EAGLES LE3-4244
Preliminary Recommendations on Lexical Semantic Encoding
Final Report

The EAGLES Lexicon Interest Group

Antonio Sanfilippo <antonio@lux.dg13.cec.be> [ex SHARP] (*Chair*)
Nicoletta Calzolari <glottolo@ilc.pi.cnr.it> (*Coordinator*)
Sophia Ananiadou <Sophia.Ananiadou@eml.org> (*Subgroup coordinator*)
Rob Gaizauskas <R.Gaizauskas@dcs.shef.ac.uk> (*Subgroup coordinator*)
Patrick Saint-Dizier <stdizier@irit.fr> (*Subgroup coordinator*)
Piek Vossen <piek.vossen@let.uva.nl> (*Subgroup coordinator*)
Antonietta Alonge <aalonge@pg.tecnonet.it>
Nuria Bel <nuria@gilcub.es>
Kalina Bontcheva <kalina@dcs.shef.ac.uk>
Pierrette Bouillon <Pierrette.Bouillon@issco.unige.ch>
Paul Buitelaar <paulb@dfki.de>
Federica Busa <federica@cs.brandeis.edu>
Andrew Harley <aharley@cup.cam.ac.uk>
Mouna Kamel <kamel@irit.fr>
Maite Melero Nogues <mmelero@iec.es>
Simonetta Montemagni <simo@ilc.pi.cnr.it>
Pedro Diez-Orzas <pdiez@uax.es>
Fabio Pianesi <pianesi@itc.it>
Vito Pirrelli <vito@ilc.pi.cnr.it>
Frederique Segond <Frederique.Segond@grenoble.rxrc.xerox.com>,
Christian Sjögreen <sjogreen@svenska.gu.se>
Mark Stevenson <marks@dcs.shef.ac.uk>
Maria Toporowska Gronostaj <svemt@svenska.gu.se>
Marta Villegas Montserrat <tona@gilcub.es>
Antonio Zampolli <glottolo@ilc.pi.cnr.it>

February, 1999

Contents

Foreword

This report provides a final account of work carried out by the EAGLES Lexicon Interest Group on the standarisation of lexical semantic encoding for Human Language Technology applications.

The group operated from May 1997 to December 1998 through a series of working sessions to which nearly 25 experts from European academic and industrial research laboratories provided regular input.

Interim results of this work were validated through discussion panels at the *First International Conference on Language Resources & Evaluation* (Granada, Spain, 28-30 May 1998) and *COLING-ACL '98* (Montreal, Canada, 10-14 August 1998). An interim report was also distributed to interested parties within the community and published on the Web site of the coordinating partner (http://www.ilc.pi.cnr.it/EAGLES/home.html).

As a result of these dissemination activities, the group kept growing in size with an increasingly active participation from the United States. In addition to the regular contributor listed in the title page, the current interest group includes over 30 members from the US. The group is currently one of the major foci of EU-US cooperation activities in the area of Language Technology.

The main objective of the work carried out is to accelerate the provision of lexical semantic standards for Human Language Technology applications. The social and economic topicality of such an activity are demonstrated by the following factors.

Advanced content processing will be a major enabler of Human Language Technologies in the next 3-5 years. During the last decade, HLT applications have acquired commercial viability through the use of robust word-based analysis tools such as stemmers, lemmatisers and statistical part of speech and phrase recognisers. The integration of concept-based processing capabilities is the next logical step to make available more intelligent language-enabled technologies, which are better suited to satisfy users' needs.

Language Technology Applications enhanced with semantic processing capabilities will have a major commercial impact. The globalisation of economies and societies in combination with a bolstering Internet penetration provide a fertile terrain for the advancement of the digital economy. In this context, intelligent online services and information management tools provide a strong competitive advantage to sellers and buyers alike across the business fabric, by facilitating the outsourcing of transaction services to the client and bridging linguistic, social and geographical barriers.

The evaluation of lexical resources will be a crucial step in the development of Human Language Technologies for the new millennium High quality lexical resources will be increasingly important for the development of language-enabled applications able to support the digital economy and facilitate information access and dissemination in the Information Society. Lexical semantic standards provide an essential tool in the assessment of the lexical resources needed for such applications.

MT and IS applications provide an ideal focus in standardising lexical semantic encoding. Through the increasing ubiquity of the World Wide Web in our everyday life, search engines and online translation services are assuming a primary role in securing information access to citizens. At the same time, the enabling technologies which underly these applications are mature enough to employ various degrees of semantic processing successfully.

Early lexical semantic standards pave the way towards common protocols for word sense identification. Full lexical semantic standardisation requires a complete suite of consensual criteria for word sense identification leading to comparable word meaning characterisations. Early lexical semantic standards help identifying such criteria.

The report provides a survey in the area of focus and a set of guidelines. The survey describes:

- linguistic aspects of lexical semantics (§2),
- lexical semantic resources (§3)
- semantic requirements for Machine Translation, Information Systems (§4) and their enabling technologies (§5).

The guidelines (§6) give an early specification of lexical semantic standards with specific reference to

- thesaural relations (e.g. synonymy, hyponymy, meronymy, etc.);
- ontological classification and subject domains, and
- linguistic characterisation with reference to aspect, quantification and predicate frames.

Readers who wish to view the guidelines now can proceed directly to §6.9.

Chapter 1

Introduction

This report provides a final account of work carried out by the EAGLES Lexicon Interest Group on the standarisation of lexical semantic encoding for Human Language Technology applications.

EAGLES is a project sponsored by DGXIII of the European Commission to promote the creation of de facto standards in the area of Human Language Technologies (HLT). EAGLES work has been carried out on selected priority areas by several interest groups bringing together representatives from industry and academia. Interest areas tackled so far include the following.[1]

Text corpora: morphological and syntactic annotation, and corpus mark-up.

Computational lexicons: morphosyntactic specifications (language independent and dependent); specifications for syntactic and semantic subcategorisation, and common specifications for lexical data in corpora and lexicons.

Spoken language resources: spoken language system and corpus design; spoken language characterisation; spoken language system assessment; spoken language reference materials; terminology; off-the-shelf product evaluation; audiovisual and multimodal systems, and speech databases.

Evaluation of NLP systems: development of a general methodological framework for user oriented adequacy evaluation; user profiling, and cooperation with ISO.

Integrated resources: dialogue representation and annotation for text and spoken language corpora.

Computational linguistic formalisms: Major trends and commonalities in popular formalisms.

The overall goal of the Computational Lexicons Interest Group is to provide guidelines for the standardization of lexical encoding. The work reported here is intended to

[1] Further information on earlier and later phases of the EAGLES project can be found at the *HLTCentral* web site ¡http://www.linglink.lu/le/projects/eagles/index.html¿ and the EAGLES home page ¡http://www.ilc.pi.cnr.it/EAGLES/home.html¿.

provide preliminary recommendations on lexical semantic encoding. These recommandations are meant to extend the results of standardization activities in the area of syntactic subcategorization previously carried out by the EAGLES Lexicon Interest Group, see http://www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html.

The proposed extension addresses the annotation of

- semantic class of subcategorizing expressions (e.g. verbs),
- semantic class of subcategorized terms (e.g. nouns), and
- integration of semantic and syntactic information in subcategorization structures.

The work has a special focus on requirements for Machine Translation and Information Systems. These applications have had a major impact on the electronics and computing industries and are therefore excellent candidates in providing a focus for standardization in the area of language engineering.

The workplan includes the survey of current practices in the encoding of semantic information and the provision of guidelines for standards in the area of focus with reference to:

- approaches to Lexical Semantics;
- lexicographic resources and Natural Language Processing lexicons, and
- semantic requirements for Machine Translation and Information Systems (e.g. Information Extraction/Retrieval and Summarization).

The survey provides a description of lexical semantic notions as

- developed in the linguistic literature (Part I, 2);
- used in the compilation of lexical resources (Part II, 3), and
- applied in Machine Translation, Information Systems, and related enabling technologies (Part III, 4 and 5).

The goal of the survey is to provide the basic environment in which to table proposals for standardization in lexical semantic encoding. More specifically, the survey is intended to be indicative of the sort of issues at stake in standardization, e.g. what are the basic notions and how they are utilized. Such an indication is achieved through a comparative assessment of the status of lexical semantics in theoretical linguistics, lexicography and language engineering.

Although we have tried to cover all major topics and applications concerned with equal attention, we recognize that the survey may fail to address several relevant areas and possibly provide too specific a treatment of others. Meaning is such a pervasive aspect of linguistic knowledge that a totally unbiased and perfectly balanced survey of theoretical and applied work involving lexical semantic notions is a very difficult, perhaps impossible, project to carry out. Fortunately, such a limitation is not likely to affect our purposes. Indeed, our choice of lexical semantics issues and language resources to be surveyed was driven by the decision to concentrate on the requirements of applications such as Machine Translation and Information Systems. By narrowing the focus of application, we hope to have avoided the danger of unmotivated biases.

The chapter on guidelines (6) describes a preliminary proposal for standards in the area of lexical semantic encoding with specific reference to the areas, tools and applications discussed in the survey parts.

Part I

Linguistic aspects of lexical semantics

Chapter 2

Linguistic aspects of lexical semantics

2.1 Introduction

In this chapter, we survey a number of areas of lexical semantics which are often referred to and used with some adaptations in NLP applications. One of the aims of this section is to provide the reader with a background of linguistic issues in lexical semantics so that he/she can have a better view of what the main definitions, the hypotheses, the foundations and the limitations of each area are. These linguistic issues are often not used directly in NLP systems and in the description of lexical resources, they are usually adapted at various degrees to make the descriptions e.g. more comprehensive, and better suited for the problem or the domain been addressed.

The elements presented here are basic issues in lexical semantics and have in most cases an inter-theoretical perspective. They have been judged to be sufficiently stable to be presented here. They can therefore be used in a number of frameworks; they can also be integrated or combined with a large number of syntactic systems. A choice criterion has been that lexical semantic descriptions should be as easy to use as possible and non-ambiguous. They should also be as language independent as possible and easy to encode in a lexicon. It should be noted that we do not claim to be comprehensive, and the absence of a framework is certainly not an *a priori* judgement of its value. Lexical semantics is indeed a vast area with many ramifications, impossible to cover in such a document.

