parent
refresco -----> drink	Bilingual dictionary
drink -----> DRINK_L_2_1	LKB entry
absinth -----> ABSINTH_L_0_1	LKB entry
===>	
CARIDAD_X_I_4 / DRINK_L_2_1:	
PARTIAL-TLINK.	

6.3 Linking DGILE to LDOCE

Several experiments, corresponding to rather "closed" and narrow semantic domains, have been performed using TGE in the (semi)automatic approach. We discuss in this Section those experiments corresponding to "drinks" [Ageno et al. 94]. We will illustrate the tlink generation process with an example of an entry for which a number of different tlinks have been generated, namely *batido_X_I_5.* In the figure 3 where *batido_X_I_5* appears with the tlink options, we had selected the option *all*, and subsequently, all the possible tlinks have been suggested by the system. TGE allows, however, other selection criteria. As we can see in Screen 6.1, five tlinks are proposed by the system for this particular example:



Screen 6.1, a typical TGE session.

1) The first option is not a correct one. Among the various translations given for the source LKB entry *batido_X_I_5* the adjective *shot* appears; another syntactic realisation of *shot* is that of a noun denoting a drinkable thing as such it is included in the target subset.

2) The second is a simple-tlink type linking *batido_X_I_5* with the target LKB entry *milk_shake_L_0_0* . In this case, we have an example of the application of the compound-tlink-ruleset.

3) The third is a phrasal-tlink type, linking *batido_X_I_5* with the target LKB entries *milk_L_1_1* and *shake_L_2_3* composed by the + sign. This is an example of the application of the phrasal-noun-tlink-ruleset.

4) Both the fourth and fifth, are partial-tlink-types, linking *batido_X_I_5* with the target LKB entries *shake_L_2_3* and *milk_L_1_1* respectively. This is an example of the application of the general-tlink-ruleset.

The Spanish taxonomy of drink-nouns, extracted from VOX dictionary, consists of 235 noun senses, and has 5 levels. The English taxonomy of drink-nouns, extracted from LDOCE, consists of 192 noun senses.. Some of the obtained results are the following:

• Going from Spanish to English, 223 out of 235 drink-nouns have been linked by means of different, often more than one, tlinks (95 %). However, only 52 English nouns have been linked with Spanish nouns (27%). Out of these 223 drink-nouns mentioned above, 210 have been linked by using (mainly) the bilingual dictionary as a translation resource while the rest, that is, 13, have been linked by means of the orthographic-tlink ruleset, and, consequently, the gap of the bilingual dictionary has been bridged in the end, because in both languages the same word with exactly the same spelling is used. For example, *chartreuse_X_I_1* and *chartreuse_L_I_0*, *sherry_X_I_1* and *sherry_L_0_0,* etc.

• 74 out of 235 source LKB entries for drink-nouns are also bilingual entries (31,5%). Consequently, 161 source LKB entries have no corresponding bilingual entries (68,5%). This big gap in the bilingual dictionary is because the one used VOX/Harrap´s, is an essential one, and as such, it only contains 32,463 senses. By contrast, the VOX monolingual Spanish dictionary covers 143,700 senses.

• 30 out of the translations of the 74 source LKB entries that were found in the bilingual dictionary are also target LKB entries. Consequently, the translations of 44 bilingual entries have no corresponding target LKB entries.

• 13 out of 161 source LKB entries are also target LKB entries (8 %).

• For most entries, more than one tlink type has been extracted. The total number of tlinks that have been generated and selected for the taxonomy of bebida _X_I_3 (drink) with the explained software is 372 tlinks. We show next the different tlinks generated by each ruleset and the amount of lexical entries of each language involved.

	Tlinks	Spanish entries	English entries
simple-tlinks (14,5%)	55		
by simple-tlink-ruleset	41	26	31
by compound-tlink-ruleset	1	1	1
by orthographic-tlink-ruleset	13	13	13
phrasal-tlinks (0.5 %)	2		
by phrasal-noun-tlink-ruleset	2	1	3
partial-tlinks (85 %)	320		
by parent-tlink-ruleset	268	149	15
by grandparent-tlink-ruleset	44	30	10
by general-tlink-ruleset	8	7	6

Table 6.1, figures for the (semi)automatic use of TGE on drink domain.

All the tlink-rulesets have worked satisfactorily, therefore resulting in a considerable part of the subsets linked (95% of the source lexicon). However, these PRE tlink-rulesets have only been tested over limited subsets of specific semantic fields. Its real potential will be tested on a later stage, once its application to larger and less restricted sets of word senses (including categories different from nouns) takes place.

6.3.3 Linking DGILE to WordNet

The main drawbacks of the previously described methodology, as discussed in [Ageno et al. 94] are: 1) the poor coverage of English entries (only 27%) partially explained by the limited coverage of the bilingual dictionary used; and 2) the need of huge specialised human intervention for selecting the appropriate tlinks from those proposed by the system. This second point will be addressed in the following section.

What is presented now is a heuristic method based on conceptual distance that uses information from an external wide-coverage semantic taxonomy (WordNet). The main goal is to overcome the problem in an automatic way or to provide the user with complementary information in order to make the choice easier. The proposal is based on the use of a Conceptual Distance between the alternatives. The base for computing this distance is the use of WordNet.

In our previous approach all the tlinks extracted by means of the corresponding kowledge sources (basically the bilingual dictionary and the LKB) were offered to the user in order to allow the selection of the appropriate ones. This process was relatively high time consuming and needed knowledge of both source and target Languages by the user. Our proposal is to measure the conceptual distance between the lexical entry corresponding to the source language and the different lexical entries corresponding to the target language. Three modes of performance are then allowed to the user: 1) select automatically the most feasible, 2) select automatically all the tlinks over a determined threshold and 3) rank the tlinks and allow the user to make the selection manually.

The TGE environment using Conceptual Distance performs the creation of tlinks among lexical entries placed into the LKB and synsets in WordNet in a top-down fashion. Starting from the top lexical entry of the Spanish taxonomy the user selects the most feasible synsets of WordNet from those proposed by the rulesets using the bilingual dictionaries. Once the user has selected the equivalent synsets of the Spanish lexical entry in WordNet no further selection is required. Then the program applies recursively the TGE rulesets to all the hyponym lexical entries of the Spanish taxonomy. The Conceptual Distance among the equivalence translations proposed by the TGE environment and those selected previously (normally hypernym synsets) is computed, selecting those closest (a Conceptual Distance threshold can be used for selecting a set of feasible synsets), and so on. Applying the Conceptual Distance measure the tlinks proposed by one ruleset can also be rejected. In this

situation the TGE control mechanism decides what other ruleset must be launched. The automatic tlinks generation process is illustrated with the following example:

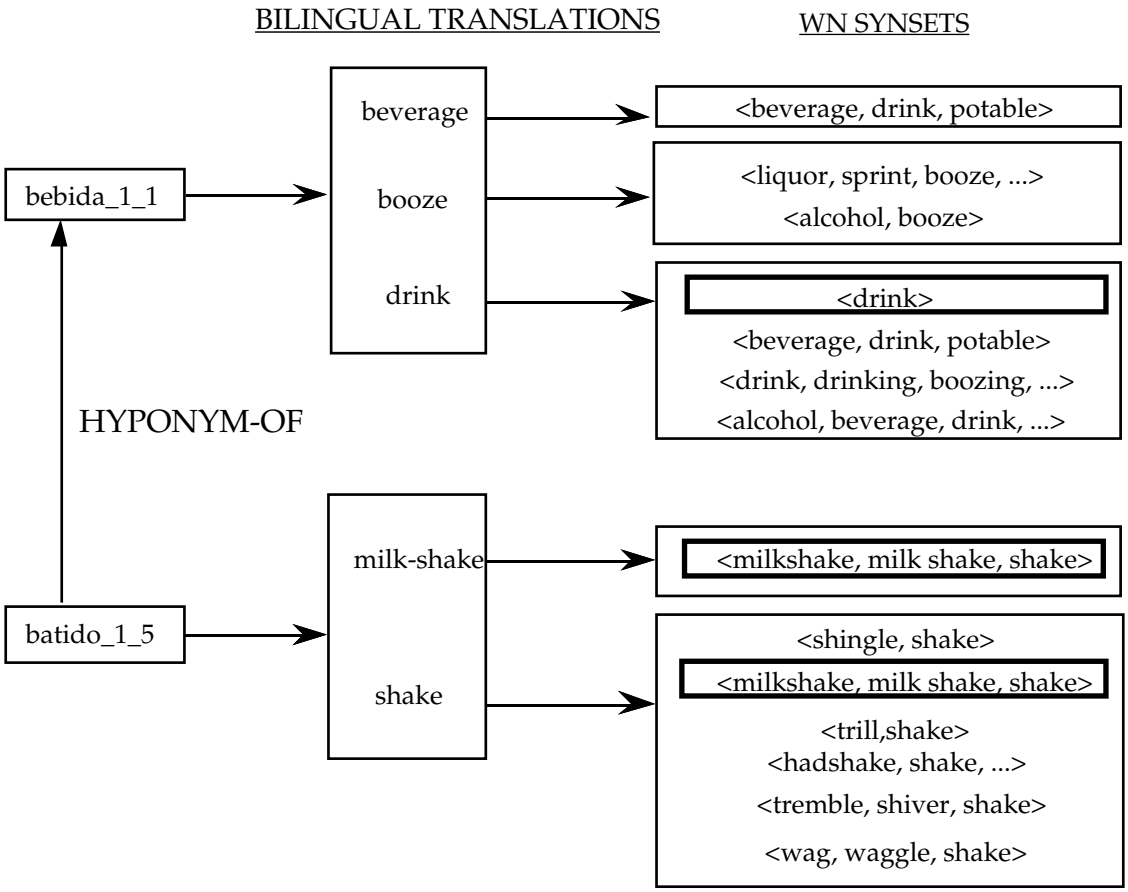


Figure 6.3, translation equivalence selection.

Once the translation links for *bebida_1_1* have been selected, all the possible translations of *batido* are looked up from the bilingual dictionary (if no translations are found in the bilingual dictionary, other rulesets are launched in order to overcome this lexical gap, such as *parent-tlink-ruleset*, etc.). Applying the disambiguation module using the Conceptual Distance among those synsets proposed for *batido_1_5* and those previously attached for *bebida_1_1*, the closest ones are selected (in bold squares). These selected synsets act as constraints for further disambiguation processes with the hyponyms of *batido* .

Several experiments have been undertaken on the same domains of precedent ones. In the food domain from 140 source lexical entries, up to 54 lexical entries (only 39%) has direct (by means of bilingual dictionaries) and correct (a correct sense for the translation is placed in WordNet) equivalent synsets in WordNet. This result is good taken into account the different sources of error: 1) no existence of translation in the bilingual dictionary (50 cases), 2) there is a translation but not the correct one (30 cases), 3) there is no correct sense into WordNet (6 cases), 4) the translation does not appears in WordNet (no errors detected in this taxonomy).

Although the lexical gap among the three lexical knowledge sources used in his experiment, all the lexical entries that belongs to the taxonomy of *comida* have been linked to WordNet synsets using the rulesets presented in [Ageno et al. 94] in a fully automatic way. The results have been the following:

simple-tlinks	**57**
simple-tlink-ruleset	52
compound-tlink-ruleset	2
orthographic-tlink-ruleset	3
phrasal-tlinks	**1**
phrasal-noun-tlink-ruleset	1
partial-tlinks	**84**
parent-tlink-ruleset	78
grandparent-tlink-ruleset	6

Table 6.2, figures for the automatic use of TGE on FOOD domain.

6.5 Linking Bilingual Dictionaries to WordNet

Our third experiment can be considered as an initial attempt to build WordNets from bilingual dictionaries. It's commondly agreed that WordNet has become a de-facto standard wide-coverage ontology for a wide range NL tasks.

WordNet success has encouraged several projects in order to build WordNets (WNs) for other languages or to develop multilingual WNs. The most ambitious of such efforts is EuroWordNet (EWN)[1], a project aiming to build a multilingual WordNet for several European languages[2]. The work we present here is included within EWN and presents our approach for (semi)automatically building a Spanish WN (see [Atserias et al. 97], [Benítez et al. 98] and [Farreres et al. 98]). The main strategy within our aproach is to map WN1.5 thus creating for Spanish a parallel-in-structure network. Therefore, our main goal is to attach Spanish word meanings to the existing WN1.5 concepts. This paper describes automatic techniques that have been developed in order to achieve this goal for nouns.

This section explores the automatic construction of a multilingual Lexical Knowledge Base directly from a pre-existing lexical structure. First, a set of automatic and complementary techniques for linking Spanish words collected from monolingual and bilingual MRDs to English WordNet synsets are described. Second, we show how resulting data provided by each method is then combined to produce a preliminary version of a Spanish WordNet with an accuracy over 85%. Both coarse-grained (class level) and fine-grained (synset assignment level) confidence ratios are used and evaluated. Finally, the results for the whole process are presented.

Our approach for building the Spanish WN (SpWN) is based on the following considerations:

• The close conceptual similarity of English and Spanish allows for the preservation of the structure of WN1.5 in order to build the SpWN. Moreover, when necessary, lexicalization mismatches are solved using multi-word translations (collocations) supplied by bilingual dictionaries.

• An extensive use of pre-existing structured lexical sources is performed in order to achieve a massive automatic acquisition process.

• The accuracy of cross-language mappings is validated by hand on a sample. Each attachment to WN bears a confidence score (CS).

• Only attachments over a threshold are considered. Moreover, a manual inspection of attachments in a given range will be carried out.

---

1 *EuroWordNet: Project LE- 4003 of the EU.*

2 *Initially three languages, apart from English, were involved: Dutch, Italian and Spanish. The project has been recently extended for covering French and German.*

Undoubtfully, following this approach most of the criticisms placed to WN1.5 also apply to SpWN: too much sense fine-grainedness, lack of cross-POS relationships, simplicity of the relational information, not purely lexical but lexical-conceptual database, etc. Despite of these drawbacks, WN1.5 is widely used and tested and supports few but the most basic semantic relations. Our approach ensures that most of the huge networking effort, which is necessary to build a WN from scratch, is already done.

The different sources involved in the process show a different accuracy. High CSs can be assigned to original sources, as MRDs, but derived sources, which result from the performance of automatic procedures, come to bear substantially lower CSs. Our major claim is that multiple source/procedures leading to the same result will increase the particular CS, while when leading to different results the overall CS will decrease.

Several lexical sources have been applied in order to assign Spanish WMs to WN1.5 synsets:

1) Small Spanish/English and English/Spanish bilinguals
2) A large Spanish monolingual dictionary DGILE
3) English WordNet (WN1.5).

By merging both directions of the bilingual dictionaries what we call homogeneous bilingual (HBil) has been obtained. The maximum synset coverage we can expect to reach by using HBil due to its small size is 32%. In the table 6.3, the summarised amount of data is shown.

	English nouns	Spanish nouns	synsets	Connections[5]
WordNet1.5	87,642	-	60,557	107,424
Spanish/English	11,467	12,370	-	19,443
English/Spanish	10,739	10,549	-	16,324
HBil	15,848	14,880	-	28,131
Maximum Reachable Coverage - of WordNet - of bilingual	12,665 14% 80%	13,208 - 90%	19,383 32% -	66,258 - -

Table 6.3, some figures of the bilingual mapping onto WordNet.

6.5.1 Methods

Bilingual entries must be disambiguated against WN. The different procedures developed for linking Spanish lexical entries to WN synsets can be classified in two main groups according to the kind of knowledge sources involved in the process[1]:

- **Class methods**: use as knowledge sources individual entries coming from bilinguals and WN synsets.
- **Conceptual Distance methods**: makes use of knowledge relative to meaning closeness between lexical concepts.

All the methods have been manually inspected in order to measure its CS. Such tests have been performed on a random sample of 10% using the Validation Interface (VI), an environment designed to allow hand validation of Spanish word forms to WN synsets

---

[5] *Connections can be word/word or word/synset. When there are synsets involved the connections are Spanish-word/synset,(except for WordNet itself), otherwise Spanish-word/English-word.*

[1] *We can consider also other types of methods, as the structural methods, presented in [Atserias et al. 97].*

assignment. It allows to consult and to navigate through the monolingual and bilingual lexical databases and WN. The following diagnostics can result from the performance of this validation:

ok:	correct links.
ko:	fully incorrect links.
hypo:	links to a hyponym of the correct synset.
hyper:	links to a hyperonym of the correct synset.
near:	links to near synonyms that could be considered ok.

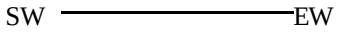
6.5.1.1 Class Methods

Following the properties described in [Rigau & Agirre 95] Hbil has been processed and 2 groups of 4 different cases have been collected depending on whether the English words are either monosemous or polysemous relative to WN 1.5. Afterwards two hybrid criteria are considered as well.

a) Monosemic Criteria

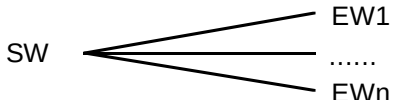
These criteria apply only to monosemous EW with respect to WN1.5. As a result, this unique synset is linked to the corresponding Spanish words.

- **Monosemic-1 criterion (1: 1):**



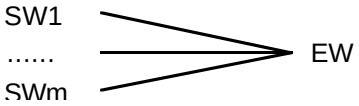
A Spanish Word (SW) has only one English translation (EW); symmetrically, EW has SW as its unique translation.

- **Monosemic-2 criterion (1:N with n>1)**



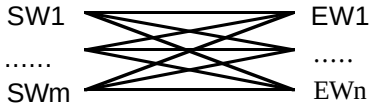
A SW has more than one translation; each EW has SW as its unique translation.

- **Monosemic-3 criterion (M:1 with m>1):**



Several SWs have the same translation; EW has several translations to Spanish.

- **Monosemic-4 criterion (M:N with m>1 & n>1)**



Several SWs have different translations; EWs also have several translations.

b) Polysemic Criteria

Applying the corresponding cases for those English polysemic nouns in WN1.5, criteria polysemic 1 to 4 have been obtained.

Although more complex strategies can be considered in order to decide which of the multiple synsets related to the English words must be linked to their Spanish translations, on the four polysemic criteria all synsets have been selected, assuming some of them are incorrect.

c) Hybrid Criteria

- **Variant criterion**

For a WN1.5 synset which contains a set of variants EWs[1], if it is the case that two or more of the variants EWi have only one translation to the same Spanish word SW, a link is produced for SW into the WN1.5 synset.

- **Field criterion**

This procedure makes use of the existence of a field identifier in some entries (over 4,000) of the English/Spanish bilingual. For each English entry bearing a field identifier (EW), if it is the case that both occur in the same synset, for each EW translation to Spanish a link is produced. Results of the manual verification for each criterion are shown in the table 6.4.

Criterion	#links	#synsets	#words	%ok	%ko	%hypo	%hyper	%near
mono1	3697	3583	3697	92	2	2	0	2
mono2	935	929	661	89	1	5	0	3
mono3	1863	1158	1863	89	5	0	2	1
mono4	2688	1328	2063	85	3	6	2	4
poly1	5121	4887	1992	80	12	0	0	6
poly2	1450	1426	449	75	16	2	0	5
poly3	11687	6611	3165	58	35	0	1	5
poly4	40298	9400	3754	61	23	5	1	9
Variant	3164	2195	2261	85	4	4	1	6
Field	510	379	421	78	9	2	2	9

Table 6.4, resulting figures of the class methods.

6.5.1.2 Conceptual Distance Methods

As in other parts of the work presented in this thesis, the Conceptual Distance formula used in this work is shown in formula 6.1 (see Section 4.3.4 for details).

(6.1)
$$CD(w_1, w_2) = \min_{\substack{w_1 \in c_i \\ w_2 \in c_j}} \sum_{k \in shortestpath(c_i, c_j)} \frac{1}{depth(c_k)}$$

where w_i are words and c_i are synsets representing those words. Conceptual Distance between two words depends on the length of the shortest path that connects the concepts and the specificity of the concepts in the path. Then, providing two words, the application of the Conceptual Distance formula selects those closer concepts that represent them. Following this approach, three different sources has been used.

[1] *A variant is a word of a synset.*

a) Using Co-occurrence words collected from DGILE (CD1)

Following [Wilks et al. 93] two words are coocurrent in a dictionary if they appear in the same definition. For DGILE, a lexicon of 300,062 coocurrence pairs among 40,193 Spanish word forms was derived and the affinity between these pairs was measured by means of the Association Ratio (AR), which can be used as a fine grained CS.

Then, the Conceptual Distance formula for all those pairs has been computed using HBil and the nominal part of WN. Consider for instance the following coocurrence word pair, the association ratio and their possible translations, synsets and semantic files:

AR	Sw1	Ew2	synset	SF	Sw2	Ew2	synset	SF
1.8524	**bebida**	beverage	05074818	13	**chocolate**	chocolate	03472382	07
		bozze	02004443	06			04861776	13
			05089006	13			05106900	13
		drink	00418859	04		dope	02672720	06
			05074818	13			04338801	10
			05076795	13			06061223	18
			05077192	13		hash	02673273	06
							05064828	13

The method computes the Conceptual Distance for all possible synset combinations and selects the one that return the minimal score. In this case, synset 05074818 for bebida and 05106900 for chocolate (the second is direct hyponym of the first).

05106900 <cocoa, **chocolate**, hot chocolate> -- (made from baking chocolate ...)
=> 05074818 <**beverage**, drink, potable> -- (any liquid suitable for drinking)
=> <food, nutrient> -- (any substance that can be metabolized by an organism...)
=> <liquid> -- (a substance that is liquid at room temperature and pressure)

b) Using Headword and genus of DGILE (CD2)

Computing the Conceptual Distance formula on the headword and the genus term of 92,741 nominal definitions of DGILE dictionary (only 32,208 with translation to English). This proccess has been described in Sections 5.2.2.1 and 5.3.2.4.8.

c) Using Spanish entries with multiple translations in the bilingual dictionary (CD3)

In this case, we have derived a small but closely related lexicon of 3,117 translation equivalents with multiple translations from the Spanish/English direction of the bilingual dictionary (only 2,542 with connection to WordNet1.5). Consider the next bilingual entry:

chocolate *m* chocolate. 2 *arg* (*hachís*) dope, hash. • *fam fig l**as cosas claras y el ~ espeso***, let´s get things clear •• ***~a la taza***, drinking chocolate; ***~con leche***, milk chocolate; ***tableta de*** ~, bar of chocolate.

In this case, we can perform the disambiguation process considering sense 2, that is, between dope and hash, selecting in this case synset 02672720 for dope and synset 02673273 for hash (the second is direct hyponym of the first).

02673273 <hashish, **hash>** -- (purified resinous ...; used as a hallucinogen)
=> 02672720 <cannabis, marijuana, ganja, pot, grass, marihuana, **dope**...>-- (...)
=> <soft drug> -- (a narcotic that is considered relatively mild)

Table 6.5 summarises the performance of the three Conceptual Distance methods described above.

Criter.	#links	#synsets	#words	%ok	%ko	%hypo	%hyper	%near
CD - 1	23,828	11,269	7,283	56	38	3	2	2
CD - 2	24,739	12,709	10,300	61	35	0	0	3
CD - 3	4,567	3,089	2,313	75	12	0	2	8

Table 6.5, resulting figures for Conceptual distance methods.

6.5.2 Combining Methods

Collecting those synsets produced by the methods described above with an accuracy greater than 85% (mono1, mono2, mono3, mono4, variants, field), we obtain a preliminary version of the Spanish WordNet containing 10,982 connections (2,830 polysemous) between 10,786 synsets and 9,986 Spanish nouns with an overall CS of 87,4%. However, combining the discarded methods we can take profit of portions of them precise enough to be acceptable.

All files resulting from discarded methods were crossed and their intersections were calculated. Using VI, a tool designed to perform this task, a manual inspection of samples from each intersection was carried out. Results are shown in the table 6.6.

		method2					
method1		cd2	cd3	p1	p2	p3	p4
cd1	size	15736	**1849**	**2076**	**556**	3146	15105
	%ok	79	**85**	**86**	**86**	72	64
cd2	size	0	**2401**	**2536**	**592**	3777	13246
	%ok	0	**86**	**88**	**86**	75	67
cd3	size	0	0	**205**	**180**	**215**	3114
	%ok	0	0	**95**	**95**	**100**	77
p1	size	0	0	0	0	**77**	**178**
	%ok	0	0	0	0	**100**	**88**
p2	size	0	0	0	0	28	**78**
	%ok	0	0	0	0	77	**96**

Table 6.6, performance of the intersection.

In bold appear intersections with a CS greater than 85%. Up to 7,244 connections (5,877 polysemous) can be selected with 85.63% CS, 4,780 of which are new with an overall CS of 84% resulting in a 30% increase. It must be pointed out that most of these connections correspond to highly polysemous words (4,553 new connections). Then a second version of the Spanish WordNet has been obtained containing 15,535 connections (7,383 polysemous) among 10,786 synsets and 9,986 Spanish nouns with a final accuracy of 86,4%. Table 6.7 shows the overall figures of the resulting SpWNs.

Criterion	#links	#synsets	#word	#CS	#poly links
SpWN v0.0	10,982	7,131	8,396	87.4	1,777
Combination	7,244	5,852	3,939	85.6	2,075
SpWN v0.1	15,535	10,786	9,986	86.4	3,373

Table 6.7, overall figures of SpWNs.

The application of these combinations results on an increment of the extracted connections of a 30% without losing accuracy.

The approach seems to be extremely promising, attaching up to 75% of reachable Spanish nouns and 55% of reachable WN1.5 synsets. Currently we are performing complementary experiments for extending the approach for covering other lexical sources, especially wider-coverage bilinguals.

Other lines of research we are following by now include: 1) dealing with mismatches, i.e., when coming from different method/source an Spanish word is assigned to different synsets. If in the former case the overall CS increases, in the last one it should decrease. 2) A fine grained cross-comparison of methods and sources (intersections of more than two classes, decomposition of classes into finer ones, etc.) will be performed to obtain a more precise classification and CS assignment. 3) We are trying to obtain an empirical method for CS calculation of intersections. Methods based on bayesian inference networks or quasiprobabilistic approaches have been tested giving promising results.

6.6 Conclusions

In this Chapter we have presented TGE, an environment designed and built in order to aid in the recovery of cross-linguistic relations. We have reported and described results of an experiment for (semi)automatically and automatically extracting equivalence relations for Spanish and English drink-nouns by using the TGE software.

The first experiment has been carried out on the "drink" domain in a (semi)automatic way, as the tlink generation is performed automatically, whilst the selection of the desired tlinks is done manually.

Moreover, a fully automatic method selecting the most likely tlink among a set of candidates has been also presented. The proposal tries to overcome the main problem found on (semi)automatically extracting translation links between multilingual lexical entries using as main knowledge sources bilingual dictionaries. The system mechanism is based on calculating the conceptual distance between the competing lexical entries in the target language. Then, we select the higher ranked concept from those previously linked that appears higher in the taxonomy.

Finally, an approach for building multilingual Wordnets combining a variety of lexical sources as well as a variety of methods has been proposed which tries to take profit of the existing WordNet structure for attaching words from other languages in a way guided mainly by the content of bilingual lexical sources. A central issue of our approach is the combination of methods and sources in a way that the accuracy of the data obtained from the combined sources overcomes the accuracy obtained from the individual sources. Several families of methods have been tested, each of them bearing its own confidence score. Only those methods offering a result over a threshold (85%) have been considered. In a second phase of our experiments, intersections between the results provided by the different individual methods have been performed. However, it is clear that valuable set of entries, owning an insufficient, in some cases rather bad, individual CS can be extracted if they occur as a combination of several methods. In this way, using the same threshold, the amount of synsets attached to Spanish entries has increased. It must be pointed out that **all** these new connections correspond to highly **polysemous** words.

Currently, using the approaches and techniques provided in this Chapter, we are working on the construction of the first versions of the Spanish and Catalan WordNets (see [Benítez et al. 98] for further details). Furthermore, we are planning to use during the mapping process the taxonomies derived from the monolingual dictionary (first attempts in this direction has been also performed [Farreres et al. 98]).

Chapter 7

Conclusions and Further Work

7.1 Introduction

This thesis focuses on the massive acquisition of lexical knowledge from monolingual and bilingual conventional dictionaries (on-line dictionaries or Machine-Readable Dictionaries, MRDs). A complete productive methodology for acquiring useful lexical knowledge from MRDs has been designed. SEISD, a powerful, complete and flexible software system allowing us to acquire massive lexical knowledge from on-line monolingual and bilingual dictionaries and to represent and validate the lexical knowledge acquired in a Multilingual Lexical Knowledge Base, has been designed and implemented to perform the methodology. The proposed methodology has been applied to the extraction of lexical information from DGILE: a huge, loosely structured Spanish monolingual dictionary. Both issues 1) the methodology and 2) the application to DGILE constituted the main objectives of the thesis and its main contributions. Finally, we have proposed, implemented and experimentally tested various techniques in different methodological steps, obtaining improvements for several of them.

The lexicon, which represent lexical information reliably and precisely enough for automated use, is recognised as one of the major problems in NLP applications both because of the need for substantial vocabulary in habitable NLP systems and because of the increasing complexity. The "lexical bottleneck" [Briscoe 91] is even worse for languages other than English. The work presented here tries to lay down solutions to overcome or alleviate this problem. In comparison with other methodologies for acquiring massive lexical knowledge (from introspection or corpora) we think that the main advantages of the approach taken here lie in: economy (little human-labour is involved), productivity (richness and variety of the resulting data, even applying little effort) and modularity (the current results can be improved using different techniques in a stepwise refinement).

This chapter summarises the contributions described in the thesis, presents general conclusions, comments on the results of the experiments reported, and suggests directions for further work. Thus, in Section 2 the main goals achieved during this work are shown. Section 3 lists the main lexical resources acquired from the MRDs during the work presented here. Section 4 describes the further work we are planning to do, and at the end, Section 5 presents a final summary.

7.2 Main contributions

In this thesis we set out to achieve the massive automatic acquisition of lexical knowledge from conventional dictionaries allowing the easy construction (or derivation) of a large set of rich lexicons (from MTDs to multilingual LKBs) suitable for use in a wide range of NLP systems (morphological analysers, Information Retrieval systems, Machine Translation applications, etc.). While for English a huge set of rich lexical resources are available (highly coded MRDs such as LDOCE, Lexical Data Bases such as Comlex, Lexical Knowledge Bases such as WordNet, etc.) this is not the case for the majority of languages. However, a

great deal of monolingual and bilingual dictionaries is available for many languages. The possibility of obtaining large computational lexicons for NLP tasks from them using automatic techniques (even for less coded and structured dictionaries than LDOCE) has been explored in this thesis.

MRDs are not, of course, the only lexical knowledge resource suitable to be exploited to obtain lexical knowledge. Although some researchers have pointed out that dictionaries are inadequate as a source of some kinds of lexical knowledge for sophisticated Natural Language Processing [1] it is our belief that (as has been demonstrated in the Acquilex I, Acquilex II and EuroWordNet projects) dictionaries are the main lexical knowledge resource available for building large, useful lexicons for NLP quickly.