The areas surveyed here are the following:

- **Aspect in the lexicon**, the aim is to define a common notation for actionality/Akstionart and Aspect. The concrete use in lexicons of aspect is to mark verbs, in syntactic contexts, in order to identify the nature of the action: e.g. perfective/imperfective,

- **Lexicalization**, the aim is to identify and to organize lexicalization patterns across languages. Uses are numerous, e.g.: the study of the cooperation syntax-semantics, and its use in generation of NL, the investigation of mismatches between 2 languages, and the taking into account of lexicalization preferences. The questions addressed are:
  - What is a lexicalization pattern?
  - What are the meaning components at stake and their combinatorial properties?

  – What notions are used in applied work and how to normalize them?

- **Semantic roles** (or thematic roles), the aim is to classify the arguments of predicates into a small and closed set of types of 'participants' modelling the role played by the argument w.r.t. the predicate. This approach provides a level of semantic representation. Semantic roles can also be viewed as a bridge between syntax and semantics. They have been used in a variety of applications ranging from Machine Translation (see §4.1) to Information Retrieval (see §4.2.2).

- **Verb semantic classes** and the semantics of verbs, the aims are to identify and to compare different classifications methods for verbs, to outline their properties, advantages and limitations, and to isolate semantic universals. This study contributes to the identification of senses and sense extensions. Uses are quite numerous, e.g.: a better organization of semantic lexicons where verb classes share a number of semantic properties, and a contribution to lexical acquisition. The contents are: the verb syntactic alternations systems, classifications based on semantic roles and Lexical Conceptual Structure-based classifications.

- **The semantics of nouns and compound nouns**, the aims are the same as for verbs; in addition, the following points are studied: nouns and nominalizations, semantic properties: WordNet classes, quantification, semantic alternations. Different forms of representations (features, networks) are presented. An analysis and a specification of noun compounds, relational nouns and deverbal nouns is also introduced

- **Adjectives**, the aim is to classify different types of adjectives, and to normalize syntactic alternations involving adjectives. A second aim is to identify generic properties modified by adjectives and to normalize them.

- **Prepositions**, the aims are to classify prepositions w.r.t. uses, to identify general semantic classes, prepositions being very polysemic. Another goal is to relate prepositions to semantic roles, and to selectional restrictions. The main use considered here is the introduction of generic classifications in verb subcategorization frames.

For each area, a description of the main approaches is first given, followed by comparisons aimed at identifying basic notions and components. Then relations to other areas of lexical semantics are emphasized. Finally, as a form of pointer towards the two next chapters, indications are given on how lexical semantics data is encoded in lexical knowledge bases and used in large-size NL applications (see §4).

Lexical semantics being a very vast domain, it has not been possible to include every approach. The absence of a framework should not be considered as an a priori judgement of its value. Among the missing frameworks that we foresee to add in the future are the Mel'cuk functional system (somewhat evoked in some paragraphs) and the Script/Frame system.

2.2 Lexical aspect

2.2.1 Introduction

The notion of aspectuality has traditionally come to cover a number of distinct but closely related phenomenon and approaches strictly intertwined with the notion of event and event structure. Consider the following examples:

(1) a. John loved Mary for/in three years

b. John ran for/*in three hours

c. John ate an apple *for/in three minutes

d. John reached the top of the mountain *for/in three days

As can be seen, the reported verbs differ as to their capability to combine with *for-* and *in-* adverbials. This and many other examples leadscholars to revive and renew a classification of verbal predicates dating as far back as to Aristotle. For instance, it is now common to talk about stative predicates (*to love*, *to know*, ...), activities (*to run*, *to push*, ...), accomplishments (*to eat*, *to drink*, *to write*, ...) and achievements (*to find*, *to note*, ...), this way following suggestions elaborated by [Ken63], [Ven67], [Dow79] and many others.

It has been noted that these predicates differ as to their internal constitution and their temporal properties. Consider, for instance, (1a). There is a clear indication of a three-years period in which the predicate (*to love*) was true. However, it can be noted that much the same obtains for any subperiod of it. That is, for any subperiod t of the three-years period, the predicate *to love* can still truthfully apply. Similar considerations hold, at least to a certain extend, of activities. Surely, there are subperiods of a runnings that still qualify as runnings; that is, they show the **subinterval property**. On the other hand, predicates such as *eating an apple* do not display the same behaviour. If it took three minutes to eat an apple, as in (1c), then there is no subpart of this period to which the predicate *eat an apple* can be applied. On this respect achievements pattern with accomplishments. In the literature, these facts have been described by resorting to the notion of **cumulativity** and/or **divisibility**. Statives and activities are divisive whereas accomplishments and achivements are not. Other properties have been discovered, which are similar in nature to the one just discussed, e.g. **summativity** and so on. [1]

The explanation of these facts crucially depends on a previous understanding of the ontological and semantic features of the involved entities, that is **events**. In this respect, two main approaches deserve consideration, the Montagovian one and the Davidsonian one.

According to the Montagovian tradition, events are properties of time, whereas Davidsonians regard them as primitive individuals in the ontology of natural languages. For a Montagovian, actionality properties are crucially explained by resorting to time entities. Thus, verbal predicates are true at time points or intervals, and a stative verb corresponds to a predicate which, whenever true at a time (interval) t, is also true at every subinterval included in t. All the other properties can be similarly reduced to properties of properties of times. For our purposes, it is important to notice that this way of explaining events require the full blown apparatus of higher order logic typical of the Montagovian tradition. This means that the approach might be of little value (at least in the generality of cases) for NLP, because of the well known problems of implementing such kind of logics.

In the Davidsonian tradition, as elaborated by, i.e. [Hig85], [Hig97], [Kri89], [Kam93], not only events are primitive individuals, but event variables are introduced in the logical forms of sentences by verbs, nouns (as, e.g. in eventive nominalitation) and so on. This paves the way for a rich, and relatively simple (basically first order) semantics of events, in which eventive variables can be quantified over by such devices as adverb of quantification (*often*, *never*, *always*, etc., see [Lew75], [Swa91]), modifiers directly attach and events (e.g.,

[1] It will not be possible, here, to go into any depth on these questions, so we refer the reader to the available literature, among which: [Kri89], [Dow79], [Hig97], etc.

by conjunctive modification, as in [Dav67], [Hig85], [Hig97], and so on. In this respect, and partially independently from theoretical reasons, the Davidsonian approach recommends itself for NLP purposes, exactly for the reasons that made Montagovian approaches indigest to NLP.

Turning to actionality-like properties in a Davidsonian framework, they can be explained by directly resorting to properties of events, leaving time on the background. One way of pursuing this approach consists in singling out some structural primitive, e.g. **part-of**, which gives rise to appropriate algebraic structures (**lattices** and/or **semi-lattices**). The properties discussed above can then be seen as properties of predicates defined on these algebraic structures. This approach has been explored and made popular by [Kri89], who follows similar suggestions for the objectual domain advanced by [Lin83] and by previous work by [Bac86].[2] One main point of Krifka proposal concerns the existence of formal relations between objectual and eventive domains (both conceived as algebraic structures) which take the form of homomorphism. These relationships are taken to be responsible for such contrast as that in (2):

(2) a. John ate an apple in/*for three minutes

b. John ate apples *in/for three minutes

It has been noticed that the nature and properties of the direct object of such verbs as *to eat* yield different acceptability results when confronted with *in-* and it for-adverbials. Verkuyl [Ver93] observes that the existence of a specific quantity of matter as a referent for the direct object can affect the telic nature of a predicate. According to Krifka, this is correct and amounts to a passage of formal properties, via homomorphism, from the objectual domain to the eventive one. The nominal *an apple* in (2) identifies a specific quantity of a certain matter (**+SQA** for Verkuyl, a quantised predicate for Krifka). Via the homomorphism, this induces a similar specification on the eventive predicate, which becomes quantised. On the other hand, given that the bare plural *apples* is not quantised, by the same token the corresponding complex predicate *eat apples* is not quantised as well. Thus, the contrast is explained by stipulating: that *in-/for-*adverbials select for quantised/non-quantised eventive predicate; that there can be a passage of formal properties from arguments to predicates, via homomorphism; and that such a property is encoded by thematic-relations (see §2.4.5).

Besides the merit of this, or other, particular theory, the relevant point is that actionality emerges not as a property of isolated verbs (or verbal stems) but, rather, as a compositional facet of complex structures. For our purposes, the morale is that verbal stems need to bear appropriate specification that can be used by a parser/generator and/or by a semantic processor to compute the correct actional value of a complex phrase, taking into account the properties of designated arguments.

The latter point needs further clarification. In (2), the argument affecting the actional properties of the (complex) predicate was the direct object. In other cases, however, other complements/arguments can play a similar role. This is the case of directional PPs with verbs of movement:

(3) a. John ran for/*in two hours

b. John ran home *for/in two hours

[2] More information on the importance of the part-of relation can be found in [Sim87]. For events and parthood, see also [Pia96a], [Pia96b].

These cases are basically similar to those discussed above, with the difference that the affecting factor is not a specified quantity of matter, as with many accomplishments, but a **spatial path** (in the sense of Jackendoff) provided with a specific end point. Notice the following example:

(4) John walked around the fountain for/in ten minutes.

Both the *in*- and the *for*- adverbials are acceptable. However, the sentence changes its meaning accordingly. If the *for*-adverbial is chosen, then no completion of the around-the-fountain tour is entailed. That is, John might have been walking around the fountain for a certain period of time, without completing its tour. On the other hand, the *in*-adverbial forces a **completive (telic)** reading. Although the explanation of this contrast calls for such notions as **telicity** and **perfectivity** which will be discussed in the next section, it is important to notice that the contrast is consistent with the discussion above. The PP *around the fountain* can specify the end point of a path and it does so only with the *in*-adverbial and not with the *for*-adverbial.

2.2.2 Perfectivity and telicity

As can be seen from the discussion above, actionality is the result of the interaction between lexical properties (basically, aktionsarten of the verb) and structural ones (the characteristics of the designated argument/complement). Aspectuality, in the way this term is mostly used in the literature, is a morphosyntactic property, deriving from specifications bore by certain verbal forms, such as the Italian simple past, or from free/bound morphems that can (more or less) freely combine with the verb, as in Slavonic and many other languages.[3] The basic distinction is that between **perfectivity** and **imperfectivity**. Interpretively, perfective verbal forms refer to complete/finished events, as in:

(5) John ate an apple.

The eating, as reported in (6), is finished, completed. On the other hand, imperfective forms lack this meaning and depict ongoing, unfinished processes:

(6) Mario mangia/mangiava una mela.
Mario eats/ate(IMPF) an apple.

The present and the imperfect tense in Italian are both imperfective. As a consequence, a sentence such as (7) does not say anything about the completion/termination of the event described. Even if the imperfect, is a past tense, the event need not be conceptualised as finished:

(7) Alle tre Mario scriveva la lettera e la sta ancora scrivendo.
At three, M. wrote(IMPF) the letter and he is still writing it.

Aspectuality heavely interacts with actionality. Thus, the aforementioned tests with *in*- and *for*-adverbials only make sense with perfective forms:

(8) Mario scriveva una lettera *per/*in tre ore.
M. wrote(IMPF) a letter for/in three hours.

Furthermore, the perfective/imperfective distinction also plays a crucial role for telic readings. Teloses (also called **culminations**, **natural end points**, etc..) only arise with perfective

[3]For a discussion of aspectuality, see [Gio97].

aspect and can be revealed by the *in- for*-adverbials test.

Also, teloses depend on the actional value of the complex predicate. Thus, such predicates as *eat an apple* yield teloses, if perfective. On the other hand, predicates such as *eat apples*, *run*, etc.. remain non-telic even when perfective. In a sense, and for many purposes, the telic-atelic distinction overrides and encompasses many traditional aktionsarten ones, along the following schema:

(9) a. Process & +SQA = telic

b. Process & -SQA = atelic

Here process is to be understood as a term of art encompassing vendelrian activities and accomplishments. Notice, that, according to what said above, the fact that a predicate is telic, in the sense just defined, does not entail that any *actual* telos is given, because aspectuality must be considered:

(10) a. telicity & perfectivity $\rightarrow$ +telos

b. telicity & imperfectivity $\rightarrow$ -telos

Atelic predicates, on the other hand, never give raise to teloses:

(11) atelic $\rightarrow$ -telos

Finally, there are predicate which always give produce teloses, that is achievements. In this respect, they can be seen as being both lexically telic and perfective.

To sum up. Actionality results from the combination of the lexical properties of the verb together with those of a designated argument. Aspectuality, on the other hand, is a morphosyntactic properties, depending on the presence of appropriate specifications bore by particular verbal forms or by bound and/or free morphemes attaching to the verb. Finally, telicity lies somehow in between actionality and perfectivity, in many respect depending on both. However, we saw that there is a class of verbs, the so called achievements, which are always telic and always actualise teloses.

Before concluding this sketchy introduction to actionality and aspectuality, one point must be stressed. First of all, actionality seems to be a cross-linguistically rather stable property. Thus corresponding verbs in different languages tend to have the same actional properties and much the same can be said about complex actionality. On the other hand, the perfective/imperfective distinction is highly language-dependent. This holds both with respect to the means available for expressing the distinction itself (complex verbal forms, as in Italian, French, etc., vs. free or bound morphemes attaching to the verb, as in Slavonic, etc.) and with respect to the range of possibilities available. Thus, [Gio97] argue that English does not have the perfective/imperfective distinction. In this language all eventive verbal forms are perfective. Similar conclusions can be drawn for many non-Indoeuropean languages (Creole languages, African languages, Chinese, etc.)

2.2.3 Overlapping area and phenomena

In this section we will be very briefly consider area and phenomena which overlap those discussed above. The first point to be addressed is the progressive form (found in English, Italian, and, to a certain extent, in other Romance and Germanic languages). Two different perspectives can be taken on this respect. According to the first (see [Gio97]) the progressive should be distinguished from true imperfective forms. The former, in fact, has an intensional meaning that ordinary imperfective forms lack.[4] The progressive should be regarded as a way to imperfectivise an otherwise perfective form by intensionalising the completeness/finiteness of the event. The distinction is subtle, but very clear empirically:

(12) John is reaching the top of the mountain.
Gianni sta raggiungendo la cima della montagna.
Gianni raggiunge la cima della montagna.

Both in Italian and in English, an achievement predicate such as *reach the top/raggiungere la cima* can be progressivised in the present tense. However, in Italian, the imperfective present tense form is not acceptable.

On the other hand, other scholars more or less tacitly assume that there is no real distinction between progressive and imperfective verbal forms, collapsing them under the same heading of imperfectivity.

As said above, the distinction is subtle and can probably be neglected for many purposes. It cannot be ignored with such tasks as text or dialogue generation and machine translation, though, at pain of producing a wrong or ungrammatical form. So care must be taken in the design of the basic aspectual classes for an NLP system.

Another important overlapping area is constituted by all such aspectual forms and values as **inchoatives**, **resultatives** and so on. It should be noticed that many of the deviant examples discussed above can have one of this readings. Thus (3a) becomes acceptable in following context:

(13) After the car accident, John ran in three days.

(13) has the inchoative meaning that John was able to run again only three days after the accident he suffered. We suggest that these cases be left unaddressed within this document. To be sure, the availability of such readings depends on a number of structural (syntactic) facts that are beyond the scope of our contribution.[5] As such they should be treated by the syntactic and/or by the semantic analyser. In the case such aspectual values are borne by lexical items (**aspectual verbs**) or by free/bound morpheme, specifications similar to those we are going to propose for the basic aspectual opposition can be easily devised.

2.2.4 Relation to other Areas of Lexical Semantics.

Most obviously, lexical aspect and actionality is connected with Semantic (or Thematic) Roles (see §2.4). For instance, within the approach advocated by [Kri89], the transfer of formal properties from the objectual domain to the eventive one depends on the nature of thematic relation involved, as analysed according to the suggestion in [Dow89] and [Dow91]. See §2.4 for a survey of the relevant problems.

2.2.5 Encoding in Lexical Databases.

Attempts have been made at capturing lexical aspect and aktionsarten within LDBs. An example of a *static* definition can be found in DIONYSUS [Nir93] where aspectual specifications are provided which distinguish between phasal properties (e.g., **begin**), durational properties ($\pm$ **prolonged**), and so on.

[4]See [Dow79] and [Lan92] for a discussion of the intensional properties of the progressive.
[5]For more information, see [Hig97].

In ACQUILEX an attempt is made at a more dynamic classification of aktionsarten which also tries to address the interactions with thematic roles, [San91]. All this is done within a TFS formalism. Thus, the traditional vendlerian distinction of verbal aktionsarten is reconstructed in the type hierarchy by means of the two types **stative** and **dynamic**. These are then combined with the two krifkean types **cumulative** and **quantized** to obtain a quadripartite classification of verbal predicates. Such a classification is then connected with the properties of thematic (proto-)roles to allow for the transfer of properties from the objectual to the eventive domain.

2.2.6 Relevance to LE Applications

Concerning NL analysis in general, lexical aspect and actionality play a role in determining the grammatical and semantical interactions between predicates and other parts of the sentence, both arguments and modifiers. With respect to the former, besides the facts discussed in the previous sections, we want to mention genericity, habituality, event quantification, the mass/count distinction and so on, as cases in which the aspectual and actional properties of the predicate play a relevant role in determining syntactic and semantic facts. Concerning adjuncts, the actional properties of a (simple or complex) predicate determines the kind of temporal modifiers it can combine with (*for-* vs. *in-* adverbials), the kind of adjunct clauses (purpose and rationale clauses), etc... For all these reasons, lexical aspect and actionality play an important role in parsing (e.g., to restrict and control attachment ambiguities) and in semantic computations.

For the purposes of NL generation, and independenlty of the framework chosen to express the meaning of the text to be generated, the aspectual and actional properties of lexical items play a crucial role in lexical choice and in microplanning (see §4.4, 4.5).[6] These considerations are even the more important when the system deals with multilingual generation. In such cases, in fact, stipulation of direct correspondencies between actional and aspectual configurations and predicates on the basis of a single language inevitably fails to generalise to other languages, so that more principled approaches are necessary.

Concerning classes of applications, it is worth mentioning that because of the kind of (semantic) information provided, lexical aspect is important for all the system which must be able to detect and reason on events. This is true of information and knowledge extraction systems (§4.2), dialogue management systems, etc... and in general for all the cases in which it is important to tell whether an action/event, as presented in the text or during the dialogue, is terminated/finished or whether it can continue at the speech time or at any other focalisation time. Furthermore, actionality is crucial to establish the appropriate relationships between the various events presented in the text/dialogue: simultaneity, overlapping, precedence, etc. Within machine translation (§4.1), given the various means available to different languages to express actionality and aspect, a proper mastering of these notions is crucial to improve the quality of the translation, [Dor93], by selecting/producing the appropriate constructions in the target language, etc.., irrespectively of the chosen approach (interlingua, transfer, mixed techniques based). Finally, we want to emphasise that the strict relationship between the eventual domain and the temporal one makes the aspectual information indispensable for systems performing any form of temporal reasoning: e.g., task scheduling, dialogue management, population and maintainance of temporal databases, etc..

[6] Microplanning is the process by which more fine grained syntactic structures for the sentence are produced from the results of discourse and text planning activities.

2.3 Lexical Semantics Relations

In this section, we consider a more global level of lexical organization. Lexical semantics relations play an essential role in lexical semantics and intervene at many levels in natural language comprehension and production. They are also a central element in the organization of lexical semantics knowledge bases. Most of the material presented here is borrowed from [Cru86].

2.3.1 Introduction

Congruence Relations

Two words W1 and W2 denoting respectively sets of entities E1 and E2, are in one of the following four relations:

- identity: E1 = E2,
- inclusion: E2 is included into E1,
- overlapp: E1 and E2 have a non-empty intersection, but one is not included into the other,
- disjunction: E1 and E2 have no element in common.

These relations supports various types of lexical configurations such as the type/subtype relation.

Hierarchical Relations

There are basically three major types of hierarchical relations: taxonomies, meronomies and proportional series.

Taxonomies The taxonomy relation is the well-known isa relation which associates an entity of a certain type to another entity (called the hyponym) of a more general type. Taxonomy introduces a type/subtype relation which can be characterized by one of the following linguistic tests:

- X is a subtype of Y if the following expressions are correct:

  X is a kind of Y or X is a type of Y for nouns,

  X-ing is a way of Y-ing for verbs.