In particular, we have designed a complete methodology to build and validate a multilingual LKB from a set of monolingual MRDs using bilingual MRDs to aid the linking process between languages. We have applied this methodology to a concrete set of monolingual and bilingual MRDs (with their own particular characteristics: size, encoding, information content, etc.) without loosing generality. However, our methodology can be applied to any monolingual descriptive dictionary[2] in any language[3]. The main issues for delineating the base methodology have been the characteristics of the lexical resources used, the information to be extracted from them, how to carry out the process and how to represent and exploit the information extracted.

As the majority of MRDs are not built for computational purposes (e.g., they contains circularity, errors and inconsistencies, etc.) we designed a mixed methodology. We described a set of semantic primitives using the LKB (prescriptive approach) and we placed this in a natural classification of the concepts represented implicitly in the MRD definitions (descriptive approach). Thus, first, we developed several techniques for detecting (and/or selecting) the main semantic subsets underlying MRD definitions. Second, we justified that a small set of dictionary senses are not enough to lead to full coverage of a semantic subset, and thirdly, we studied a novel and productive method for discovering the main top dictionary senses representative of a given semantic subset.

We covered the whole methodology implementing a complete modular computer system called **SEISD** (*Sistema d'Extracció d'Informació Semàntica de Diccionaris*) which provides a user-friendly interface with five subsystems and also a way of integrating these subsystems with the management of the multiple sources of heterogeneous data used by the system. SEISD was designed as a medium for the extraction methodology and is fully integrated with Acquilex representational formalisms and their supporting software tools. SEISD covers the main functions of the proposed methodology, that is, the extraction of semantic information implicitly located in DGILE (performed by **TaxBuild** and **SemBuild**), the mapping process of the information extracted to the LKB (covered by the **CRS**), the multilingual acquisition process (performed by the **TGE**) and the validation and exploitation of the lexical knowledge acquired (carried out by the **LDB/LKB** System).

A central guideline was to build the whole system so as to perform each process (semi)automatically. Once SEISD was finished, each module was tested in order to analyse its performance (the test on the (semi)automatic use of SEISD was reported in [Castellón 93] and [Taulé 95]). Then, some improvements in both methodology and techniques applied were introduced in some modules for efficacy (to obtain more information) and efficiency (to obtain this information more easily). A second test was performed to compare the results with the previous ones, improvements being obtained in both aspects (efficacy and efficiency).

Now, for each of the subsystems we will summarise the main contributions reported in this thesis.

---

[1] *Obviously, knowledge like "You have to be awake to eat" (from [Lenat 95]) are unlikely to be published in textbooks, dictionaries, magazines, or encyclopaedias, even those designed for children.*

[2] *Other dictionaries (e.g., synonym dictionaries, acronym dictionaries, etc.) could also be useful lexical resources for acquiring lexical knowledge automatically.*

[3] *In fact, this methodology is also being applied to French, Euskera and Catalan dictionaries.*

- **On semantic knowledge acquisition from genus terms (TaxBuild)**

The semantic knowledge acquisition function of SEISD on genus terms is performed by Taxonomy Builder (**TaxBuild**) [Ageno et al. 91b], [Ageno et al. 92b], one of the most important mechanisms of SEISD. This module produces complete disambiguated and partially analysed (using SegWord morphological analyser and FPar syntactic-semantic analyser) dictionary sense taxonomies from DGILE.

A central issue in the thesis is the genus disambiguation task allowing the automatic construction of taxonomies from traditional monolingual dictionaries without any special encoding. We reported a very successful result with a set of different informed heuristics, combining their results (see [Rigau et al. 97]). We improved the results of two previous large-scale disambiguation heuristics based on those described in [Yarowsky 92] and [Wilks et al. 93]. We developed a successful new heuristic based on that reported in [Rigau 94], which is the combination of wide-range large-scale lexical resources (WordNet and bilingual dictionaries) and the notion of conceptual distance to enrich monolingual dictionary senses with WordNet semantic tags. We created and tested various conceptual distance formulae for this purpose (see [Agirre & Rigau 95], [Agirre & Rigau 96a]). We applied the heuristics together, obtaining better results combining them rather than applying each one separately.

We carried out a new approach (see Section 5.2 and 5.3) for selecting those main genus terms for a given semantic primitive, and after a filtering process, we applied another novel technique for deriving fully automatic and accurate taxonomies from any conventional dictionary (these results are also published in [Rigau et al. 97] and [Rigau et al. 98]).

- **On semantic knowledge acquisition from the differentia (SemBuild)**

We proposed and implemented a methodology for performing a deeper analysis of the implicit information located in each dictionary sense belonging to a semantic subset once the construction of all its taxonomies has finished. Because of the lack of complete grammars and robust analysers for Spanish, we proposed a cycling methodology for enriching partial grammars systematically. That is, given all dictionary senses belonging to a prescribed semantic subset (e.g., FOOD) and its former representation in the LKB, we studied highly frequent syntactic patterns, which denote conceptual relations between concepts. Furthermore, using lexical knowledge acquired previously (i.e., taxonomies) some partial syntactic analysis can now be semantically interpreted. Thus, following (semi)automatic techniques we acquired in-depth formal semantic representations of dictionary senses.

We are currently using a broad range morphological analyser of Spanish [Acebo et al. 94] and a tagger of Spanish [Padró 98] and a shallow DCG grammar to parse all dictionary definitions completely. Perhaps an in-depth grammar/parser of Spanish could lead to better results, but building such a tool is beyond the scope of this research and given the kind of material to be parsed (because of the sublanguage used in dictionaries) and the goals of the acquisition, partial coverage does not seem to be a serious limitation. Thus, rather than analyse small parts of the dictionary definitions (i.e., [Alshawi 89], [Artola 93], [Castellón 93]) we propose (when no full parse can be performed with high accuracy) the complete analysis of the dictionary definition using a shallow parser which provides a fully analysed set of longest chunks for an input definition (see Section 5.3).

- **On the mapping process from the analysed taxonomies to the LKB (CRS)**

The main aim of the Conversion Rules System (**CRS**) [Ageno et al. 92c], [Ageno et al. 92d] in the SEISD environment is to perform the conversion of the semantic information extracted from the partially analysed dictionary senses to lexical entries constrained by the Type System of the LKB. That is, taking the analysed taxonomy generated by the TaxBuild and SemBuild systems, the CRS was designed in order to perform the translation from one structure to the other in the most declarative way. Thus, we implemented the CRS using the PRE, a rule-oriented general purpose interpreter deeply adapted to natural language

applications and capable of managing several complex and heterogeneous lexical knowledge resources (taxonomies, Type System, bilingual dictionaries, etc.).

- **On the multilingual lexical knowledge acquisition (TGE)**

Translation Links between lexicons can, of course, be established manually, but the multiplicity of cases occurring and the existence of several heterogeneous knowledge sources, such as bilingual dictionaries, monolingual LDBs and multilingual LKBs, motivates the automation of the process. To help perform this task we developed the Tlinks Generation Environment (**TGE**) [Ageno et al. 94].

Like the CRS, the TGE was implemented using the PRE and may be considered a toolbox and, thus, it does not impose a single methodological strategy. We designed an initial set of modules according to the typology of Translation Links. It included several sorts of Translation Links that showed different conceptual correspondences between the two languages.

We tested the module by applying two different methodologies on several massive lexical resources, the first one semi-automatically between DGILE and LDOCE taxonomies [Ageno et al. 94] and the second one between DGILE and WordNet in a fully automatic way using the notion of conceptual distance developed for sense identification purposes [Rigau et al. 95]. In both cases we used bilingual dictionaries as a large-scale lexical resource to aid the linking process.

Besides using the TGE for bilingual sense linking between lexicons derived from monolingual MRDs we performed several experiments for mapping directly bilingual dictionaries to a Lexical Knowledge Base. That is, we attached Catalan and Spanish words to WordNet synsets using, as in the previous case, bilingual dictionaries producing noun preliminary versions of Catalan and Spanish WordNets [Atserias et al. 97], [Benítez et al. 98] and [Farreres et al. 98].

- **On the validation and exploitation of lexical knowledge acquired (LDB/LKB integration)**

We developed the LDB/LKB merging system [Rigau et al. 94] to allow the evaluation and validation of the lexical knowledge acquired and placed in the LKB.

Once the information contained in the dictionary definitions has been represented as a lexicon in the LKB, some testing processes should be performed on the lexicon acquired in order to improve the information extracted (e.g., detect possible errors or inconsistencies, extract more information, etc.) to determine which changes to make in the next acquisition loop. The LKB guarantees the appropriateness of the lexicon against the Type System and provides some generative inference mechanisms (e.g., the inherence mechanism distributes the information from the top level lexical units to the most specific ones, lexical rules produce new lexical entries from the pre-existing ones, etc.) but no facilities are provided for performing complex queries on the content of the lexical entries represented in the lexicon.

For the purposes of both the validation and the exploitation of the information acquired, the LDB/LKB merging system has the function of both systems: LDB-like access to an LKB lexicon.

- **On lexical knowledge acquired**

For testing purposes we applied the whole methodology to restricted subsets, obtaining large and rich Spanish lexicons placed in the MLKB. Although the only resulting data we expected after applying the whole methodology were these lexicons, throughout process a large number of extensive Spanish MTDs (ready-usable Spanish lexicons) were derived from the dictionaries themselves to aid with some methodological steps (e.g., lexicons for the morphological analysis, cooccurrence vectors as source data for a particular heuristic, harmonised bilingual dictionaries, etc.), among others, word frequency lists, bigrams, trigrams, cooccurrence lexicons, part-of-speech lexicons, word taxonomies, Spanish salient

word lists for every WordNet semantic file, word-sense disambiguated taxonomies, semantically tagged dictionary sense lexicon, bilingual lexicons, etc.

7.3 Main results

This section describes the main quantitative results achieved in the thesis. That is, we will summarise the results following the main experiments carried out.

- **On semantic knowledge acquisition from genus terms (TaxBuild)**

Selecting the correct genus term. We developed two different specialised grammars for detecting the genus term for noun and verb definitions. The noun grammar was applied to all 93,394 noun definitions of DGILE, the obtaining the genus terms for 92,693 (99%) of them (97% accuracy). We also obtained the genus terms for the 26,465 verb definitions.

Selecting the main top beginners for a selected semantic primitive. We proposed a novel methodology which combines several structured lexical knowledge resources for acquiring the most important genus for a given semantic primitive.

We labelled automatically the noun dictionary twice, the first time, computing the conceptual distance between headword and genus of the noun definitions. Assigning WordNet synsets to Spanish headwords, the program classified 29,205 of the noun DGILE definitions (31% of the whole nominal part) into 24 different semantic classes (corresponding to the 24 WordNet lexicographer's files) with 64% accuracy. Following the method proposed by [Yarowsky 92], we used this preliminary classification to partition DGILE into 24 subcorpora. We used this classification to acquire the salient words for each semantic class the subcorpus was representing. Using these salient words we labelled DGILE again, classifying the 86,759 noun definitions (93% of the nominal part of DGILE) with an overall accuracy of 80%. Finally, for each semantic category, after a filtering process we collected all its representative genus terms. All the genus terms gathered for a semantic category are the main top beginners for the semantic primitive we were looking for.

To bridge the language gap between WordNet and DGILE we used a Spanish/English bilingual dictionary.

(Semi)automatic genus sense identification. In the (semi)automatic approach, our attention was focused on different semantic subsets of nouns and verbs. For nouns, we derived lexicons for *substancia* (*substance*, including *food*), *persona* (*person*), *lugar* (*place*) and *instrumento* (*instrument*). Using TaxBuild we constructed (semi)automatically (that is, in a supervised mode, see [Castellón 93]) the complete disambiguated noun taxonomies, taking these words as a starting point containing 3,210 dictionary senses (382 belonging to the FOOD domain).

Automatic genus sense identification. We performed several experiments on the performance of the eight different heuristics applied for genus sense identification. The automatic construction of taxonomies from conventional dictionaries without either special encoding or a supervised technique accounted for much of the effort put into this thesis. By applying several robust and informed heuristics, we achieved very successful results (83% correct hypernym sense identification). In this way, we derived a completely disambiguated taxonomy of Spanish. This taxonomy contains 111,624 dictionary senses and has only 832 dictionary senses which are tops of the taxonomy (these top dictionary senses have no hypernyms), and 89,458 leaves (which have no hyponyms). That is, 21,334 definitions are placed between the top nodes and the leaves.

Furthermore, using the most relevant genus terms for a particular semantic taxonomy gathered previously and applying a filtering process, we are able to construct fully automatic taxonomies from any conventional dictionary. We applied the methodology to the FOOD semantic primitive. Thus, using the first set of criteria (LABEL2+F2+F3>9, 100% accuracy) we

acquire a FOOD taxonomy with 952 senses (more than twice the size than if the operation is done manually). Using the second one (LABEL2+F2+F3>4, 96% accuracy), we obtain another taxonomy with 1,242 (more than three times larger). Using the first set of criteria, the 33 genus terms selected produces a taxonomic structure with only 18 top beginners, while the second set, with 68 possible genus terms, produces another taxonomy with 48 top beginners.

The results show that the construction of taxonomies using lexical resources is not limited to highly structured dictionaries. Applying appropriate techniques, monolingual dictionaries such as DGILE could be useful resources for building substantial pieces of an LKB automatically.

- **On the analysis of the differentiae (SemBuild)**

In order to compare the performance of the current shallow parsing process we analysed the FOOD taxonomies. While [Castellón 93] captures 883 chunks of information we collected 2,760 chunks for taxonomies derived from LABEL2+F2+F3>9 filtering criteria, and for taxonomy derived from LABEL2+F2+F3>4 filtering criteria, a total amount of 3,270 chunks. That is, on average we are doubling the total amount of pieces of information acquired.

- **On the mapping process from the analysed taxonomies to the LKB (CRS)**

The system was tested by [Castellón 93] and [Taule 95] and no further methodological improvements have been performed. A technical improvement has been performed derived from the use of PRE rather than an ad-hoc mapping engine. We are currently evaluating the results produced by the CRS using the information acquired in the taxonomy acquisition process.

- **On the multilingual lexical knowledge acquisition (TGE)**

We tested the TGE by applying two different methodologies to several lexical resources, the first one semi-automatically between DGILE and LDOCE DRINK taxonomies [Ageno et al. 94]. The second one between DGILE FOOD taxonomies and WordNet in a fully automatic way using the notion of conceptual distance developed for sense identification purposes [Rigau et al. 95]. In both cases, we used bilingual dictionaries as a large-scale lexical resource to help the linking process. During the first experiment, using seven informed modules we were able to produce 372 completely disambiguated translations links (of three types) between the Spanish taxonomy of drinks with 235 dictionary senses and the English one with 192 senses. In the second experiment, using the same seven modules as in the previous experiment and the conceptual distance formula, we selected automatically, from a set of possible candidates, the closest dictionary senses to that being linked (following the taxonomic structure). That is, we obtained a single Translation Link from each of the 140 senses of the DGILE FOOD taxonomy.

We also performed several experiments for mapping bilingual dictionaries directly to a Lexical Knowledge Base. That is, we attached Spanish words to WordNet synsets using, as in the previous case, bilingual dictionaries. Combining 17 different methods, we produced a preliminary version of a Spanish WordNet [Atserias et al. 97]. Collecting those synsets with accuracy greater than 85% we obtained a preliminary version of the Spanish WordNet containing 10,982 connections between 7,131 synsets and 8,396 Spanish nouns. However, combining the discarded methods and adding the resulting data to the preliminary version of the Spanish WordNet, we obtained a final Spanish WordNet with 15,535 connections (a 41% increase) between 10,786 synsets and 9,986 Spanish nouns.