Taxonomies usually have up to 7 levels that correspond to different levels of genericity (as in natural taxonomies). However, taxonomies of technical terms may be much deeper. It is also important to note that in some cases, certain nodes do not have any corresponding word in a given language; whereas they have one in another language. A taxonomy may thus have holes. Methods to define taxonomies are given in §3.6.3. The main property of a taxonomy is transitivity of properties from the type to the subtype. This property can also be viewed as a well-formedness criterion for taxonomies.

Most levels of a certain degree of genericity have a large number of subtypes, each of them having different possible realizations as words. The notion of subtype is however difficult to qualify in an homogeneous way. There is indeed a problem of prototypicality which is raised: some subtypes are more prototypical than others of their hyponym (the type above them). Let us recall the famous example of the blackbird which is more prototypical of a bird than a hen which is itself more prototypical of that same class than a penguin.

Meronomies Meronomies describe the part-whole relation. It is a fairly complex relation which attempts to take into account the degree of differenciation of the parts with respect to the whole and also the role that these parts play with respect to their whole. For example, elements such as spatial cohesion and spatial differenciation, functional differenciation and nature of the links between the parts are crucial elements for determining meronomies. In fact, depending on the quality of these elements, we may have different kinds of meronomies, with different types of properties.

Meronomies can be characterized may be in a slightly too resrictive way, by the following linguistic tests. A is a part of B if one of these sentences is correct:

- *A has B (or A has a B),*
- *B is part of A.*

The meronomy relation has itself some properties (or attributes) which must be taken into account in any realistic model:

- optionality of a part,
- cardinality of a part with respect to the whole, e.g. a human has 2 legs, a car has 4 wheels,
- [Win87] distinguish 6 kinds of meronomies which differ according the functionalities, the spatial cohesion and the degree of dissimilarity between the parts and their whole. We have the following classes:
  - component / integral object: there is a clear structural and functional relation between the whole and its parts, e.g. handle/cup, phonology/linguistics.
  - member / set or group: parts do not necessarily have a structural or functional relation with respect to the whole, parts are distinct from each other. In this class fall for example tree/forest, student/class.
  - portion / mass: There is a complete similarity between the parts and between parts and the whole. Limits between parts are arbitrary and parts do not have any specific function a priori with respect to the whole. We have in this class for example: slice/bread, centimeter/meter. This subrelations is often called a mereology.
  - object / material: This type of relation describes the materials from which an object is constructed or created, or the constitutive elements of an object, e.g. alcohol/wine, steel/car.
  - subactivity / activity or process: describes the different subactivities that form an activity in a structured way, for example in a temporally organized way. Fall in this class examples such as: pay/buy, give exams/teach.

- precise place / area: parts do not really contribute to the whole in a functional way. This subrelation expresses spatiality, as in: oasis/desert, Alps/Europe.

Similarly to taxonomies, the meronomy relation cannot really be concieved between two elements, but should be concieved with respect to the set of all the parts forming the whole. This also permits to introduce a kind of point of view in a meronomic description. Meronomies do not, in general, allow transitivity at logical and linguistic levels. However, some authors tend to allow transitivity at linguistic level between elements which are linked by the same subtype of meronomic relation described above.

Non-Branching Hierarchies Non-branching hierarchies allow for the ordering of elements that correspond to different levels of organization or of dimensionality. The structure does not correspond to a type/subtype organization, but could have in somes cases some similarity with a meronomic relation. Non-branching hierarchies are often related to a spatial, a temporal or an abstract notion of dimensionality.

We can distinguish three kinds of non-branching hierarchies:

- a continuous hierarchy where limits between elements are somewhat fuzzy, as in: frozen - cold - mild - hot; small - average - large, and in most topological relations,
- a non-continuous and non-gradable hierarchy, in general not based on any measurable property such as institutional hierarchies and technical hierarchies: sentence - proposition - phrase - word - morpheme.
- a non-continuous and gradable hierarchy, organized according to a given dimension, such as units of measure.

In some cases, non-branching hierarchies may reflect a more linguistic than common-world knowledge.

Non-Hierarchical relations

Among non-hierarchical relations we mainly distinguish synonymies and the different forms of opposition. These relations, as we shall see it, are either binary or ternary. The ternary character reflects the context-dependence of some of these relations.

Synonyms Two words are synonyms if they have a significant similar semantic content. Synonyms have a significant semantic overlapp, but the degree of synonymy is not necessarily related to that overlapp. There are very few absolute synonyms, if any, in a language, but words may be synonyms in given contexts. We then view the synonymy relation as a ternary relation: W1 and W2 are synonyms in the context C. Synonyms often do not depend on the degree of precision of the semantic descriptions, but their degree of synonymy may however change at different levels of granularity.

Antonyms and Opposites Antonyms and opposites cover a very large variety of phenomena, more or less clearly defined. A basic definition could be that W1 and W2 are antonyms or opposites if they have most semantic characteristics in common but if they also differ in a significant way on at least one essential semantic dimension. Similarly to synonyms, antonyms and opposites are highly contextual and thus introduce a kind of ternary relation. There also various degrees of opposition, some pairs of word-senses are more prototypically opposites than others. Antonyms refer to gradable properties and opposites to non-gradable ones.

For example, with respect to the context 'to start' to keep on and to stop are opposites. Similarly, good and bad are generally admitted as antonyms, and are more prototypical than the opposition between father and mother.

Antonyms do not necessarily partition the conceptual space into two mutually exclusive compartments which cover the whole conceptual domain. Some overlap or space in between is possible, as in good and bad, since it is indeed possible to say that something is neither good nor bad, or, possibly, to say that something is both good and bad. A special class of antonyms are complementaries which divide the whole conceptual space into two non-overlapping compartments. In [Cru86] several classes of complementaries are defined, such as the class of interactives, which represent a relation of the type stimulus-response, as in: *grant - refuse with respect to the context of request.*

Another interesting class among opposites are directional opposites. They represent either basic, topological, or conceptual (metaphorical) directional oppositions. In this class, which is conceptually relatively simple, fall examples such as: start-finish, top-bottom, descend-ascend.

The role of opposites in a lexical semantics knowledge base is somewhat difficult to define. Similarly to synonyms, opposites and antonyms may certainly play the role of integrity constraints. Their use in natural language generation, for example to avoid the use of too many negations, is somewhat hard to make explicit, because of numerous pragmatic factors that may intervene, such as the polarity of an element in a pair of opposites or antonyms. We can say, for example *how expensive is this book?*, but probably not *how cheap is this book?*. Finally, the linguistic tests or the analysis methods for defining exactly if two elements are opposites or antonyms and to what degree remain to be defined precisely.

2.3.2 Lexical semantics relations in lexical knowledge bases

Lexical semantics relations are of much use to structure lexical data, in particular hierarchically. They have been extensively used and evaluated in WordNet (§3.6.2) and EuroWordNet (§3.6.3). They have also been used more experimentally in projects such as Acquilex (§3.12.3) and Delis (§3.12.5).

2.3.3 Lexical semantics relations in applications

Lexical semantics relations are used in the linguistic design of lexical databases such as wordnets and thesauri (see §3, especially §3.6). Lexical resources enciding information about semantic relations among words such as synonymy and hyponymy have been employes in text indexing for summarization, information retrieval/extraction (see §4.4.3, 4.2, 4.3.3 and [Sus93, Voo93]).

2.4 Semantic Roles

2.4.1 Introduction

Semantic relations were introduced in generative grammar during the mid-1960s and early 1970s ([Fil68], [Jac72], [Gru67]) as a way of classifying the arguments of natural language predicates into a closed set of participant types which were thought to have a special status in grammar. A list of the most popular roles and the properties usually associated with them is given below.

Agent: A participant which the meaning of the verb specifies as doing or causing something, possibly intentionally. Examples: subjects of *kill, eat, hit, smash, kick, watch.*

Patient: a participant which the verb characterizes as having something happen to it, and as being affected by what happens to it. Examples: objects of *kill, eat, smash* but not those of *watch, hear, love.*

Experiencer: A participant who is characterized as aware of something. Examples: subject of *love*, object of *annoy.*

Theme: A participant which is characterized as changing its position or condition, or as being in a state or position. Examples: objects of *give, hand*, subjects of *walk, die.*

Location: The thematic role associated with the NP expressing the location in a sentence with a verb of location. Examples: subjects of *keep, own, retain, know*, locative PPs.

Source: Object from which motion proceeds. Examples: subjects of *buy, promise*, objects of *deprive, free, cure.*

Goal: Object to which motion proceeds. Examples: subject of *receive, buy*, dative objects of *tell, give.* (adapted from ([Dow89])

In linguistic theory, thematic roles have traditionally been regarded as determinant in expressing generalizations about the syntactic realization of a predicate's arguments (see [EAG96]).

2.4.2 Approaches

The theoretical status of semantic roles in linguistic theory is still a largely unresolved issue. For example, there is considerable doubt about whether semantic roles should be regarded as syntactic, lexical or semantic/conceptual entities. Another open issue, connected with the previous one, is whether semantic roles should be considered a primitive part of linguistic knowledge (see, among others, [Fil68] [Fil77], [Dik89], [Wil81], [Cho86], [Bel88]) or as a derivative notion of some specific aspect of the form-meaning mapping ([Jac72], [Jac90], [Rap88]). However, the most common understanding is that semantic roles are semantic/conceptual elements (see, among others, [Jac72], [Jac90], [Rap88], [Dik89]).

Most characterizations of thematic roles have been carried out in terms of primitive semantic properties of predicates. For example, [Jac72] suggested that thematic relations should be defined in terms of the three semantic subfunctions CAUSE, CHANGE and BE which constitute some of the primitive building blocks of lexical conceptual representations. According to this treatment, the lexical-conceptual representation of a transitive verb like *open* would be as shown below where NP^1 is interpreted as agent and NP^2 as theme.

$$\textsc{cause}\left(\mathrm{NP}^1, \begin{bmatrix} \textsc{change} \\ \textit{physical} \end{bmatrix} (\mathrm{NP}^2, \textsc{not open}, \textsc{open})\right)$$

An analogous proposal was developed by [Dow79] within a Montague Grammar framework and later adopted and extended by [Fol84].

In addition to [Dow79], Other model-theoretic formalizations are presented in [Car84] and [Dow89]. In [Dow89], Dowty defines thematic role types as abstractions above *individual thematic roles* of specific verbs. The *individual thematic role* of a verb is defined as the set of all properties which the verb entails for a given argument position:

Given an n-place predicate δ and a particular argument x_i, the *individual thematic role* $\langle \delta, i \rangle$ is the set of all properties α such that the entailment

$\Box[\delta(x_1, ..., x_i, ..., x_n) \rightarrow \alpha(x_i)]$

holds.

For example the individual role $\langle love, 1 \rangle$ would correspond to the set of properties which can be attributed to the first argument of the predicate *love* through semantic entailment, e.g. the properties which characterize the *lover* participant:

$\langle love, 1 \rangle = \lambda P \exists x \exists y \Box[love(x, y) \rightarrow P(x)]$
where $P \in \{\lambda x \exists y[like(x, y)], \lambda x \exists y[desire(x, y)], ...\}$

A *thematic role type* can then be defined as the intersection of some set of individual thematic roles:

Given a set T of pairs $\langle \delta, i_\delta \rangle$ where δ is an n-place predicate and i_δ the index of one of its arguments (possibly a different i for each verb), a *thematic role type* τ is the intersection of all the individual thematic roles determined by T.

For example, the role type $RECIP$ (recipient) would be the set of all entailed properties shared by a particular individual role of verbs such as *give, sell, buy, receive and tell*, e.g.

$$\begin{array}{lll} RECIP = \lambda Q[& \Box[\lambda x_{1_1}, ..., x_{1_i}, ..., x_{1_n}[\delta_1(x_{1_1}, ..., x_{1_i}, ..., x_{1_n}) \rightarrow Q(x_{1_i})]] & \text{where} \\ & \wedge, ..., \wedge & \\ & \Box[\lambda x_{n_1}, ..., x_{n_i}, ..., x_{n_n}[\delta_n(x_{n_1}, ..., x_{n_i}, ..., x_{n_n}) \rightarrow Q(x_{n_i})]] \quad] & \end{array}$$

$\{\delta_1, ..., \delta_n\} = \{give, sell, buy, receive, tell...\}$

As Dowty himself hastens to point out, this method is not guaranteed to yield useful results. Even assuming that each individual role will effectively intersect with at least another individual role, the number of resulting role types might just be too big to be useful at all. More generally, the identification of an appropriate set of semantic roles [Pal94] is problematic; in practice this means the number of roles varies significantly across different proposals.

The problems just pointed out have led several scholars — e.g. [Som87], [Roz89], [Dow91] — to put forward alternative conceptions of semantic roles. Although from different angles, these all criticize the use of necessary and sufficient conditions for the identification of roles, and advocate more flexible approaches. These approaches appear particularly suitable for the construction of large scale lexicons since they overcome many problems of role identification inherent to traditional approaches, i.e. the difficulty in enumerating precise criteria which qualify the conceptual makeup of a given semantic role.

[Dow91] proposes to abandon the use of discrete role types to provide a total indexing of verbal arguments in favour of a weaker method where the relation between role types and clusters of entailments of verb meanings need not be unique. Dowty assumes that there are only two "thematic-role-like concepts" for verbal predicates: the *proto-agent* and *proto-patient* role. Proto-roles are conceived of as "cluster-concepts" which are determined for each choice of predicate with respect to a given set of semantic properties. The properties which contribute to the definition of the proto-agent and proto-patient roles are listed below.

Contributing Properties for the Proto-Agent Role
volition
sentience (and/or perception)
causes event
movement

Contributing Properties for the Proto-Patient Role
change of state (including coming-to-being, going-out-of-being)
incremental theme (i.e. determinant of aspect, *see section on lexical aspect below*)
vcausally affected by event
stationary (relative to movement of Proto-Agent)

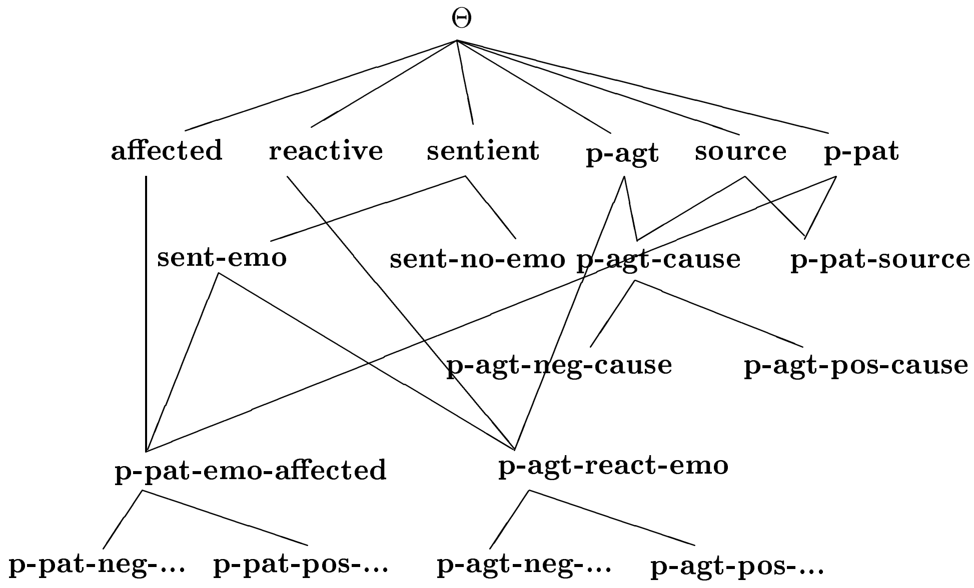
According to Dowty, proto-roles are essentially meant for argument selection, e.g. lexical assignment of grammatical functions to subcategorized arguments.

The work of [Dow91] has been taken as the starting point of the EAGLES recommendations on the encoding of thematic roles [EAG96]).

[San92a, San92b, San93a, San93b] propose to extend the functionality of Dowty's prototype roles by including in the defining clusters properties which are instrumental for the identification of semantic verb (sub)classes. For example, it is well known that at least six subtypes of psychological verbs can be distinguished according to semantic properties of the stimulus and experiencer arguments (see [Jac90] and references therein):

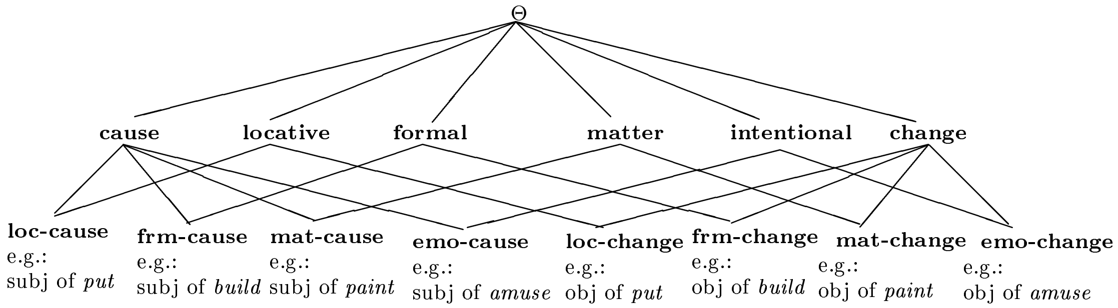
STIMULUS		EXPERIENCER		EXAMPLE
non-causative source		neu.	reactive emotive	*experience*
"		pos.	"	*admire*
"		neg.	"	*fear*
neutral	caus. source	neu.	affected emotive	*interest*
positive	"	pos.	"	*delight*
negative	"	neg.	"	*scare*

This characterization of psychological verbs can be rendered by defining a lattice of thematic sorts relative to the stimulus and experiencer arguments which extend prototype agent and patient roles, e.g.



[Ash95] assume that both *causation* and *change* can be specified along the following dimensions so as to yield a thematic hierarchy such as the one described in the lattice structure below:

- **locative** specifying (the causation of) motion, e.g. subj/obj of *put*
- **formal** specifying the creation and destruction of objects, e.g. subj/obj of *build*
- **matter** specifying (the causation of) changes in shape, size, matter and colour of an object, e.g. subj/obj of *paint*
- **intentional** specifying causation and change of the propositional attitudes of individuals, e.g. subj/obj of *amuse*



[SanFC] proposes to enrich this characterization by

- extending the semantic parameters along which causation and change can be specified using insights from the Austin/Searle's taxonomy of illocutionary acts ([Aus62], [Sea69]), e.g.

verb class	REPRESENTATIVE PREDICATES
assertive	hypothesize, insists, suggest
directive	beg, order, persuade
commissive	agree, promise
expressive	congratulate, apologize, welcome
declarative	declare, fire, resign
perceptive	hear, see, touch
emotive	fear, like, please
formal	build, eat, demolish
matter	paint, knead, carve
locative	send, swim, sit

- providing a specification of *manner*, *direction* and *motion* for the *locative* predicate class, using Talmy's insights on the characterization of locative predicates ([Tal85]) , e.g.

locative verb class	REPRESENTATIVE PREDICATES
$+$*motion*, $+$*manner*	swim, walk
$+$*motion*, $+$*direction*	go, reach
$+$*motion*, $+$*direction*, $+$ *manner*	swim/wal across
$-$*motion*	sit

- providing a definition of source, goal, path and stationary participant in terms of eventuality phases (ONSET, MID, CODA), e.g.

$+$*onset*	source
$+$*mid*	path
$+$*onset*, $+$*mid*	stat
$+$*coda*	goal

2.4.3 Mapping between Semantic Roles and Grammatical Relations

Semantic roles are assumed to be the source of grammatical relations in many linguistic theories. Grammar frameworks such as Government and Binding (GB), Lexical Functional Grammar (LFG) and Functional Grammar (FG) all posit a level of semantic, or thematic, relations to which grammatical relations systematically relate. In particular, semantic roles are the standard devices used for organising predicate argument structures within the lexicon, where arguments are identified on the basis of semantic roles. GB, LFG and FG follow a lexicalist approach to grammar which makes the lexicon the source of syntactic representations; this implies that grammatical relations are, one way or another, projected from predicate argument structures specified within the lexicon.