- **On the validation and exploitation of lexical knowledge acquired (LDB/LKB system)**

The prototype was tested with a sample Type System and lexicon providing the function we designed (see [Rigau et al. 94] for further details).

- **On lexical knowledge acquired**

From frequency word list to multilingual lexical knowledge lexicons, a large variety of lexicons (in terms of the knowledge and size they contain) have been derived in the thesis. Most of them have been developed as an intermediate result to be applied as a lexical knowledge resource to facilitate one or other methodological steps (part-of-speech lexicons for the morphological analysis, bigrams or trigrams for detecting conceptual patterns, cooccurrence vector for computing similarity of definitions, etc). These lexicons form a set of large complete Spanish MTDs (ready-usable Spanish lexicons) that have been derived from the dictionaries themselves. A large list of intermediate monolingual and bilingual lexicons are provided as an appendix in the final thesis.

7.3 Further work

From the beginning, our methodology has been regarded as being evolutionary and our system as modular, allowing further technical and methodological improvements elsewhere in the acquisition process. This means that the theoretical limit of our work is the acquisition of all lexical knowledge contained in the MRD.

We used a new technique for the genus sense identification task [Rigau et al. 97] based on, among others, the notion of **conceptual distance** in a hierarchical net of concepts. This measure appears as one of the most important tools for facilitating the construction process of large-scale lexicons from MRDs. Although we used this approach successfully to enrich dictionary definitions [Rigau 94], link taxonomies from different languages [Rigau et al. 95], attach bilingual MRDs to preexisting semantic nets [Atserias et al. 97] and attach monolingual senses to preexisting semantic nets [Rigau et al. 98], we think that more in-depth analysis could be performed to obtain better results (for instance, exploring other lexical knowledge measures such as conceptual density [Agirre & Rigau 95], [Agirre & Rigau 96a]).

Another natural extension of the work reported in this thesis involves the issue of combining heuristics for the genus sense identification task. We reported a very successful performance selecting the correct genus sense automatically without having any explicit semantic codes on DGILE definitions (see [Rigau et al. 97]). We achieved this result by combining a set of different informed heuristics and adding the result of one heuristic to the other. We think that better combinations could be adopted to improve the current results.

It is clear that by improving the parsing process on definitions, more in-depth lexical knowledge could be extracted. A wide-range parsing grammar for Spanish is currently under development (e.g., [Climent 97] and [Climent & Moré 97]). We are planning to apply this grammar with a robust chart parsing analyser to all DGILE definitions [Ageno & Rodríguez 96] to obtain a complete syntactic analysis rather than partial ones. Meanwhile, we are improving the consistency and efficiency of the parsing process of dictionary definitions using the SinPar rather than the FPar analyser.

Obviously, the complete taxonomy structure of DGILE could aid the semantic analysis of the definitions and the conversion process from analysed taxonomies to the LKB because of the more in-depth conceptual knowledge acquired. That is, more general inference mechanisms could be made using semantic classes rather than only the words contained in the analysed dictionary definitions. Thus, it is preferable first to obtain the whole set of taxonomies from a dictionary and then to use these taxonomies to extract more in-depth implicit knowledge from dictionary definitions. Currently, we are working in this direction.

Although we have not made any comparison of the performance of the lexical knowledge extracted from MRDs versus other lexicons constructed manually, we have used several lexicons extracted automatically from the MRDs themselves to obtain intermediate results in several steps of our methodology with a very good performance.

The most immediate extension to our work concern the massive acquisition of lexical knowledge from our monolingual and bilingual dictionaries in order to obtain more in-depth semantic knowledge about the Spanish lexicon. Although we have built substantial lexicons

from the monolingual and bilingual MRDs, we are planning to merge all data in the framework of the EuroWordNet project[1]. First, covering WordNet 1.5 with Spanish words using the bilingual dictionaries, and second, completing the sparser parts by attaching Spanish taxonomies extracted from the monolingual dictionary (see [Farreres et al. 98] for further details). This process will provide the lexicographers with a great amount of accurate lexical knowledge extracted automatically from the MRDs.

Although we have focused our attention throughout the thesis on noun lexical knowledge, MRDs also contain lexical knowledge on other part-of-speech categories. A natural extension of the work presented here could be centring the automatic acquisition process on verbs, adjectives and adverbs, and functional words (no-content words). Obviously, each category is described using different schemes and different knowledge can be extracted from them. In conventional dictionaries, verbs are described similarly to nouns, allowing the automatic construction of verb taxonomies, parsing verb definitions and placing the acquired knowledge on the MLKB. An in-depth study on acquisition of verb lexical knowledge from MRDs can be found in [Taulé 95]. Adjective and adverb definitions have no taxonomic structure but the same approach could be taken to analyse them. Other lexical knowledge in the MRD such as idioms, compounds and other lexical items have neither been exploited nor analysed.

A final extension on our work would be to extend the lexical knowledge acquisition process to languages other than Spanish, especially to Catalan, using monolingual and bilingual MRDs such as the *Diccionari Contemporani de la Llengua Catalana*. The first steps in this direction have already been taken (see [Benítez et al. 98]).

[1] *The main aim of EuroWordNet project, LE Reference 4003, is to develop a generic multilingual database with WordNets for several European languages (English, Dutch, Italian and Spanish). The European WordNets will as far as possible be built from available existing resources and Lexical Data Bases with semantic information developed in various projects.*

Dictionaries

CDEL	Collins Dictionary of the English Language.
CIDE	Cambridge International Dictionary of English. Cambridge University Press, Cambridge, United Kingdom, 1995.
DGILE	Diccionario General Ilustrado de la Lengua Española VOX. Alvar M. (ed.). Biblograf S.A. Barcelona, Spain, 1987.
DILEC	Diccionario Ideológico de la Lengua Española J. Casares
DILEV	Diccionario Ideológico de la Lengua Española VOX. Biblograf S.A. Barcelona, Spain, 1995.
EEI	Diccionario VOX/Harrap's Esencial Español/Inglés. Biblograf S.A. Barcelona, Spain, 1992.
EIE	Diccionario VOX/Harrap's Esencial Inglés/Español. Biblograf S.A. Barcelona, Spain, 1992.
LDOCE	Longman Dictionary of Contemporary English. Procter P. et al. (eds). Longman, Harlow and London. 1987.
LLOCE	Longman Lexicon of Contemporary English. MacArthur T. (ed). Longman, Group (Far East) Ltd. Hong Kong, 1992.
NCT	New Collins Thesaurus, Collins, London and Glasgow, 1984,
NDIG	Il Nuovo Dizionario Italiano Garzanti, Garzanti. Milano, 1984.
OALD	Oxford Advanced Learner's Dictionary Hornby (ed.), 1980.
RT	Roget's International Thesaurus (Fourth Edition), Chapman R. , Harper and Row, New York, 1977.
RTII	Roget's II: The New Thesaurus, Houghton Mifflin, Borston, 1980.
VLI	Vocabulario della Lingua Italiana, Zanichelli. Bologna.
W7	Webster's Seventh Collegiate Dictionary.
W7N	Webster's Seventh New Collegiate Dictionary.
NMW	New Merriam-Webster Pocket Dictionary. Pocket Books.

References

[Abney 91] Abney S., *Parsing by Chunks*, In Bernick R., Abney S. and Tenny C. (eds.), Principle-Based Parsing, Kluwer Academic Publishers, 1991.

[Acebo et al. 94] Acebo S., Ageno A., Climent S., Farreres J., Padró L., Ribas F., Rodríguez H. and Soler O., *MACO: Morphologycal Analyzer Corpus-Oriented*, Research Report LSI-94-30-R. Computer Science Department. UPC. Barcelona. 1994. Also as Esprit BRA-7315 Acquilex II Working Paper 31. 1994.

[ACM 95] Lenat D., Miller G. and Yokoi T., Communications of the ACM 38:(11), pages 33-48, 1995.

[Ageno et al. 91a] Ageno A., Castellón I., Martí M.A., Rigau G., Rodríguez H., Taulé M. and Verdejo M.F., *Análisis de las definiciones del diccionario* Vox, in proceedings of the 7th annual meeting of SEPLN. Valencia, Spain, 1991.

[Ageno et al. 91b] Ageno A., Castellón I., Martí M., Rigau G., Rodríguez H., Taulé M. and Verdejo M., *SEISD: An environment for extraction of semantic information from on-line dictionaries*, Research Report LSI-91-33. Computer Science Department. UPC. Barcelona. 1991.

[Ageno et al. 92a] Ageno A., Cardoze S., Castellón I., Martí M.A., Rigau G., Rodríguez H., Taulé M. and Verdejo M.F. *Un entorno para la extracción de información semántica de diccionarios.* in proceedings of the Annual Meeting of the Asociación Española para la Inteligencia Artificial (AEPIA'91). Madrid, 1991. published also in Novática 18(98). 1992.

[Ageno et al. 92b] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M. and Verdejo F., *SEISD: An environment for extraction of Semantic information from on-line dictionaries,* in proceedings of the 3th Conference on Applied Natural Language Processing (ANLP'92), Trento, Italy. 1992.

[Ageno et al. 92c] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M. and Verdejo F., *A semiautomatic process to create LKB entries*, Esprit BRA-3030 Acquilex I Working Paper 38. 1992.

[Ageno et al. 92d] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M. and Verdejo F., *From the LDB to the LKB,* Esprit BRA-3030 Acquilex I Working Paper 39. 1992.

[Ageno et al. 93] Ageno A., Ribas F., Rigau G., Rodríguez H. and Verdejo F., *TGE: Tlink Generation Environment*, Esprit BRA-7315 Acquilex II Working Paper. 1993.

[Ageno et al. 94] Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H. and Samiotou A., *TGE: Tlink Generation Environment.* in proceedings of the 15th International Conference on Computational Linguistics (COLING'94). Kyoto, Japan. 1994.

[Ageno & Rodríguez 96] Ageno A., and Rodríguez H., *Using Bidirectional Chart Parsing for Corpus Analysis*, LSI-96-12-R, Computer Science Department. UPC. Barcelona. 1996.

[Agirre et al. 94] Agirre E., Arregi X., Artola X., Díaz de Ilarraza A. and Sarasola K., *Conceptual Distance and Automatic Spelling Correction*, proceedings of the workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, 1994.

[Agirre & Rigau 95] Agirre E. and Rigau G., *A Proposal for Word Sense Disambiguation using Conceptual Distance*, in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'95). Tzigov Chark, Bulgaria. 1995.

[Agirre & Rigau 96a] Agirre E. and Rigau G., *Word Sense Disambiguation using Conceptual Density*, in proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark. 1996.

[Agirre & Rigau 96b] Agirre E. and Rigau G., *An Experiment in Word SenseDisambiguation of the Brown Corpus Using WordNet*. Memoranda in Computer and Cognitive Science, MCCS-96-291, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico. 1996.

[Alshawi 89] Alshawi H., *Analysing the Dictionary Definitions*, in Boguraev B. and Briscoe E. eds. Computational Lexicography for NLP, chaper 7. Longman, London. 1989.

[Alshawi 92] Alshawi H., The Core Language Engine, Alshawi H. (ed.) ACL-MIT Press Series in Natural Language Processing. 1992.

[Alvar & Villena 94] Alvar M. and Villena J., Estudios para un Corpus del Español, Universidad de Málaga. Malaga, Spain, 1994.

[Amsler 81] Amsler R., *A Taxonomy for Enghish Nouns and Verbs,* in proceedings of the 19th Annual Meeting of tha Association for Computational Linguistics, (ACL'81), pages 133-138, Stanford, California, 1981.

[Anick & Pustejovsky 90] Anick P. and Pustejovsky J., *An Application of Lexical Semantics to Knowledge Acquisition from Corpora*, in proceedings of the 13th International Conference on Computational Linguistics (COLING'90). pages 7-12, Helsinki, Finland, 1990.

[Appelt et al. 93] Appelt D., Hobbs J.,Bear J., Israel D., Kameyama M. and Tyson M., *Fastus: A Finite-state Processor for Information Extraction from Real-world Text*, in proceedings of International Joint Conference of Artifical Inteligence (IJCAI'93). pages 1172-1178. Chambery, France. 1993.

[Arranz et al. 95] Arranz M., Radford I., Ananiadou S. and Tsuji J., *Towards a Sublanguage-Based semantic Clustering Algorithm*, in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'95), pages 110-117. Tzigov Chark, Bulgaria. 1995.

[Artale et al. 97] Artale A., Magnini B. and Strapparava C., *Lexical Discrimination with the Italian Version of WordNet*, in proceedings of a ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources. Madrid, Spain. 1997.

[Artola 93] Artola X. *Conception et construccion d'un systeme intelligent d'aide diccionariale (SIAD)*. Ph.D. Thesis, Euskal Herriko Unibertsitatea, Donostia, 1993.

[Atkins et al. 86] Atkins B. Kegl T. and Levin B., *Explicit and Implicit Information in Dictionaries,* in proceedings of the Second Annual Conference of the UW Centre for the New OED, Waterloo, Canada, 1986.

[Atkins & Levin 88] Atkins B. and Levin B., *Admiting Impediments,* in proceedings of the 4th Annual Conference of the UW Centre for the New OED, Waterloo, Canada, 1988. Also published in Zernik U. (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Atkins et al. 92] Atkins S., Clear J. and Ostler N., Corpus Design Criteria. Literary and Linguistic Computing 7:1-16, 1992.

[Atserias et al. 97] Atserias J., Climent S., Farreres X., Rigau G. and Rodríguez H., *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*, in proceedings of of International Conference "Recent Advances in Natural Language Processing" (RANLP'97). Tzigov Chark, Bulgaria, 1997.

[Atserias et al. 98] AtseriasJ, Castellón I and Civit M., *Syntactic Parsing of Unrestricted Spanish Text*. In Proceedings of 1st International Conference on Language Resources and Evaluation, LREC'98. Granada, Spain. 1998.

[Barrière & Popowich 96] Barrière C. and Popowich F., *Concept Clustering and Knowledge Integretion from children's dictionary*, in proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark. 1996.

[Basili et al. 92a] Basili R., Pazienza M. and Velardi P., *Computational Lexicons: the Neat Examples and the Odd Exemplars*, in proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92), pages 96-103. Trento, Italy. 1992.

[Basili et al. 92b] Basili R., Pazienza M. and Velardi P., *Combining NLP and Statistical Techniques for Lexical Acquisition*, in proceedings of the AAAI Fall Symposium on Statistically-Based NLP Techniques. 1992.

[Basili et al. 95] Basili R., Della Rocca M., Pazienza M. and Velardi., *Contexts and categories: tuning a general purpose verb classification to sublanguages,* in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'95), pages 118- 127. Tzigov Chark, Bulgaria, 1995.

[Bateman 90] Bateman J., *Upper modeling: Organizing knowledge for Natural Language Processing*. in proccedings of Fifth International workshop on Natural Language Generation, Pittsburg, PA. 1990.

[Benítez et al. 98] Benítez L., Cervell S., Escudero G., López M., Rigau G., Taulé M., *Methods and Tools for Building the Catalan WordNet*, in proceedings of the ELRA Workshop on Language Resources for European Minority Languages, Granada, Spain. 1998.

[Boguraev & Briscoe 87] Boguraev B.K. and Briscoe E.J. , *Large Lexicons for Natual Language Processing: Utilising the Grammar Coding system of the Longman Dictionary of Contemporary English*, Computational Linguistics 13(4):219-240. 1987.