The principles guiding the mapping of lexical representations onto syntactic structures vary across the different theories. A first distinction can be made between multi-stratal frameworks such as Government and Binding and mono-stratal ones such as Lexical Functional Grammar and Functional Grammar: whereas in the former the mapping is onto D-structure representations, in the latter the projection is directly onto surface representations. Hence, in GB the attention is focused on the way in which thematic roles are mapped onto structural positions at D-structure; the actual syntactic realization of these roles in the surface of the sentence is then accounted for at the level of the mapping between D- and S-structure. By contrast, LFG and FG link semantic relations directly to their surface syntactic realization. From this it follows that the mapping conditions in the two cases are different. In multi-stratal frameworks, D-structure appears to be a pure structural representation of thematic relations, regardless of their syntactic expressions; this could explain why GB lexical representations do not systematically have to specify the syntactic expression of arguments. In mono-stratal frameworks, such mapping conditions have to account for the variation in the syntactic realization of the same semantic relation in the surface of the sentence.

In spite of these different conceptions of the mapping between semantic and syntactic relations, all frameworks considered here share the general assumption that the relationship between semantic and syntactic relations is constrained by some sort of hierarchy of semantic roles. This idea dates back to [Fil68] who first formulated the view that subject selection is in some way sensitive to a hierarchy of "cases", i.e. semantic relations. Following Fillmore, most theories invoke (though to a different extent) a mapping between an ordered list of semantic (i.e. a hierarchy) and an ordered list of grammatical relations (either expressed as different positions within phrase markers, or explicitly or implicitly organized in hierarchical terms). Given a thematic/semantic hierarchy (agent > theme ...) and a syntactic hierarchy (subject > object ...), the general form of the mapping is as follows: map the semantic roles of a given argument structure, which have been ordered according to the hierarchy, into the syntactic hierarchy from left to right. Under this view, the mapping is controlled by hierarchical, i.e. relative, strategies (that is, "higher" semantic roles are mapped onto "higher" syntactic relations), rather than invariable correspondence relations (such as a given semantic role always maps onto a given grammatical relation).

However, this mapping between hierarchies is not always sufficient to predict the syntactic realization of an argument. Whenever this is the case, the mapping is constrained through additional information on the syntactic realization of arguments; this practice is adopted, though to a different extent, within the GB and LFG frameworks. In GB, when the syntactic expression of an arguments cannot be predicted on the basis of general rules, as in the case of psychological verbs, this information is specified at the level of lexical representations in the form of Case-grid [Bel88]. By contrast, LFG lexical representations systematically include the "syntactic function assignment", i.e. the explicit stipulation of the syntactic realization of verb's arguments; according to latest developments [Bre89], this specification is made in underspecified form. Due to its peculiar conception of grammatical relations, FG never contains specifications of this kind: since subject and object selection is made on the basis of pragmatic considerations, the mapping between semantic and syntactic functions only defines the range of possible syntactic realizations, thus stating preferences rather than constraints within the range of possible mappings.

A kind of regular mapping between grammatical relations and semantic roles is also assumed in Dowty's conception of proto-roles and further developments. In fact, proto-roles are related to argument selection through the so-called "Argument Selection Principle", according to which the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate and the argument having the greatest number of Proto-Patient properties will, all else being equal, be lexicalized as the direct object. The basic idea underlying this approach to argument selection is that the ranking according to which the arguments of a verb compete with one another with respect to subjecthood and objecthood is provided by the clustering of semantic properties, rather than by the mapping between specific positions (say between Agent and Subject). This is to say that argument selection of subject and object is determined by the total number of Proto-Agent entailments and Proto-Patient entailments shown by each argument of a verb.

To sum up, three different aspects have been taken as defining features of the kind of mapping between lexical and syntactic representations, i.e. whether:

- lexical representations are mapped either onto D-structure or directly onto the surface structure of the sentence;
- the mapping is constrained by a hierarchy of semantic roles;
- the mapping conditions express constraints or preferences on the syntactic expression of arguments.

All grammar frameworks considered in this brief survey have been characterised with respect to these features.

2.4.4 Comparing Approaches

Semantic roles are dealt with under a number of different names in the linguistic literature, including thematic relations, participant roles, deep cases, semantic case/roles and theta roles. Though many of these terminological distinctions carry specific implications regarding theoretical status, there is a shared underlying intuition that certain characteristic "modes of participation" in an event can be identified and generalised across a variety of verbs. Such modes of event participation are spelled out in terms of basic concepts such as *cause, change, be*. All the approaches reviewed do in some sense follow this practice, although they differ as to the number and type of basic concepts used.

Concerning formalization, three treatments can be distinguished. First, approaches which rely on an informal specification such as Jackendoff's lexical conceptual structures [Jac90]. Second, approaches such as those proposed by [Dow79], [Dow89] which are developed within a model-theoretic framework. Third approaches which provide an algebraic specification within a typed feature structure formalism; these tend to be more oriented towards NLP applications (e.g. [San92b], [San93a], [San93b], [SanFC]).

2.4.5 Relation to other Areas of Lexical Semantics

Lexical aspect Since Dowty's and Verkuyl's pioneering work on aspect compositionality during the early seventies [Dow72], [Ver72], it has been a well known fact that many non-stative verbs can give rise to either a telic or atelic interpretation according to whether their theme NP has *quantized* or *cumulative* reference. Informally, a nominal predicate has quantized reference if it responds positively to the additivity test, e.g. sugar and sugar makes sugar. Conversely, with a quantized nominal there is no proper subpart of the nominal which has the same denotational properties of the NP, e.g. no proper subpart of a chair is a chair. The basic generalization concerning the interaction of nominal and temporal reference with respect to aspect compositionality can be briefly stated as follows:

> a theme NP which has cumulative reference induces a durative reading at the sentential level, while with a theme NP which has quantized reference a terminative reading obtains.

This pattern is shown in the examples below where a sequence of two question marks indicates incompatibility — under a single event reading — between a quantified NP and the durative adverbial *all day* which forces an atelic (i.e. durative) interpretation on the sentences.

(14) a. John drank beer all day

b. ?? John drank a glass of beer all day

[Kri89], [Kri90] has argued that the contribution of nominal reference to sentence aspect should be characterized by establishing a homomorphism between algebraically structured NP and event denotata in such a way that subsequent stages of change in the NP are reflected in developmental stages of the event. Consider, for example, a sentence such as *John drank a glass of wine*. An intuitive way of characterizing Krifka's approach would be to say that by monitoring the level of wine in the glass we would be able to establish whether the event is at its outset, halfway done, completed etc.: for each subpart of liquid in the glass there is a subevent of drinking. Another important feature of Krifka's approach consists in regarding thematic roles as links between nominal reference and temporal constitution. More specifically, some thematic roles are regarded as having transfer properties which allow the object-to-event homomorphism referred above to take place.

2.4.6 Encoding in Lexical Databases

The classification of suggested in [San92b] and [San93a] has been used in the ACQUILEX lexicons (§3.12.3) to provide representations for psychological and motion verbs in English (see [Cop93]).

Thematic roles are also used in the EUROTRA MT lexica (§3.11.1), DELIS (§3.12.5), the EDR Concept Description Dictionary (§3.8).

2.4.7 Relevance to LE Applications

Thematic or semantic roles represent another type of grammatical relation - a semantic one - holding between a predicate and its arguments, which can be usefully exploited in the framework of NLP applications. Information on the thematic role borne by arguments with respect to their predicates is useful to abstract away from the issue of their concrete syntactic realization (e.g. in terms of grammatical functions such as subject, object indirect object and the like). In Machine Translation, for instance, this information can be used to map the predicate-argument structures of two translation equivalents such as, for example, English *like* and Italian *piacere*, although the syntactic realization of the two verbs is radically different: namely, what is the subject of English *like* becomes an indirect object in Italian *piacere*, while what is the object of *like* is turned into the subject in the Italian translation. A characterization of the argument structure of the two verbs in terms of semantic roles attains the purpose of neutralizing their different syntactic behaviour in the specific language. Thematic roles have been used in Machine Translation (§4.1) to express generalizations about complex translation equivalence (see [Dor90], [Dor93], [San92b], [Cop93]).

Another possible use of semantic roles in the framework of NLP applications is based on their semantic content which, in principle, can be used to infer the semantic role borne by a given constituent in a sentence. In a syntactically - either structurally or functionally - ambiguous context, recognition of the semantic role borne by a given constituent can help to resolve the syntactic ambiguity due to the pervasive regularities observed in the mapping between semantic roles and grammatical relations (see §2.4.3). This kind of hypothesis is entertained and corroborated by psycholinguistic studies on language comprehension showing that thematic information is highly instrumental in the resolution of local ambiguities and garden-paths [Car88], [Pri88], [Sto89]. In particular, the results of the experiments carried out in these studies show that there appears to be a persistent default pattern of thematic assignment throughout natural language - animate subjects are Agents and objects are Themes, inanimate subjects are Themes. These results validate the hypothesis that thematic assignments have a bearing on syntactic parsing which is thus performed on the basis of semantic information. Yet, when the exploitation of this hypothesis is considered for disambiguation purposes in the framework of wide coverage NLP systems, the picture emerging from psycholinguistic studies changes radically due to the impossibility of having a coherent characterisation of semantic roles while keeping their being the source of coherent syntactic representations [Mon95].

2.4.8 Glossary

Thematic role, semantic role, deep case, semantic case, agent, patient, experiencer, theme, location, source, goal, incremental theme.

2.5 Lexicalization

2.5.1 Introduction

One of the basic goals of lexical semantic theory is to provide a specification of word meanings in terms of semantic components and combinatory relations among them. Different works in lexical semantics converge now on the hypothesis that the meaning of every lexeme can be analysed in terms of a set of more general meaning components, some or all of which are common to groups of lexemes in a language or cross-linguistically. In other words, meaning components can be identified which may or may not be lexicalized in particular languages. The individuation of the meaning components characterising classes of words in a language and of the possible combinations of such components within word roots leads to the identification of lexicalization patterns varying across languages. Moreover there is a strong correlation between each combination of meaning components and the syntactic constructions allowed by the words displaying them (e.g., [Tal85], [Jac83], [Jac90]).

A trend has recently emerged towards addressing the issues of

i identifying meaning components lexicalized within verb roots;

ii stating a connection between specific components characterizing semantic classes of verbs and syntactic properties of the verbs themselves (e.g. [Lev93], [Alo94b], [Alo95]).

The basic goals of research on lexicalization of meaning components are:

- to define a (finite?) set of (primitive? universal? functionally discrete?) meaning components;
- to provide a description of word meanings in terms of meaning components and combinatory relations among them;
- to identify 'preferences' displayed by (groups of) languages for lexicalization patterns;

- to identify linkings between each meaning components 'conflation' pattern and syntactic properties of words.

The main aims of this section are firstly to point out some problematic issues raised in works dealing with the identification and discussion of meaning components; then, to briefly discuss proposals concerned with the identification of lexicalization patterns of semantic components both in a language and cross-linguistically. Furthermore, we shall point out how information on lexicalization is eventually encoded in lexical databases and useful for LE applications.

2.5.2 Description and comparison of different approaches

Works on meaning components

Representing complex meanings in terms of simpler ones has generally been considered as one of the fundamental goals of semantic theory, however different positions have been taken with respect to various aspects of the issue in works devoted to it. In any case, the following hypotheses are shared by the various positions:

- the meaning of every lexeme can be analysed in terms of a set of more general meaning components;
- some or all of these components are common to groups of lexemes in a language/cross-linguistically.

It was [Hje61] componential analysis of word meaning which gave rise to various researches of the same type in Europe (e.g., [Gre66], [Cos67], [Pot74], etc.). These researches, although different with respect to the specific hypotheses put forward, tried to identify semantic components shared by groups of words by observation of paradigmatic relations between words. The semantic components identified in the various proposals differentiate both a group of words from another and, by combining in various ways, a word from another. A standard example is reported in the following, showing the kind of analysis usually performed:

woman : man : child :: mare : stallion : foal :: cow : bull : calf

In this set of words, all the words in a group contrast with the words in another group in the same way (i.e. because of the same semantic components), and the first word in each group contrasts with the other words in its group in the same way, etc. Thus, for instance, all the words in the first group will be assigned a component HUMAN, vs. EQUINE and BOVINE assigned respectively to the second and third group. Then, the first word in each group will be characterized by a component FEMALE, vs. MALE characterizing the second word, etc.

Componential analysis in America developed independently firstly among anthropologists (e.g., [Lou56], [Goo56]) who described and compared kinship terminology in various languages. Their research was taken up and generalized by various linguists and in particular by scholars working within the framework of transformational grammar (cf. [Kat63]), who aimed at integrating componential analyses of words with treatments of the syntactic organization of sentences. Generative semanticists (e.g. [McC68], [Lak70]) tried to determine units of meaning, or 'atomic predicates', by means of syntagmatic considerations. Thus, for instance, the components BECOME and CAUSE were identified by analysing pairs of sentences displaying similar syntactic relationships such as the following:

(15) a. The soup cooled.

b. The metal hardened.

c. John cooled the soup.

d. John hardened the metal

Afterwards, scholars working in different fields of the research on language dealt with various issues connected with the identification/definition of meaning components. Within this survey we do not intend to report on all the similarities/differences among the various hypotheses put forward. We shall instead point out problematic aspects which have been dealt with and which can be of interest for our work.

The most important issues raised in the works on semantic components are the following:

- first there is the question if the meaning components which have been identified/can be identified should be considered as 'primitives' or not: i.e., if they are linguistic/conceptual units of some kind from which all possible meanings in a language can be derived, but which in turn are not themselves derivable from any other linguistic unit;

- strictly linked up to the above issue is that of the 'universality' of primitives, i.e. if such primitives are the same across languages;

- then, there is the question if it is possible to identify a finite set of (universal) primitives;

- finally, there is the question of identifying a procedure of definition of semantic components.

These issues have been explicitly or implicitly dealt with in theoretical semantic research, in the computational field, in philosophy, in psycholinguistics. The 'strongest' proposal put forward with respect to them is probably that presented by Wierzbicka in a number of works on *semantic primitives* (cf. [Wie72], [Wie80], [Wie85], [Wie89a], [Wie89b]). The avowed goal of these works is just to arrive at a definition of a complete and stable set of *semantic primitives*, by means of cross-linguistic research on *lexical universals*. These are concepts which are encoded in the lexica of (nearly or possibly) all natural languages. While lexical universals are not necessarily universal semantic primitives (e.g., a concept such as *mother*), according to Wierzbicka the contrary is true, i.e. all semantic primitives are universal. Decisive for succeeding in identifying such semantic primitives are large scale lexicographic studies. These studies should not rely on research of the most frequent words recurring in the definitions of conventional dictionaries, due to all the limits, inchoerences and lack of data which are typically evidenced in these sources. In the various stages of her research, Wierzbicka postulated different sets of primitives. While the first set included only 14 elements, in [Wie89b] a set of twenty-eight universal semantic primitive candidates was proposed:

I, you, someone, something, this, the same (other), two, all, I want, I don't want, think, say, know, would (I imagine), do happen, where, when, after, like, can (possible), good, bad, kind (of), part, like, because, and very.

According to the author, this list should not necessarily be considered as 'final'. In any case, the set 'works' in semantic analyses and has been validated through research into lexical universals.

In general, studies dealing with meaning components treat them as 'primitives', i.e. as units which cannot be further defined. However, the components indicated as primitives in certain works are not always accepted as such in others (cf. Jackendoff's discussion in [Jac83] of the proposal for a primitive ALIVE in [McC68]).

Sometimes, a strong relation between 'primitivity' and 'universality' is not explicitly stated. For instance, [Mel89] conceives semantic primitives simply as 'elementary lexical meanings of a particular language' without wondering if they are the same for all the languages. However, others, and especially scholars working within a Chomskian framework, assume the universality of semantic primitives, in that they share the position that the meaning components which are lexicalized in any language are taken from a finite inventory, the knowledge of which is innate (e.g. [Jac90]).

The main problem remains, then, to decide which the universal semantic primitives are: i.e., to (eventually) define a finite and complete set of them. Indeed, while Wierzbicka proposes a complete and 'stable' (although not necessarily definitive) set of (pure) primitives, strong hypotheses like hers have not in general been presented. Rather, analyses of portions of the lexicon have been proposed: for instance, [Tal76] describes the various semantic elements combining to express causation; [Tal85] discusses the semantics of motion expressions; [Jac83], [Jac90] extends to various semantic fields semantic analyses provided for motion or location verbs (e.g., verbs of transfer of possession, verbs of touching, etc.); etc. In any case, by analysis and comparison of different works on the issue, we cannot circumscribe a shared set of primitives which could also be seen as 'complete'.

Finally, no clear procedure of identification of semantic components has been so far formalized.

An approach which deliberately wants to avoid a strong theoretical characterization of semantic components is that chosen by [Cru86], which, for this reason, could be taken as the starting point for the Eagles recommendations on the encoding of semantic components. According to Cruse's 'contextual approach', the meaning of a word can be described as composed of the meanings of other words with which it contracts paradigmatic and syntagmatic relations within the lexicon. These words are named *semantic traits* of the former word. Thus, for instance, *animal* can be considered a semantic trait of *dog*, since it is its hyperonym. Moreover, *dog* is implied in the meaning of *to bark*, given that it is the typical subject selected by the verb. Cruse clearly states that his 'semantic traits' are not claimed to be

> primitive, functionally discrete, universal, or drawn from a finite inventory; nor it is assumed that the meaning of any word can be exhaustively characterised by any finite set of them ([Cru86], p. 22).

A similarly weakly theoretically characterised approach has been taken by [Dik78], [Dik80] with his 'stepwise lexical decomposition'. No semantic primitive/universal elements are postulated. Lexical meaning is reduced to a limited set of basic lexical items of the object language, identified by analysing a network of meaning descriptions.

Works on Lexicalization patterns

Relying on the basic assumption that it is possible to identify a discrete set of elements (semantic components) within the domain of meaning and combinatory relations among them, [Tal85] carried out a study on the relationships among such semantic components and morphemes/words/phrases in a sentence/text. In particular, he deeply investigated the regular associations (lexicalization patterns) among meaning components (or sets of meaning components) and the verb, providing a cross-linguistic study of lexicalization patterns connected with the expression of motion. He was mainly interested in evidencing *typologies*, i.e. small number of patterns exhibited by groups of languages, and *universals*, i.e. single patterns shared cross-linguistically.

According to Talmy, a motion event may be analysed as related, at least, to five basic semantic elements:

- MOTION (the event of motion or location),
- PATH (the course followed or site occupied),
- MANNER (the manner of motion),
- FIGURE (the moving object),
- GROUND (the reference object).

These may be found either lexicalized independently of one another, or variously *conflated* in the meaning of single words, as can be seen in the examples below (all taken from [Tal85], except the last one):

(16) a.

The rock	moved	down	the hill	rolling
FIGURE	MOTION	PATH	GROUND	MANNER

b.

The rock	rolled	down	the hill
FIGURE	MOTION + MANNER	PATH	GROUND

c.

La botella	entró	a	la cueva	flotando
the bottle	moved-in	to	the cave	floating
FIGURE	MOTION + PATH	PATH	GROUND	MANNER

d.

She	powdered	her nose
	MOTION + PATH + FIGURE	GROUND

e.

I	shelved	the books
	MOTION + PATH + GROUND	FIGURE

f.

L'uomo	fuggì
the man	escaped
FIGURE	MOTION + PATH + MANNER

Firstly, Talmy presents three basic lexicalization types for verb roots which are used by different languages in their most characteristic expression of motion:

1. **MOTION + MANNER/CAUSE**
2. **MOTION + PATH**
3. **MOTION + FIGURE**

Talmy provides examples of these patterns of conflation:

- the first one is found in the roots of e.g.

- *stand* in *The lamp stood on the table*;
- *roll* in *The rock rolled down the hill*;
- *push* in *I pushed the keg into the storeroom*[7].

This pattern is typical of English but not, for instance, of Spanish (or, we could also say, of Italian), which expresses the same meanings with different constructions as in e.g. *Metí el barril a la bodega rodandolo* = I rolled the keg into the storeroom.

- The second pattern is typically displayed by Semitic, Polynesian and Romance languages, but not by English: whereas e.g. in Spanish we find *El globo bajó por la chimenea flotando* and *La botella cruzó el canal flotando*, in English we would find *The ballon floated down the chimney* and *The bottle floated across the canal.*
- Finally, the third major typological pattern is displayed in a few English forms (e.g. *I spat into the cuspidor*, but an example *par excellence* of this type is Atsugewi, a Hokan language of northern California.

Another interesting issue discussed by Talmy is the possibility to have an extension of the first pattern seen far beyond the expression of simple motion in English, in which, e.g. MOTION and MANNER can be compounded with mental-event notions (e.g. *I waved him away from the building*), or with specific material in recurrent semantic complexes (e.g. *I slid him another beer*), etc.