[Boguraev & Briscoe 89a] Boguraev B. and Briscoe T. editors, *Computational Lexicography for Natural Language Processing*. Longman, Cambridge, England. 1989.

[Boguraev & Briscoe 89b] Boguraev B. and Briscoe T., *Utilising the LDOCE grammar codes*, in Boguraev B and Briscoe T., eds. *Computational Lexicography for Natural Language Processing*. Longman, Cambridge, England. 1989.

[Boguraev & Pustejovsky 90] Boguraev B. and Pustejovsky J., *Lexical Ambiguity and The Role of Knowledge Representation in Lexical Design*, in proceedings of the 13th International Conference on Computational Linguistics (COLING'90). pages 36-41. Helsinki, Finland, 1990.

[Boguraev et al. 91] Boguraev B., Briscoe E., Carroll J. and Copestake A., *Database Models for Computational Lexicography*, IBM T.J. Watson Research Department RC 17120. Also in proceedings of 4th Euralex International Congress (Euralex'90). pages 59-78. Benalmádena, Spain, 1990.

[Boguraev & Pustejovsky 95] Boguraev B. and Pustejovsky J., Corpus Processing for Lexical Acquisition. The MIT Press. Cambridge, Massachusetts. 1995.

[Brill 92] Brill E., *A simple rule-based part of speech tagger*, in proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92), pages 152-155, Trento, Italy. 1992.

[Briscoe et al. 90] Briscoe E., Copestake A. and Boguraev B., *Enjoy the paper: Lexical Semantics via lexicology*, in proceedings of the 13th International Conference on Computational Linguistics (COLING'90),pages 42-47. Helsinki, Finland, 1990.

[Briscoe 91] Briscoe E., *Lexical Issues in Natural Language Processing*. In E. Klein and F.. Veltman (eds.), Natural Language and Speech, pages 39-68, Springer-Verlag. 1991.

[Briscoe & Carroll 91] Briscoe E. and Carroll J., *Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars*, Technical Report No. 224, Computer Laboratory, University of Cambridge, Cambridge, UK, 1991.

[Briscoe & Carroll 93] Briscoe E. and Carroll J., *Generalised probabilistic LR parsing of natural language corpora with unification-based grammars*, Computational Linguistics 19(1), 1993. Also as Technical Report No 224, University of Cambrige, Computer Laboratory. Cambrige, UK, 1993.

[Briscoe & Carroll 95] Briscoe E. and Carroll J., *Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels*. In Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies, pages 48-58. Prague, Czech Republic, 1995.

[Briscoe & Carroll 97] Briscoe T. and Carroll J., *Automatic Extraction of subcategorization from Corpora*, in proceedings of 5th International Conference on Applied Natural Language Processing (ANLP'97). Washington, DC, 1997.

[Brown et al. 91a] Brown P., Della Pietra S., Della Pietra V. and Mercer R., *Word-Sense Disambiguation using Statistical Methods*, in proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pages 264-270. Berkeley, California, 1991.

[Brown et al. 91b] Brown P., Lai J. and Mercer R., *Aligning sentences in parallel corpora*, in proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pages 169-176. Berkeley, California, 1991.

[Brown et al. 92] Brown P., deSouza P., Mercer R., Della Pietra V. and Lai J., *Class-based n-gram models of natural language*, Computational Linguistics, 18(4). 1992.

[Brownston et al. 86] Brownston L., Farrell R., Kant, E. and Martin N., *Programming Expert Systems in OPS5.* Addison-Wesley. 1986.

[Bruce & Guthrie 91] Bruce R. and Guthrie L., *Bulding a Noun Taxonomy from a Machine Readable Dictionary*, Research Report MCCS-91-207. Computing Research Laboratory, New Mexico State University, Las Cruces, NM, 1991.

[Bruce & Guthrie 92] Bruce R. and Guthrie L., *Genus disambiguation: A study in weigthed preference,* in proceedings of the 14th International Conference on Computational Linguistics (COLING'92). Nantes, France, 1992.

[Bruce et al. 92] Bruce R., Wilks Y., Guthrie L., Slator B. and Dunning T., *NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour*, Research Report MCCS-92-246. Computing Research Laboratory, New Mexico State University. Las Cruces, NM, 1992.

[Bruce & Wiebe 94] Bruce R. and Wiebe J., *Word Sense DIsambiguation Using Decomposable models*, in proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94). Las Cruces, New Mexico, 1994.

[Byrd 89] Byrd R., *Discovering Relationship among Word Senses*, in proceedings of the 5th Annual Conference of the UW Centre for the New OED, pages 67-79. Oxford, England. 1989.

[Calzolari 88] Calzolari N., *The dictionary and the thesaurus can be combined*, in M. Evens (ed.), Relational Models of the Lexicon, Studies in NLP, pages 75-96. Cambridge University Press, 1988.

[Calzolari & Bindi 90] Calzolari N. and Bindi R., *Acquisition of lexicalinformation from a large textual italian corpus,* in proceedings of the 13th International Conference on Computational Linguistics (COLING'90). Helsinki, Finland. 1990.

[Calzolari 91] Calzolari N., *Acquiring and Representing Semantic Information in a Lexical Knowledge Base*, in proceedings of the Workshop oon Lexical Semantics, Berkeley. Also in Esprit BRA-3030 Acquilex I Working Paper 16. 1991.

[Calzolari et al. 93] Calzolari N., Cotoneschi P., Montemagni S. and Spanu A., *Extraction and normalization of noun taxonomical chains to create a thesaurical set for sense disambiguation tasks*, Esprit BRA-7315 Acquilex II Working Paper 15. 1993.

[Carlson & Niremburg 90] Carlson L. and Niremburg S., *World modeling for NLP*. Technical Report CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University. 1990.

[Carmona et al. 98] Carmona J., Cervell S., Márquez L., Martí M.A., Padró L., Placer R., Rodríguez H., Taulé M. and Turmo J. *An Environment for Morphosyntactic Processing of Unrestricted Spanish Text*. In Proceedings of 1st International Conference on Language Resources and Evaluation}, LREC'98. Granada, Spain. 1998.

[Carroll & Grover 89] Carroll J. and Grover C., *The derivation of a large computational lexicon for English from LDOCE*, in Boguraev B. and Briscoe E. (ed.), Computational Lexicography for Natural Language Processing, pages 117-134, Longman, London. 1989.

[Caroll 90a] Carroll J. *Lexical Database System. User Manual*. Deliverable 2.3.3(a), Esprit Bra 3030. Computer Laboratory. University of Cambridge.

[Carroll 90b] Carroll J., *Flexible Pattern Matching Parsing Tool (FPar) Technical Manual*, Esprit BRA-3030 Acquilex I Six-Month Deliverable Release. 1990.

[Carroll 92] Carroll J., *The Acquilex lexical database system: system description and user manual*, in Sanfilippo A. (ed.), The (other) Cambridge Acquilex Papers, Technical Report No. 253, pages 24-48, University of Cambridge, Computer Laboratory, Cambridge, UK, 1992.

[Carpenter 92] Carpenter B, The Logic of Typed Feature Structures, Cambridge University Press, Cambridge, England, 1992.

[Castellón et al. 91] Castellón I., Martí M.A., Rigau G., Rodríguez H. and Verdejo V., *Loading MRD into LDB. Characteristics of the Vox Dictionary*. Research Report LSI-91-24. Computer Science Department, UPC, Barcelona, 1991.

[Castellón 93] Castellón I., *Lexicografia Computacional: Adquisición Automática de Conocimiento Lèxico,* Ph.D. Thesis, Universitat de Barcelona, Barcelona, 1993.

[Cavazza & Zweigenbaum 95] Cavazza M. and Zweigenbaum P., *Lexical Semantics: Dictionary or encyclopedia?*, Patrick Saint-Dizier & Evelyne Viegas (eds.) Computational Lexical Semantics, Cambridge University Press. Cambridge, UK. 1995.

[Charniak 93] Charniak E., Statistical Language Learning. The MIT Press. Cambridge, Massachusetts. 1993.

[Chen & Chang 98] Chen J. and Chang J., *Topical Clustering of MRD Senses Based on Information Retrieval Techniques*. Computational Linguistics, Special Issue on Word Sense Disambiguation, 24(1)61:95.

[Chodorow et al. 85] Chodorow M.S., Watson T.J. and Byrd R.J. *Extracting Semantic Hierarchies from a Large on-line Dictionary*, in proceedings of the 23th Annual Meeting of Association for Computational Linguistics, pages. 299-304. 1985.

[Chomsky 70] Chomsky N., *Remarks on Nominalization*, in Jacobs R. and Rosenbaum P. (eds.), Readings in English Transformational Grammar, Ginn, Watham, Mass, 1970.

[Church & Hanks 90] Church K., and Hanks P., *Word association norms, mutual information, and leogicography*, Computational Linguistics, vol. 16, ns. 1, 22-29. 1990. Also in proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89). Pittsburg, Pennsylvania, 1989.

[Church et al. 91] Church K., Gale W., Hanks P. and Hindle D., *Using Statistics in Lexical Analysis*, in Zernik U. (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Church 93] Church K., *Char_align: A program for aligning parallel texts at the character level*, in proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (ACL'93). Columbus, Ohio, 1993.

[Climent 97] Climent S., *CHAOS (Chunk Analyser of Spanish)*, Report UB-LG.1997-1, Sección de Linguistica General, Facultad de Filologia, Universidad de Barcelona. Barcelona. 1997.

[Climent & Moré 97] Climent S. and Moré J., *(DC)GATOS. A definite clause grammar to parsecorpora of Spanish*, Report UB-LG.1997-2 Sección de Linguistica General, Facultad de Filologia, Universidad de Barcelona. Barcelona. 1997.

[Coates-Stephens 92] Coates-Stephens S., The Analysis and Acquisition of Proper Names for Robust Text Understanding, Ph.D. Thesis, Departament of Computer Science of City University, London, England. 1992.

[Cohen & Loiselle 88] Cohen P. and Loiselle C., *Beyond ISA: Structures for Plausible Inference in Semantic Data*, in proceedings of 7th Natural Language Conference AAAI'88. 1988.

[Copestake 90] Copestake A., *An approach to building the hierarchical element of a lexical kwoledge base from a machine readable dictionary*, in proceedings of the Fisrt Intenational Workshop Inheritance in NLP ,Tilburg, Netherlands, pages 19-29, 1990.

[Copestake 92a] Copestake A., *The Acquilex LKB: represention issues in semi-automatic acquisition of large lexicons*. in proceedings of 3th Conference on Applied Natural Language Processing (ANLP'92), Trento. Italy. 1992.

[Copestake 92b] Copestake A., The Representation of Lexical Semantic Information, Ph.D. thesis, Cognitive Science Research Report Paper 280, University of Sussex at Brighton. September 1992.

[Copestake et al. 92] Copestake, A., Jones B., Sanfilippo A., Rodríguez H. and Vossen P., *Multilingual Lexical Representation*. Esprit BRA-3030 Acquilex Working Paper nº38. 1992.

[Copestake et al. 94] Copestake A., Briscoe T., Vossen P., Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., Samiotou A., *Acquisition of Lexical Translation Relations from MRDs*.. Esprit BRA-7315 Acquilex-II Working Paper Nº.040. September 1994. Also as Research Report LSI-94-35-R. Computer Science Department, UPC, Barcelona, 1994. Also in Machine Translation: Special Issue on the lexicon, 9:3,33-69.

[Cottrel & Small 89] Cottrel G. and Small S., *A Connectionist Scheme for Modelling Word Sense Disambiguation*, Cognition and Brian Theory, 6(1), 89-120. Also in Neural Networks for NLP, Fourth European Summer School in Logic, Language and Information. University of Essex, Chochester, UK, August, 1992.

[Cowie et al. 92] Cowie J., Guthrie J. and Guthrie L., *Lexical Disambiguation using Simulated annealing,* in proceedings of DARPA WorkShop on Speech and Natural Language, 238-242. New York, February 1992.

[Dagan et al. 91] Dagan I., Itai A. and Schwall U., *Two languages are more informative than one*, in proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pages 130-137. Berkeley, California, 1991.

[Dagan et al. 94] Dagan I., Pereira F. and Lee L., *Similarity-based Estimation of Word Co-occurrences Probabilities*, in proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94). Las Cruces, New Mexico, 1994.

[Dagan et al. 97] Dagan I., Lee L. and Pereira F., *Similarity-Based Methods for Word Sense Disambiguation*, in proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), pages 56-63. Madrid, Spain, 1997.

[Dietterich 97] Dietterich T., *Machine Learning Research: Four Current Directions*, AI Magazine 18(4), 97-136. 1997.

[Dolan et al. 93] Dolan W., Vanderwende L. and Richard son S., *Automatically deriving structured knowledge bases from on-line dictionaries.* in proceedings of the first Conference of the Pacific Association for Computational Linguistics (Pacling'93), April 21-24, Simon Fraser University, Vancouver, Canada. 1993.

[Dolan 94] Dolan W., *Word Sense Clustering: Clustering Related Senses*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94), pages 712-716. Kyoto, Japan, 1994.

[Eijk 93] van der Eijk, P., *Automating the Acquisition of Bilingual Terminology*, in proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93), pages 113-119. 1993.

[Eizirik et al. 93] Eizirik L., Barbosa V. and Mendes S., *A Bayesian-Network Approach to Lexical Disambiguation*, Cognitive Science 17, pages 257-283. 1993.

[Farreres et al. 98] Farreres X., Rigau G. and Rodríguez H., *Using WordNet for Building WordNets*. In Proceedings of COLING-ACL Workshop "Usage of WordNet in Natural Language Processing Systems". Montreal, Canada. 1998.

[Farwell et al. 95] Farwell D., Helmreich S. and Casper M., *SPOST: a Spanish Part-of-Speech Tagger*, in proceedings of the 11th Annual Meeting of SEPLN, Deusto, Spain. 1995.

[Fernández 95] Fernandez, A., *Mismatches and lexical gaps in Verb entries,* Esprit BRA 7315 Acquilex II Working Paper. 1995.

[Firth 57] Firth J., *A synopsis of linguistic theory 1930-1955*, in Selected Papers of J.R. Firth, Plamer F. (ed.), Longman, London, 1957.

[Fischer 97] Fischer D., *Formal redundancy and inconsistency checking rules for the lexical database WordNet1.5*, in proceedings of a ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, pages 22-31. Madrid, Spain. 1997.

[Francis & Kucera 82] Francis W. and Kucera H., *Frequency Analysis of English Usage*. Hougton Mifflin Company, Boston, Massachusetts, 1982.

[Fox et al. 88] Fox E., Nutter T., Ahlswede T. and Evens M., *Building a Large Thesaurus for Information Retrieval*, in proceedings of the Second Conference on Applied Natural Language Processing (ANLP'88), pages 101-108, Austin, Texas, 1988.

[Fukumoto & Suzuki 96] Fukumoto F. and Suzuki Y., *An Automatic Clustering of Articles Using Dictionary Definitions*, in proceedings of the 16th International Conference on Computational Linguistics (COLING'96). Copenhagen, Denmark. 1996.

[Fung 95] Fung P., *A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora*, in proceedings of the 33th Annual Meeting of tha Association for Computational Linguistics, (ACL'95). 1995.

[Gale & Church 91] Gale W. and Church K., *Identifying Word Correspondences in Parallel Texts*, in proceedings of DARPA Workshop on Speech and Natural Language, pages 152-157. Pacific Grove, CA, 1991.

[Gale et al. 92a] Gale W., Church K. and Yarowsky D., *One Sense Per Discourse*, in proceedings of DARPA WorkShop on Speech and Natural Language, pages 233-237. New York, 1992.

[Gale et al. 92b] Gale W., Church K. and Yarowsky D., *Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs*, in proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92), pages 249-256. 1992.

[Gale et al. 92c] Gale W., Church K. and Yarowsky D., *On Evaluation of Word-Sense Disambiguation Systems*, in proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92). 1992.

[Gale et al. 93] Gale W., Church K. and Yarowsky D., *A Method for Disambiguating Word Senses in a Large Corpus*, Computers and the Humanities 26, pages 415-439. 1993.

[Gatius & Rodríguez 96] Gatius M. and Rodríguez H., *A Domian-restricted task-oriented NLI to System Experts*, in proceedings of the Flexible Query Answer Systems. Rockilde, Denmark. 1996.

[Gazdar et al. 85] Gazdar G., Klein E., Pullum G. and Sag I., *Generalized Phrase Structure Grammar*, Blackwell, Oxford and Hardvard University Press, Cambridge, MA. 1985.

[Gomez et al. 94] Gomez F., Hull R. and Segami C., *Acquiring Knowledge From Encyclopedic Texts*, in proceedings of the 4th Conference on Applied Natural Language Processing, (ANLP'94), pages 84-90, Stuttgart, Germany, October, 1994.

[Grefenstette & Hearst 92] Grefenstette G. and Hearst M., *A method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results*, in proceedings of the AAAI Spring Symposium on Statistically-Based NLP Techniques. San Jose, California. 1992.

[Grefenstette 92a] Grefenstette G., *Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis*, in proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92) Student Session. pages 324-326. 1992.

[Grefenstette 92b] Grefenstette G., *Finding Semantic Similarity in Raw Text: the Deese Antonyms*, in proceedings of the AAAI Fall Symposium on Statistically-Based NLP Techniques. 1992.

[Grefenstette 93] Grefenstette G., *Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and window Based Approaches*, in proceedings of SIGLEX Workshop on Acquisition of Lexical Knowledge from Text. Columbus, OH. 1993.

[Grishman et al. 94] Grishman R., Macleod C. and Meyers A., *Comlex syntax: building a computational lexicon*. In Proceedings of the 15th Annual Meeting of the Association for Computational Linguistics. (COLING'94). 268-272. Kyoto, Japan. 1994.

[Grishman & Sundheim 96] Grishman R. and Sundheim B., *Message Understanding Conference - 6: A Brief History*, in proceedings of the 16th Annual Meeting of the Association for Computational Linguistics. (COLING'96). 466-470. Copenhage, Denmark. 1996.

[Guthrie et al. 91] Guthrie J., Guthrie L., Wilks Y. and Aidinejad H., *Subject-dependent Co-occurrence and Word Sense Disambiguation*, in proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pages 146-152. Berkeley, California, 1991.

[Guthrie et al. 93] Guthrie L., Guthrie J. and Cowie J., *Resolving Lexical Ambiguity*, in Memoranda in Computer and Cognitive Science MCCS-93-260, Computing Research Laboratory, New Mexico State University. Las Cruces, New Mexico. 1993.

[Guthrie et al. 96] Guthrie L., Pustejovsky J., Wilks Y. and Slator B., *The Role of Lexicons in Natural Language Processing*, Communications of the ACM, Vol. 39:1, pages 63-72. 1996.

[Grover et al. 93] Grover C., Carroll J. and Briscoe E., *The Alvey Natural Language Tools grammar (4th realese)*. Technical Report 284. Computer Laboratory, Cambridge University, UK. 1993.

[Hagman 92] Hagman J., *Semantic Parsing of Italian Dictionary Definitions*, Esprit BRA-3030 Acquilex I Working Paper 47. May 1992.

[Harley & Glennon 97] Harley A. and Glennon D., *Sense Tagging in Action*, in proccedings of the SIGLEX WorkShop Tagging Text with Lexical Semantics: Why, What and How?. Washington. DC, 1997.

[Hastings et al. 94] Hastings A., Rigau G., Soler C. and Tuells T., *Loading a Bilingual Dictionary into the LDB*., Research Report LSI-94-2-T. Computer Science Department. UPC. Barcelona. 1994.

[Hayes 77] Hayes P., *Some Association-based Techniques for Lexical Disambiguation by Machine*, Ph.D. Thesis, published as Technical Report No. 25 Dept. of Computer Science, University of Rochester, 1977.

[Hearst 92] Hearst M., *Automatic acquistion of hyponyms from large text corpora*, in proceedings of the 14th International Conference of Computational Linguistics (COLING'92). Nantes, France. 1992.

[Hearst & Schütze 95] Hearst M. and Schütze H., *Customizing a Lexicon to Better Suit a Computational Task*, in Boguraev B. and Pustejovsky J. (eds.) Corpus Processing for Lexical Acquisition, The MIT Press, Cambridge, Massachusetts, 1995.

[Hindle & Rooth 93] Hindle D. and Rooth M., *Structural ambiguity and lexical relations*, Computational Linguistics, 19(1), pages 103-120. 1993.

[Hirst 88] Hirst G., *Resolving Lexical Ambiguity Computationally with Spreading Activation and Polaroid Words*, in Lexical Ambiguity Resolution, Small Cotrell and Tannenhaus (Eds.), Norman Kaufmann Press, 1988.

[Hirst 95] Hirst G., *Near-synonymy and the structure of lexical knowledge*, in proceedings of the AAAI'95 Spring Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativy, pages 51-57. 1995.

[Hobbs et al. 93] Hobbs J., Appelt D., Bear J., Israel D., Kameyama M. and Tyson M., *Fastus: A System for Extracting Information from Text*, in proceedings of the ARPA Workshop on Human Language Technology, pages 133-137. Princeton. 1993.

[Hovy & Nirenburg 92] Hovy E. and Nirenburg S., *Approximating an Interlingua in a principled way*, in proceedings of DARPA Workshop on Speech and Natural Language. Harriman, NY, 1992.

[Hovy & Knight 93] Hovy E. and Knight K., *Motivating Shared Knowledge Resources: An Example from the Pangloss Collaboration*, in proceedings of the IJCAI Workshop on Shared Knowledge (IJCAI'93), Chambery, France. 1993.

[Hudson 84] Hudson D., Word Grammar, Blackwell, Oxford. 1984.

[Ide et al. 91] Ide N., Maitre J. and Véronis J., *Outline of a Model for Lexical Databases*, Technical Report 496, Centre National de la Research Scientifique, Marseille, France. 1991.

[Ide and Véronis 95] Ide N. and Véronis J., (Eds.), *The Text Encoding Iniciative: background and context*. Triple Special Issue of Computers and the Humanities, 29.

[Jacobs 91] Jacobs P., *Making Sense of Lexical Acquisition*, in Zernik U. (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Jensen 86] Jensen K., *Parsing strategies in a broad-coverage grammar of English*, IBM Research Report RC12147, IBM T.J. Watson Research Center, Yorktown Heights, New York. 1986.

[Jensen & Binot 87] Jensen K. and Binot J., *Disambiguating prepositional phrase attachments by using on-line dictionary definitions*, Computational Linguistics 13, (3-4), pages 251-260. 1987.

[Jensen & Binot 88] Jensen K and Binot J., *Dictionary Text Entries as a Source of Knowledge for Syntactic and other Disambiguations*, in proceedings of the Second Conference on Applied Natural Language Processing (ANLP'88), pages 152-159. Austin, Texas, 1988.

[Jiang & Conrath 97] Jiang J. and Conrath D., *Semantic Similarity on Corpus statistics and Lexical Taxonomy*, in proceedings of 10th International Conference Research on Computational Linguistics (ROCLING'97). Taiwan, 1997.

[Jorgenssen 90] Jorgenssen J., *The Psychological Reality of Word Senses*, Journal of Psycholinguistic Research, Vol. 19, No. 13, pages 167-190. 1990.

[Kaplan 50] Kaplan A., *An Experimental Study of Ambiguity in Context*, Cited in Mechanical Translation, 1, 1--3. 1950.

[Karlsson et al. 95] Karlsson F., Voutilainen A., Heikklä J., and Anttila A., Constraint Grammar. A lenguage-independent system for parsing unrestricted text. Mouton de Gruyter. Berlin/New York. 1995.

[Karov & Edelman 96] Karov Y. and Edelman S., Learning Similarity-Based Word Sense Disambiguation from Sparce Data, Research Report CS-TR-96-05 The Weizmann Institute of Science, Rehovot, Israel. 1996.

[Kasahara et al. 95] Kasahara K., Matsuzawa K., Ishikawa T. and Kawaoka T., *Viewpoint-Based Measurament of Semantic Similarity between Words*, in proocedings of the WorkShop on AI and Statistics, pages 292-302, Lauderdale, Florida. 1995.

[Kilgarriff 93] Kilgarriff A., *Dictionary Word Sense Distinctions: An Enquiry Into Their Nature*, Computers and the Humanities 26:365-387, 1993.

[Kilgarriff 97] Kilgarriff A., *I don't believe in word senses*, Computers and the Humanities, 31 (2). 1997.

[Klavans et al. 90] Klavans J., Chodorow M. and Wacholder N., From dictionary to knowledge base viataxonomy, in proceedings of the Sixth Conference of theUniversity of Waterloo. Centre of hte New Oxford English Dictionary and Text Research: Electronic Text Research.

[Klavans & Tzoukermann 96] Klavans J. and Tzoukermann E., *Combining Corpus and Machine-Readable Dictionary for Buliding Lexicons,* Machine Translation, 10:3-4, pages 1-34. 1996.

[Knight 93] Knight K., *Building a Large Ontology for Machine Trasnlation*, in proceedings of the ARPA Workshop on Human Language Technology, pages 185-190, Princeton. 1993.

[Knight & Luk 94] Knight K. and Luk S., *Building a Large-Scale Knowledge Base for Machine Translation*, in proceedings of the American Association for Artificial Inteligence. 1994.

[Kozima & Furugori 93] Kozima H. and Furugori T., *Similarity between Words Computed by Spreading Activation on an English Dictionary*, in proceedings of the of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL'93), pages 232-239. 1993.

[Kozima & Ito 95] Kozima H. and Ito A., *Context-Sensitive Measurament of Word Distance by Adaptative Scaling of Semantic Space*, in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'95), Tzigov Chark, Bulgaria. 1995.

[Krovetz & Croft 92] Krovetz R. and Croft B., *Lexical Ambiguity and Information Retrieval*, ACM Transactions on Information Systems, 10:2, pages 115-141. April 1992.

[Kumano & Hirakawa 94] Kumano A. and Hirakawa H., *Building an MT Dictionary from parallel Texts Based on Linguistic and Statistical Information*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94), pages 76-81. Kyoto, Japan. 1994.

[Kupiec 93] Kupiec J., *An algorithm for finding noun phrase correspondances in bilingual corpora*, in proceedings of the 31th Annual Meeting of tha Association for Computational Linguistics, (ACL'93), pages 17-22. Columbus, Ohio, 1993.

[Leacock et al. 95] Leacock C., Towell G. and Voorhees E., *Towards Building Contextual Representations of Word Senses Using Statistical Models*, in Boguraev B. and Pustejovsky J. (eds.) Corpus Processing for Lexical Acquisition, The MIT Press, Cambridge, Massachusetts, 1995.

[Lesk 86] Lesk M., *Automatic sense disambiguation: how to tell a pine cone from an ice cream cone*, in Proceeding of the 1986 SIGDOC Conference, Association for Computing Machinery, New York, 1986.

[Lenat & Guha 90] Lenat D. and Guha R., Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project. Addison Wesley. 1990.

[Lenat 95] Lenat D., *Steps to Sharing Knowledge*, Towards Very Large Knowledge Bases, Mars N. (Ed.). IOS Press. 1995.

[Li & Abe 95] Li H. and Abe N., *Generalizing Case Frames Using a Thesaurus and the MDL Principle*, in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'95), pages 239-248, Tzigov Chark, Bulgaria. 1995.

[Liddy & Paik 92] Liddy E. And Paik W., *Statistically-Guided Word Sense Disambiguation*, in proceedings of the AAAI Fall Symposium on Statistically-Based NLP Techniques. 1992.

[Marcus et al. 93] Marcus M., Santorini B. and Marcinkiewicz M. *Building a large annotated corpus of english: The Penn Treebank.* Computational Linguistics, 19(1). 1993.

[Marcus et al. 94] Marcus M., Kim G., Marcinkiewicz M., MacIntyre R., Bies A., Ferguson M., Katz K. and Schasberger B. *The Penn Treebank: Annotating predicate argument structure*, in proceedings of ARPA Workshop on Human Language Technology. 1994.

[Markowitz et al. 86] Markovitz J., Ahlswede T. and Evens., *Semantically Significant Patterns in Dictionary Definitions*, in proceedings of the 24th Annual Meeting of the Association for COmputational Linguistics (ACL'86), pages 10-13. 1986.

[McNaught 90] McNaught J., *Reusability of Lexical and Terminological Resources; Steps towards the independence*, in proceedings of the International Worshop on Electronic Dictionaries, pages 97-107, OISO, Kanagawa, Japan. 1990.

[McRoy 92] McRoy S., *Using Multiple Knwoledge Sources for Word Sense Discrimination*, Computational Linguistics 18(1), March, 1992.

[Meijs 90] Meijs M., *The Expanding Lexical Universe: Extracting taxonomies from Machine Readable Dictionaries*, Esprit BRA-3030 Acquilex I Working Paper 26. 1990.

[Miller 90] Miller G., *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

[Miller & Teibel 91] Miller G. and Teibel D., *A proposal for Lexical Disambiguation,* in Proceedings of DARPA Speech and Natural Language Workshop, 395-399, Pacific Grave, California. February, 1991

[Miller et al. 93] Miller G. Leacock C., Randee T. and Bunker R. *A Semantic Concordance,* in proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey, March, 1993.

[Miller et al. 94] Miller G., Chodorow M., Landes S., Leacock C. and Thomas R., *Using a Semantic Concordance for sense Identification,* in proceedings of ARPA Workshop on Human Language Technology, pages 232-235. 1994.

[Miller 95] Miller G., *Building Semantic Concordances: Disambiguation vs. Annotation,* in proceedings of the AAAI'95 Spring Symposium Series, working notes of the Symposium on Representation and Acquisition of Lexical Knowledge, pages 92-94. 1995.

[Nakamura & Nagao 88] Nakamura J. and Nagao M., *Extraction of Semantic Information from an Ordinary English Dictionay and its Evaluation,* in proceedings of the 12th International Conference on Computational Linguistics (COLING'88), pages 459-464, Budapest, Hungria, 1988.

[Nani & MacMillan 95] Nani I. and MacMillan R., *Identifiying Unknown Proper Names in Newswire Text*, in Boguraev B. and Pustejovsky J. (eds.) Corpus Processing for Lexical Acquisition, The MIT Press, Cambridge, Massachusetts, 1995.

[Neff et al. 93] Neff M., Blaser B., Lange J-M., Lehmann H. and Dominguez. *Get it where you can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation*, Paper presented at the AAAI Spring Simposium on Building Lexicons for Machine Translation, Stanford University. 1993.

[Ng and Lee 96] Ng H. and Lee H., *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*, in proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96). 1996.

[Nirenburg & Raskin 87] Nirenburg S. and Raskin V.,*The Subworld Concept Lexicon and the Lexicon Management System,* in International Journal of Computational Linguistics. Vol. 13(3-4), 276-289,1987.

[Nirenburg & Defrise 93] Nirenburg S. and Defrise C., *Aspects of text meaning*, in Pustejovsky J. ed. Semantics and the Lexicon, Dordrecht, Kluwer Academic Publishers, 1993.

[Niwa & Nitta 94] Niwa Y. and Nitta Y., *Co-occurence vectors from Corpora vs. Distance vectors from dictionaries*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94), pages 304-309. Kyoto, Japan. 1994.

[Normier & Nossim 90] Normier B. and Nossim M., *GENELEX Project: EUREKA for Linguistic Engineering*, in proceedings of the International Worshop on Electronic Dictionaries, pages 63-70, OISO, Kanagawa, Japan. 1990.

[Okumura & Hovy 94] Okumura A. and Hovy E., *Building Japanese-English Dictionary based on Ontology for Machine Translation*, in proceedings of ARPA Workshop on Human Language Technology, pages 236-241, 1994.

[Oostdijk & deHaan 94] Oostdijk N. and deHaan P. Corpus-based research into language. Oostdijk N. and deHaan P. (eds). Rodopi, Amsterdam. 1994.

[Padró 98] Padró L., A Hybrid environment for Syntax-Semantic tagging. Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, 1998.

[Paik et al 95] Paik W., Liddy E., Yu E. and McKenn M., *Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval*, in Boguraev B. and Pustejovsky J. (eds.) Corpus Processing for Lexical Acquisition, The MIT Press, Cambridge, Massachusetts, 1995.

[Pazienza 94] Pazienza M., *Extraction of semantic knowledge from text: a goal or a starting point?*, in proceedings of the 10th annual meeting of SEPLN, Córdoba, Spain. 1994.

[Peh & Ng 97] Peh L. and Ng H., *Domain-Specific Semantic Class Disambiguation Using WordNet*, in proceedings of the Empirical Methods for Natural Language Processing (EMNLP-2). Providence, NY. 1997.

[Pereira & Warren 80] Pereira F. and Warren D., *Definite clause grammars for language analysis - a survey of the formalism and a comparison with augmented transition networks*. Artificial Inteligence 13(3), pages 231-278. 1980.

[Pereira et al. 93] Pereira F., Tishby N. and Lee L., *Distributional Clustering of English Words*, in proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (ACL'93). 1993.

[Pollard & Sag 94] Pollard C. and Sag I., Head-Driven Structure Grammar, University of Chicago Press. 1994.

[Poznanski & Sanfilippo 93] Poznanski V. and Sanfilippo A., *Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora*, in proceedings of the SIGLEX Workshop on Extracting Lexical Knowledge from Text. 1993.

[Pustejovsky 92] Pustejovsky J., *The Adquisition of Lexical Semantic Knowledge from Large Corpora*, in of DARPA WorkShop on Speech and Natural Language, 243-248, New York, February 1992.

[Pustejovsky et al 93] Pustejovsky J., Berger S. and Anick P., *Lexical Semantic Techniques for Corpus Analysis*. Computational Linguistics, 19(2), 1993.

[Rada et al. 89] Rada R., Mili H., Bicknell E. and Blettner M., *Development an Applicationof a Metric on Semantic Nets,* in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30. 1989.

[Ravin 90] Ravin Y., *Disambiguating and interpreting verb definitions*, in proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), pages 260-267. 1990.

[Resnik 93] Resnik P., Selectional and Information: A Class-Based Approach to Lexical Relations, Ph.D. Thesis, University of Pennsylvania, 1993.

[Resnik & Hearst 93] Resnik P., *Hearst M., Structural Ambiguty and Conceptual Relations*, in proceedings of the ACL Workshop on Very Large Corpora. Columbus, OH, 1993.

[Resnik 94] Resnik P., *Using Information Content to Evaluate semantic Similarity in a Taxonomy*, in proceedings of International Joint Conference of Artifical Inteligence (IJCAI'94). 1994.

[Resnik 95] Resnik P., *Disambiguating Noun Groupings with Respect to WordNet Senses,* in proceedings of the Third Workshop on Very Large Corpora, MIT, 1995.

[Resnik 97] Resnik P., *Selectional Preference and Sense Disambiguation*, in proceedings of the SIGLEX WorkShop Tagging Text with Lexical Semantics: Why, What and How?. Washington. 1997.

[Resnik & Yarowsky 97] Resnik P. and Yarowsky D., *A Prespective on Word Sense Disambiguation Methods and their Evaluation*, in proceedings of the SIGLEX WorkShop Tagging Text with Lexical Semantics: Why, What and How?. Washington. 1997.

[Ribas 94] Ribas F., *An Experiment on Learning Appropriate Selectional Restrictions From a Parsed Corpus*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94), pages 769-774. Kyoto, Japan, 1994.

[Ribas 95] Ribas F., On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy, Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, 1995.

[Richardson et al. 94] Richardson R., Smeaton A.F. and Murphy J., *Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words*, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland. 1994.

[Richardson 97] Richardson S., Determining Similarity and Inferring Relations in a Lexical Knowledge Base, Ph.D. Thesis, The City University of New York, New York. 1997.

[Rigau et al. 94] Rigau G., Rodríguez H. and Turmo J., *LDB/LKB integration*, Esprit basic research action 7315, Deliberable 4.1 LLB1, published also as Technical Report LSI-94-32-R, Departament de llenguatges i sistemes informàtics, Barcelona, July, 1994.

[Rigau 94] Rigau G., *An Experiment on Automatic Semantic Tagging of Dictionary Senses,* in International Workshop The Future of the Dictionary, Uriage-les-Bains, Grenoble, France, 1994, published as. Research Report LSI-95-31-R. Computer Science Department. UPC. Barcelona. 1995.

[Rigau et al. 95] Rigau G., Rodríguez H. and Turmo J., *Automatically extracting Translation Links using a wide coverage semantic taxonomy*, in proceedings fifteenth International Conference AI'95 . Language Engineering '95, Montpellier, France. 1995.

[Rigau & Agirre 95] Rigau G. and Agirre E., *Disambiguating bilingual nominal entries against WordNet*, Seventh European Summer School in Logic, Language and Information. ESSLLI'95, Barcelona, Spain, August 1995.

[Rigau et al. 97] Rigau G., Atserias J. and Agirre E. *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation* in proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), pages 48-55. Madrid, Spain, 1997.

[Rigau et al. 98] Rigau G., Rodríguez H. and Agirre E. *Building Accurate Semantic Taxonomies from Monolingual MRDs,* in proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98). Montreal, Canada. 1998.

[Riloff & Shepherd 97] Riloff E. and Shepherd J., *A Corpus-Based Approach for Building Semantic Lexicons*, in proceedings of the Second Conference on Empirical Methods in Natural Language Processing. 1997.

[Riloff & Shoen 95] Riloff E. and Shoen J., *Automatically Acquiring Conceptual Patterns Without an Annotated Corpus*, in proceedings of the 3rd Workshop on Very Large Corpus, 1995.

[Rizk 89] Rizk O., *Sense Disambiguation of Word Translations in Bilingual Dictionaries:Trying to Solve The Mapping Problem Automatically*, RC 14666, IBM T.J. Watson Research Center, Yorktown Heights, NY. 1989.

[Rodríguez et al. in Press] Rodríguez H., Climent S., Vossen P., Bloksma L., Roventini A., Bertagna A., Alonge A., Peters W., *A Top-Down Startegy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*, Computers and the Humanities, in Press.

[Sager 81] Sager N., Natural Language Processing, Addison-Wesley, Reading, Mass. 1981.

[Samiotou 93] Samiotou, A., *Performance of cross-linguistic equivalence relations: A lexicon-based approach.* Msc. Dissertation. UMIST. 1993.

[Sanchez 91] Sanchez A., *Informatización de diccionarios convencionales: un sistema de consulta para el "Diccionario Ideológico de la lengua Española" de J. Casares*, in proceedings of the 7th annual meeting of SEPLN, Valencia, Spain, 1991.

[Sanchez & Nieto 95] Sanchez F. and Nieto A., *Development of a Spanish Version of the Xerox Tagger*, MLAP93/20 EC Crater Project, Working Paper nº6. 1995.

[Sanfilippo 90] Sanfilippo A., *A morphological Analyser for English & Italian*, Computer Laboratory, University of Cambridge. Esprit BRA-3030 Acquilex Working Paper n. 004. March 1990.

[Sanfilippo 94] Sanfilippo A., *The LKB Encoding of Lexical knowledge from Machine-Readable Dictionaries*, in Briscoe E. (ed.), InheritanceDefaults. Cambridge University Press. 1994.

[Schütze 92a] Schütze H., *Word Sense Disambiguation with Sublexical Representations*, in proceedings of the AAAI Spring Symposium on Statistically-Based NLP Techniques. 1992.

[Schütze 92b] Schütze H., *Context Space*, in proceedings of the AAAI Fall Symposium on Statistically-Based NLP Techniques. 1992.

[Schütze 92c] Schütze H., *Dimensions of Meaning*, In proceedings of Supercomputing'92.

[Slator 88] Slator B., *Lexical Semantics and Preference Semantics Analysis*, Memorandum in Computer and Cognitive Science, MCCS-88-143, Computing Research Laboratory, New Mexico State University. Las Cruces. 1988.

[Slator 91] Slator B., *Using Context for Sense Preference*, in Zernik U. (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Smadja 91a] Smadja F., *From N-Grams to Collocations: An evaluation of Xtract*, in proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL'91). Berkeley, California. 1991.

[Smadja 91b] Smadja F., *Macrocoding the Lexicon with Co-ocurrence Knowledge*, in Zernik U. (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Smadja 92] Smadja F., *How to Compile a Bilingual Collocational Lexicon Automatically*, in proceedings of the AAAI Spring Symposium on Statistically-Based NLP Techniques. San Jose, California. 1992.

[Smadja 93] Smadja F., *Retrieving Collocations from Text:Xtract*, Computational Linguistics 16(1):143-177. 1993.

[Soler 93] Soler, C., *Dealing with Spanish-English/English-Spanish mismatches*. Esprit BRA 7315 Acquilex II Working Paper. 1993.

[Souter & Atwell 94] Souter C. and Atwell E., *Using Parsed Corpora: A review of current practice*, in Oostdijk N. and deHaan P. (eds), Corpus-based research into language. Rodopi, Amsterdam, 1994.

[Sussna 93] Sussna M., *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network,* in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia USA, 1993.

[Sutcliffe & Slater 94] Sutcliffe R. and Slater B., *Word Sense Disambiguation of Text by Association Methods: A Comparative Study*, in proceedings of the 10th annual meeting of SEPLN. Córdoba, Spain, 1994.

[Tanaka & Umemura 94] Tanaka K. and Umemura K., *Construction of a Bilingual Dictionary Intermediated by a Third Language*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94). pages 297-303. Kyoto, Japan, 1994.

[Taulé 95] Taule M., *Representación de las entradas verbales en una Base de Conocimiento Léxico: Diátesis y Semántica Léxica,* Ph.D. Thesis, Universitat de Barcelona, Barcelona, 1995.

[Turmo 97] Turmo J., *TURBIO: Sistema de Extracción de Información a partir de Textos Estructurados*, in proceedings of the 13th annual meeting of SEPLN. Madrid, Spain, 1997.

[Uchida 90] Uchida H., *Electronic Dictionary,* in proceedings of the International Workshop on Electronic Dictionaries, OISO, Kanagawa, japan, pp. 23-43. 1990.

[Utsuro et al. 93] Utsuro T., Matsumoto Y. and Nagao M., *Verbal Case Frame Acquisition from Bilingual Corpora*, in proceedings of International Joint Conference of Artifical Inteligence (IJCAI'93). Chambery, France. 1993.

[Utsuro et al. 94] Utsuro T., Ikeda H., Yamane M., Matsumoto Y. and Nagao M., *Bilingual text Matching using Bilingual Dictionary and Statistics*, in proceedings of the 15th International Conference on Computational Linguistics (COLING'94). pages 1076-1082. Kyoto, Japan. 1994.

[Vanderwerde 95] Vanderwerde L., *Ambiguity in the Acquisition of Lexical Information*, in proceedings of the AAAI'95 Spring Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativy, pages 174-179. 1995.

[Verdejo et al. 91] Verdejo M., Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., Martí M. and Taulé M., *SEISD: User manual*, Research Report LSI-91-47. Computer Science Department. UPC. Barcelona. 1991.

[Veronis & Ide 90] Veronis J. and Ide N., *Very Large Neural Networks for Word Sense Disambiguation*, in proceedings of the European Conference on Artificial Intelligence, ECAI'90, Stockholm, August, 1990.

[Veronis & Ide 91] Veronis J. and Ide N., *An Assessment of Semantic Information Automatically extracted from Machine Readable Dictionaries*, in proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL'91). Berlin, Germany. pages 227-232. 1991.

[Voorhees 93] Voorhees E., *Using WordNet to Disambiguate Word Senses for Text Retrieval*, in proceedings of 16th International Conference on Research and Development in Information Retrieval ACM-SIGIR'93. Pittsburgh, PA. 1993.

[Vossen et al. 89] Vossen P., Meijs W. and Broeder M., *Meaning and structure in dictionary definitions*, in Boguraev B. and Briscoe E. eds. Computational Lexicography for NLP, chaper 7. Longman, London. 1989.

[Vossen & Serail 90] Vossen P. and Serail I., *Word-Devil, a Taxonomy-Browser for Lexical Decomposition via the Lexicon*, Esprit BRA-3030 Acquilex Working Paper n. 009. March 1990.

[Vossen 92] Vossen P., *The Automatic Construction of a Knowledge from dictionaries: a combination of techniques*, in Tommola H., Tarantola K., Tolonen T. and Schoop J. (eds.) Proceedings of the 5th Euralex International Congress on Lexcography, pages 311-326. Tampere, Finland, 1992.

[Vossen 94] Vossen P., *Distinguishing Levels in Noun Taxonomies*, Esprit BRA-7315 Acquilex Working Paper n. 094. 1994.

[Vossen 95] Vossen P., Grammatical and Conceptual Individuation in the Lexicon, Ph.D. Thesis, Universiteit vam Amsterdam, Amsterdam. Also publised in IFOTT Studies in Language and Language Usage 15. 1995.

[Vossen in Press] Vossen P., *Introduction to EuroWordNet,* Computers and the Humanities, in Press.

[Walker & Amsler 86] Walker D. and Amsler R., *The Use of Machine-Readable Dictionaries is Sublanguage Analysis*, in R. Grishman and R. Kittredge (eds.), Analysing Language in Restricted domains, Lawrence Erlbaum, Hillsdale, NJ, pages 69-84, 1986.

[Wilks et al. 93] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B., *Providing Machine Tractablle Dictionary Tools,* in Pustejowsky J. ed. Semantics and the Lexicon, Dordrecht, Kluwer Academic Publishers. pages 341-401, 1993.

[Wilks et al. 96] Wilks Y., Slator B. and Guthrie L., Electric Words: Dictionaries, Computers, and Meanings. The MIT Press, Cambridge, MA. 1996.

[Wilks & Stevenson 97] Wilks Y. and Stevenson M., *Sense Tagging: Semantic tagging with a Lexicon*, in proceedings of the SIGLEX WorkShop Tagging Text with Lexical Semantics: Why, What and How?. Washington. 1997.

[Wu & Palmer 94] Wu Z. and Palmer M., *Verb Semantics and Lexical Selection*, in proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94). Las Cruces, New Mexico, 1994.

[Yarowsky 92] Yarowsky D., *Word-Sense Disambiguation Using Statistical Models of Rogets Categories Traiend on Large Corpora*, in proceedings of the 14th International Conference on Computational Linguistics (COLING'92), pages 454-460, Nantes, France, 1992.

[Yarowsky 93] Yarowsky D., *One Sense per Colocation*, in proceedings of ARPA Human Language Technology WorkShop, Princeton, 1993.

[Yarowsky 94] Yarowsky D., *Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoring in Spanish and French*, in proceedings of the 32th Annual Meeting of tha Association for Computational Linguistics, (ACL'95), Las Cruces, New Mexico, 1994.

[Yarowsky 95] Yarowsky D., *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*, in proceedings of the 33th Annual Meeting of tha Association for Computational Linguistics, (ACL'95). 1995.

[Yngve 55] Yngve V., *Syntax and the Problem of Multiple Meaning,* in Machine Translation of Languages. Ed. William Locke and Donald Booth, New York, Wiley, 1955.

[Yokoi 95] Yokoi T., *The Impact of the EDR Electronic Dictionary on Very Large Knowledge Bases,* Towards Very Large Knowledge Bases, Mars N. (Ed.). IOS Press. 1995.

[Zernik 89] Zernik U., *Lexical acquisition: learning from corpora by capitalising on lexical categories*, in proceedings of International Joint Conference of Artifical Inteligence (IJCAI'89). 1989.

[Zernik 91] Zernik U., Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Zernik U. Ed. Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Zipf 45] Zipf G., *The meaning-frequency relationship of words*, The Journal of general Psychology 33, pages 251-256. 1945.

Appendix

Summary of Lexical Resources derived

A set of large complete Spanish MTDs (ready-usable Spanish lexicons) have been derived automatically from the own dictionaries to help some methodological steps (morphological Lexicons, coocurrence vectors, etc.). Most of them have been derived using simple UNIX shell programming utilities (grep, awk, Perl, etc.).

Coocurrence Lexicons

• Word frequency list containing 96,375 different words. First number of occurrences per word in the whole dictionary are shown in next table.

#	Word	#	Word	#	Word
86,667	de	21,066	y	10,982	a
32,568	que	13,788	el	9,886	las
23,634	la	12,555	se	9,599	un
23,330	en	12,376	los	9,101	una
21,320	o	12,045	del	8,845	para

• Bigram frequency list containing 10,291 different word pairs (number of occurrences bigger than 10). First bigrams of the list are shown in next table.

#	Bigram	#	Bigram	#	Bigram
9,483	de la	4,168	Acción de	1,997	Relativo a
8,799	que se	3,916	en la	1,885	en que
5,213	de un	3,903	de las	1,855	a los
5,065	de los	3,827	en el	1,623	Persona que
5,017	de una	3,810	una cosa	1,574	a una
4,686	a la	3,133	Efecto de	1,555	Que tiene

• Trigram frequency lexicon containing 4,021 different word triples (number of occurrences bigger than 10). First trigrams of the lexicon are shown in next table.

#	Trigram	#	Trigram	#	Trigram
1,237	Relativo a la	527	tiene por oficio	392	una persona o
941	en que se	480	que se hace	366	formación de palabras
845	de una cosa	478	que sirve para	360	Persona que tiene
807	con que se	428	la formación de	358	por medio de
578	que tiene por	421	a una persona	346	persona o cosa
550	en forma de	407	en la formación	335	de una persona

• A lexicon of 300,062 weighted coocurrences among 40,193 word form pairs derived from the whole dictionary. Two words are coocurrent if they appear in the same definition (word order in definitions are not taken into account). From left to right, association ratio, pair of words and number of times that they appear together in the whole dictionary.

AR	Coocurrence Words	#	AR	Coocurrence Words	#
24.6658	entra formación	1,041	23.8042	elemento significado	777
24.3899	prefijal significado	652	23.6770	formación prefijal	713
24.1675	entra significado	810	23.6357	palabras significado	821
24.0024	elemento prefijal	718	23.4981	entra palabras	935
23.9983	entra prefijal	714	23.4491	elemento entra	859
23.8511	formación significado	810	23.3636	palabras prefijal	713

• A lexicon of 192,858 weighted coocurrences among 30,765 word form pairs derived from noun definitions. Two words are coocurrent if they appear in the same definition (word order in definitions are not taken into account). From left to right, association ratio, pair of words and number of times that they appear together in the whole dictionary.

AR	Coocurrence Words	#	AR	Coocurrence Words	#
20.4493	pez teleósteo	226	19.3889	marino pez	222
20.3232	marino teleósteo	179	19.2538	dialecto hablado	66
20.2021	atómico químico	107	19.2043	perciforme teleósteo	96
20.1176	atómico símbolo	106	19.1756	atómico elemento	108
19.7993	químico símbolo	119	19.1748	dado golpe	221
19.4051	flores hojas	542	18.8308	hablada lengua	154

Part of Speech Lexicons

• Part of speech lexicon with 103,541 words derived from the monolingual dictionary. First 10 words of the lexicon are shown in next table.

Word	POS	Word	POS
a	prep.	ababa	f.
aaronita	adj.-com.	ababillarse	prnl.
aaronita	adj.-s.	ababo	m.
aarónico	adj.	ababuy	m.
aba	m.	ababábite	m.

• Part of speech lexicon with 18,947 words derived from Spanish/English bilingual dictionary. First 10 words of the lexicon are shown in next table.

Word	POS	Word	POS
abacería	f	abajo	interj
abad	m	abalanzarse	p
abadesa	f	abalorio	m
abadía	f	abandonado	adj
abajo	adv	abandonar	p

Translation Lexicons

• Harmonized bilingual dictionaries. For instance, 28,129 correct connections among 15,848 English nouns and 14, 879 Spanish nouns merging both sides of the bilingual dictionary. First 15 noun pair translations are shown in next table.

English	Spanish	English	Spanish
abacus	ábaco	abbess	abadesa
abandonment	reducción	abbey	abadía
abatement	reducción	abbot	abad
abattoir	matadero	abbreviation	abreviación
abbacy	abadía	abbreviation	abreviatura

• A version of the Spanish WordNet (disambiguated at a synset level) containing 15,535 connections (7,383 polysemous) among 10,786 synsets and 9,986 Spanish nouns with a final accuracy of 86,4%. Next table shows 20 SEA-FOOD Spanish synsets. From left to right, synset English words ans Spanish words.

04995433	bream sea_bream	besugo
04996322	grouper	mero
04996879	carp	carpa
04997880	tuna tuna_fish tunny	atún bonito
04998328	mackerel	caballa
04998878	octopus	pulpo
04998930	escargot snail	caracol
05000192	mussel	mejillón
05000578	eel	anguila
05000943	herring	arenque
05001396	kipper kippered_herring	arenque_ahumado
05003833	crawdad crawfish crayfish ecrevisse	langosta
05003969	cod codfish	bacalao
05004387	haddock	eglefino
05005170	flounder plaice turbot	platija rodaballo solla
05005730	hake	merluza
05007529	pilchard sardine	sardina
05007622	prawn shrimp gamba	langostino
05007938	trout	trucha
05008724	salmon	salmón

Semantic Lexicons

• Synonymy lexicon. 16,333 synonym sets of nouns. For instance, 6 slang ways to say *dinero* (money) in Spanish, *guita, parné, pasta, pela, peseta, tela*.

• Word taxonomies. That is, not disambiguated word tangled hierarchies. 104,900 connections among 59,755 nouns.

Headword	Relation	Genus	Headword	Relation	Genus
ábaco	IS-A	cuadro	ábaco	IS-A	artesa
ábaco	IS-A	tabla	ábaco	IS-A	tablero
ábaco	IS-A	instrumento	ábrego	IS-A	viento
ábaco	IS-A	nomograma	ábrigo	IS-A	ábrego
ábaco	IS-A	superior	ábsida	IS-A	ábside

• Salient word form lists for every WordNet lexicographer (semantic) file. 23,418 connections among 13,347 word forms and 25 coarce grained semantic tags. From left to right, word, lexicographic file (or semantic file, SF) and association ratio.

Word	SF	AR	Word	SF	AR	Word	SF	AR
ábaco	artifact	3.0479	África	animal	0.4964	álcali	artifact	0.4069
ácido	substance	3.7318	águila	animal	2.9449	álcali	substance	2.5761
ácidos	substance	2.1465	águila	possesion	2.4035	álgebra	cognition	1.3678
ácrata	person	1.7196	álamo	plant	1.4148	ámbar	substance	2.4899
África	animal	4.9555	álamos	artifact	0.9810	ámbito	attribute	2.0921

For instance, selecting those salient words of file 12 (feeling) a list of 263 Spanish words (ordered by association ratio) can be obtained. First 15 of the words are shown in next table.

Word	AR	Word	AR	Word	AR
deseo	6.8361	aversión	5.5894	propensión	4.6007
disgusto	6.3325	enojo	5.2438	temor	4.5347
desazón	6.2215	vehemente	4.9974	alegría	4.5347
repugnancia	5.7918	apetito	4.9071	odio	4.4182
asco	5.7522	nostalgia	4.7575	espanto	4.4182

• Lexicons of semantically tagged dictionary senses. For instance, following the Yarowsky approach, 86,759 DGILE noun senses (93% of total nouns in DGILE) semantically tagged with one of 25 noun semantic files of WordNet were generated. First 10 senses of file 13 (FOOD) ordered by sum association ratio of word definitions are shown in next table.

AR	Sense	AR	Sense	AR	Sense
21.6136	galleta_1_2	19.4969	pasta_1_5	18.8116	dulce_1_5
20.8583	chocolate_1_2	19.3087	rollo_1_5	18.6844	buñuelo_1_1
20.5877	jarabe_1_1	19.1253	sangría_1_9	18.6737	limonada_1_1
19.7152	churro_1_2	18.9603	pastel_1_1	18.5768	pasta_1_3
19.5872	bizcocho_1_2	18.9446	galacina_1_1	18.4373	hipocrás_1_1

• Complete Word-sense disambiguated taxonomy. Following a pure descriptive methodology, we derived (see Section 5.2.1) a noun taxonomy that contains 111,624 dictionary senses and has only 832 dictionary senses which are tops of the taxonomy (these top dictionary senses have no hypernyms), and 89,458 leaves (which have no hyponyms). That is, 21,334 definitions are placed between the top nodes and the leaves. The average number of direct hyponyms per node is 5.01. Next table shows the ten noun genus senses wih more descendants in DGILE.

#	Sense	#	Sense
14,042	ejecución_1_1	6,891	calidad_1_1
13,648	entidad_1_1	4,294	animal_1_2
10,500	resultado_1_1	2,366	línea_1_5
10,148	persona_1_1	2,349	preciso_1_1
6,909	efecto_1_2	2,012	efecto_1_1

• Partial word-sense disambigauted taxonomies. Following a mixed methodology, different sizes of taxonomies can be produced depending on the degree of accuracy and filters we apply. For instance, with accuracy near 100% (with filter LABEL2+F2+F3>9) on genus terms selected we produce a noun taxonomy of 35,099 definitions. If we reduce the level of accuracy to 96% (with filter LABEL2+F2+F3>4), we obtain a taxonomy structure of 40,754 senses. For instance, selecting filter LABEL2+F2+F3>9 next table shows the ten noun genus senses wih more descendants in FOOD DGILE classification.

#	Sense	#	Sense
80	zumo_1_1	30	leche_1_3
78	manjar_1_1	26	grano_1_1
76	bebida_1_4	26	carne_1_4
35	plato_1_2	23	comida_1_2
31	pan_1_1	22	pasta_1_2

For instance, the taxonomy for wines in FOOD DGILE classification:

zumo_1_1 vino_1_1 ablución_1_5
zumo_1_1 vino_1_1 aguapié_1_1
zumo_1_1 vino_1_1 ahumado_1_4
zumo_1_1 vino_1_1 albariño_1_1
zumo_1_1 vino_1_1 alicante_1_3
zumo_1_1 vino_1_1 aloque_1_2
zumo_1_1 vino_1_1 alpiste_1_3
zumo_1_1 vino_1_1 amontillado_1_1
zumo_1_1 vino_1_1 amoroso_1_5
zumo_1_1 vino_1_1 dolaje_1_1
zumo_1_1 vino_1_1 falerno_1_1
zumo_1_1 vino_1_1 fino_1_9
zumo_1_1 vino_1_1 fondillón_1_2
zumo_1_1 vino_1_1 garnacha_2_2
zumo_1_1 vino_1_1 jerez_1_1
zumo_1_1 vino_1_1 jerte_1_1
zumo_1_1 vino_1_1 jumilla_1_1
zumo_1_1 vino_1_1 lágrima_1_8
zumo_1_1 vino_1_1 malvasía_1_2
zumo_1_1 vino_1_1 mollate_1_1
zumo_1_1 vino_1_1 montilla_1_1

zumo_1_1	vino_1_1	morapio_1_1
zumo_1_1	vino_1_1	moriles_1_1
zumo_1_1	vino_1_1	mostagán_1_1
zumo_1_1	vino_1_1	mosto_1_2
zumo_1_1	vino_1_1	málaga_1_1
zumo_1_1	vino_1_1	navalcarnero_1_1
zumo_1_1	vino_1_1	navarra_1_1
zumo_1_1	vino_1_1	oloroso_1_2
zumo_1_1	vino_1_1	oporto_1_1
zumo_1_1	vino_1_1	pajarete_1_1
zumo_1_1	vino_1_1	pajarilla_1_3
zumo_1_1	vino_1_1	peleón_1_1
zumo_1_1	vino_1_1	penedés_1_1
zumo_1_1	vino_1_1	perojiménez_1_2
zumo_1_1	vino_1_1	peñafiel_1_1
zumo_1_1	vino_1_1	priorato_2_1
zumo_1_1	vino_1_1	purrela_1_1
zumo_1_1	vino_1_1	quianti_1_1
zumo_1_1	vino_1_1	raya_1_8
zumo_1_1	vino_1_1	requena_1_1
zumo_1_1	vino_1_1	reserva_1_12
zumo_1_1	vino_1_1	ribeiro_1_1
zumo_1_1	vino_1_1	rioja_1_1
zumo_1_1	vino_1_1	roete_1_1
zumo_1_1	vino_1_1	rosado_1_3
zumo_1_1	vino_1_1	rueda_2_1
zumo_1_1	vino_1_1	sherry_1_1
zumo_1_1	vino_1_1	tarragona_1_1
zumo_1_1	vino_1_1	tintilla_1_1
zumo_1_1	vino_1_1	tintorro_1_1
zumo_1_1	vino_1_1	toro_3_1
zumo_1_1	vino_1_1	tostadillo_1_2
zumo_1_1	vino_1_1	transfer_1_1
zumo_1_1	vino_1_1	trinque_1_1
zumo_1_1	vino_1_1	turco_1_5
zumo_1_1	vino_1_1	utiel_1_1
zumo_1_1	vino_1_1	valdepeñas_1_1
zumo_1_1	vino_1_1	verdea_1_1
zumo_1_1	vino_1_1	vinaza_1_1
zumo_1_1	vino_1_1	vinazo_1_1
zumo_1_1	vino_1_1	vinillo_1_1
zumo_1_1	vino_1_1	yecla_1_1
zumo_1_1	vino_1_1	zumaque_1_4
zumo_1_1	vino_1_1	zupia_1_2

• A lexicon containing 29,205 DGILE noun senses (31% of total nouns in DGILE) semantically tagged with one of 25 noun semantic files of WordNet obtained from appliying our method to enrich monolingual dictionary definitions using bilingual dictionaries, a lexical knowledge base (as WordNet) and the notion of Conceptual Distance. First 5 DGILE senses of tag 12 (FEELING) are shown in next table. From left to right, WordNet1.5 synset number, tag file, dgile sense identifier and genus term.

Sense	Genus Term	WN1.5 Synset	Conceptual Distance
malicia_1_1	maldad	04827166	0.0417
ojeriza_1_1	odio	04827166	0.0417
rencor_1_1	resentimiento	04825953	0.0417
angustia_1_2	temor	04812397	0.0500
apego_1_2	cariño	04823348	0.0500