Other combinatorial possibilities are considered which, however, seem to form minor systems of conflation. Furthermore, also a 'hierarchy' of conflation types is proposed, where the conflation involving PATH is considered as the most extensively represented, next there is the MANNER/CAUSE, and finally the FIGURE one. Some remarks are added on the possibility to have GROUND conflated with MOTION, which is however only sporadically instantiated (e.g. *emplane*). Further discussion is provided of lexicalization of aspect, causation etc. and of the relations between meaning components and other parts-of-speech apart from the verb. This does not however seem relevant for our purposes and, in any case, we believe that the issues treated raise problems which should not be discussed here.

Interesting discussion of lexicalization patterns are found in [Jac83], [Jac90]. His theory of Conceptual Semantics and the organization of Lexical Conceptual Structure are discussed in details in the following section. We shall only briefly recall some points of interest for our purposes. The different main elements of the LCS language are: conceptual constituents, semantic fields and primitives. Then there are other elements, like conceptual variables, semantic features, constants, and lexical functions, which play minor roles. Each conceptual constituent belongs to one of a small set of ontological categories such as *Thing, Event, State, Action, Place, Path*, etc. Among conceptual primitives the main ones are BE, which represents a state, and GO, which represents any event. Other primitives include: STAY, CAUSE, INCH, EXT, etc. A second larger set of primitives describes prepositions: AT, IN, ON, TOWARD, FROM, TO, etc.

The LCS organization incorporates [Gru67] view, according to which the formalism used for encoding concepts of spatial location and motion can be abstracted and generalized to many other semantic fields (cf. next section). Thus, Jackendoff tries to extend to a wide range

[7]Although this is not really clear from Talmy's discussion, 'motion' is used in this case to refer to what he elsewhere names *translational motion*, i.e. change of position, rather than to *contained motion*, i.e. motion along an unbounded path.

of other semantic fields semantic analyses provided for motion or location verbs. This turns out in requiring an additional elaboration of his conceptual system. At the same time, observations are added on the various correspondences between different lexicalization patterns and syntactic expressions. An interesting proposal put forward by Jackendoff (developing a suggestion from [Car88]) concerns a distinction between a MOVE-function and a GO-function: manner-of-motion verbs which cannot occur with complements referring to a PATH (more precisely, a bounded path) should only be linked to a MOVE-function. A rule is then proposed to account for (typically English) sentences containing manner-of-motion verbs allowing directional complements: a sentence like *Debbie danced into the room* expresses a conceptual structure that includes both a MOVE-function and a GO-function (indicating change of position). What differentiates English manner-of-motion verbs from, e.g., Spanish ones is the possibility to allow incorporation of what Jackendoff calls a *GO-Adjunct*.

Both Talmy and Jackendoff evidenced a strict correlation between the meaning components clustered within a verb root and the verb syntactic properties. An extensive study on the correlation between verb semantics and syntax has been provided by [Lev93]. This study shows that verb semantic classes can be identified, each characterized by particular syntactic properties (see §2.6.2).

Within the Acquilex project (§3.12.3) work has been carried out to identify information on lexicalization of meaning components and to connect such information to the syntactic properties of verbs. MRD definitions of some classes of verbs (e.g., verbs referring to motion, to change-of-state-by-cooking, etc.) were analysed in order to link recurrent patterns to specific meaning components characterizing each class in a specific language. Furthermore, connections were stated between single components and syntactic properties displayed by the verbs under analysis (e.g. [Alo94a], [Tau94]).

Within the EuroWordNet project (§3.6.3) relations between words are being encoded which allow to gather data on lexicalization. For instance, information on arguments *involved* in verb meaning are being encoded and compared cross-linguistically (cf. [AloFC]).

2.5.3 Relation to other areas of lexical semantics

The kinds of meaning components 'conflated' within verb roots are strongly correlated with the syntactic properties of the verbs themselves, i.e. with the possibility of verbs occurring with certain arguments (e.g. [Tal85, Lev93] and §2.4). Moreover, a clear identification of the semantic components conflated within verb roots in individual languages could be relevant also for isolating semantic classes displaying, or amenable to, similar sense extensions, given that amenability to yield different interpretations in context appears to be connected with semantic characteristics which verbs (words) share (cf. [San94]).

By adopting a strongly 'relational' view of the lexicon, then, we may identify lexicalization patterns by stating paradigmatic/syntagmatic relations between words (cf. work carried out within EuroWordNet). Thus, research on lexicalization is strictly linked to work on lexical relations such as hyponymy, meronymy, etc.

2.5.4 How is information encoded in lexical databases

The work carried out within the Acquilex project led to identify semantic components lexicalized within the roots of various verb classes. The information acquired is variously encoded in the language-specific LDBs. Furthermore, part of this information was encoded within the multilingual LKB by linking the relevant meaning components to the participant role types involved by verb meaning. For instance, the subject of the English verb *swim* was associated with the participant role type **proto-agent-cause-move-manner**[8], indicating that the verb involves self-causing, undirected motion for which manner is specified (cf. [San92b]).

Much information on lexicalization patterns is being encoded within the EuroWordNet database for substantial portions of the lexica of various languages. Here, information on semantic components lexicalized within word meanings is encoded by means of lexical relations applying between synsets (see §3.6.2).

2.5.5 LE applications

Results of research on lexicalization seem necessary for a variety of NLP tasks and applications. Data on lexicalization can be useful for Word Sense Disambiguation (WSD) and related applications ranging from Machine Translation (§4.1) to Information Retrieval (§4.3) and NL generation (§4.5) because of

- the strict correlation between the meaning components involved in a word root and its syntactic properties, and
- the cross-linguistic differences in the meaning components conflation within word roots,

2.5.6 Glossary

Lexicalization: the process according to which a meaning component or a set of meaning components is associated with a 'surface expression' (i.e. a word root). **Lexicalization pattern**: the regular association of sets of meaning components and word roots. **Componential analysis**: method of representation of lexical meaning, according to which the meaning of a lexical item can be decomposed into more basic semantic units (components), on the basis of either paradigmatic or syntagmatic relations to other elements in the lexicon. **Semantic/meaning component**: a basic semantic unit, generally assumed to be a 'primitive', i.e. non further definable. **Semantic traits**: sometimes used with the same meaning of 'semantic component'. [Cru86] uses this expression as a weakly theoretically characterized expression, to avoid implying reference to notions such as 'primitiveness', 'universality', etc.

One of the basic goals of lexical semantic theory is to provide a specification of word meanings in terms of semantic components and combinatory relations among them. Different works in lexical semantics converge now on the hypothesis that the meaning of every lexeme can be analysed in terms of a set of more general meaning components, some or all of which are common to groups of lexemes in a language or cross-linguistically. In other words, meaning components can be identified which may or may not be lexicalized in particular languages. The individuation of the meaning components characterising classes of words in a language and of the possible combinations of such components within word roots leads to the identification of lexicalization patterns varying across languages. Moreover there is a strong correlation between each combination of meaning components and the syntactic constructions allowed by the words displaying them (e.g., [Tal85], [Jac83], [Jac90]). A trend has recently emerged towards addressing the issues of

[8]see §2.3.2, 3.12.3 for a description of the treatment chosen to encode information on thematic relations within Acquilex

i identifying meaning components lexicalized within verb roots;

ii stating a connection between specific components characterizing semantic classes of verbs and syntactic properties of the verbs themselves (e.g. [Lev93], [Alo95], p.94, [Alo95]).

The basic goals of research on lexicalization of meaning components are:

- to define a (finite?) set of (primitive? universal? functionally discrete?) meaning components;
- to provide a description of word meanings in terms of meaning components and combinatory relations among them;
- to identify 'preferences' displayed by (groups of) languages for lexicalization patterns;
- to identify linkings between each meaning components 'conflation' pattern and syntactic properties of words.

The main aims of this section are firstly to point out some problematic issues dealt with in works which dealt with the identification and discussion of meaning components; then, to briefly discuss proposals concerned with the identification of lexicalization patterns of semantic components both in a language and cross-linguistically. Studies dealing with the interrelationship between meaning components and syntactic properties of words will also be briefly taken into consideration, although a more detailed discussion of this issue will be provided in §2.5. Furthermore, we shall point out how information on lexicalization is eventually encoded in lexical databases and used in/useful for LE applications.

2.6 Verb Semantic Classes

2.6.1 Introduction

The following approaches to building verb semantic classes are outlined in this section: verb classes based on syntactic behaviour (alternations), and verb classes formed from semantic criteria such as thematic roles and elements of the Lexical Conceptual Structure. Classifications related to WordNet criteria are disucssed in the section devoted to WordNet (§3.6.2, 3.6.3). Each of these approaches contribute to a different form of classification, whose usefulness and ease of formation will be evaluated.

The main practical aim of verb semantic classifications is to contribute to structure the lexicon and to allow for a better organized description, more homogeneous, of their semantics. On a more formal point of view, the main aims are the identification of meaning components forming the semantics of verbs, the specification of the more subtle meaning elements that differentiate closely related verbs and the study of the cooperation between syntax and semantics.

2.6.2 Description of different approaches

Syntactic Alternations and Verb Semantic Classes

In her book, B. Levin [Lev93] shows, for a large set of English verbs (about 3200), the correlations between the semantics of verbs and their syntactic behavior. More precisely, she shows that some facets of the semantics of verbs have strong correlations with the syntactic behavior of these verbs and with the interpretation of their arguments.

She first precisely delimits the different forms of verb syntactic behavior. Each of these forms is described by one or more *alternation* (e.g. alternations describe passive forms, there-insertions and reflexive forms). Then, she proposes an analysis of English verbs according to these alternations: each verb is associated with the set of alternations it undergoes. A preliminary investigation showed that there are sufficient correlations between some facets of the semantics of verbs and their syntactic behavior to allow for the formation of classes. From these observations, Beth Levin has then defined about 200 verb semantic classes, where, in each class, verbs share a certain number of alternations.

This very important work emerged from the synthesis of specific investigations on particular sets of verbs (e.g. movement verbs), on specific syntactic behaviors and on various types of information extracted form corpora. Other authors have studied in detail the semantics conveyed by alternations e.g. [Pin89] and the links between them [Gol94].

The alternation system An alternation, roughly speaking, describes a change in the realization of the argument structure of a verb. The scope of an alternation is the proposition. Modifiers are considered in some cases, but the main structures remain the arguments and the verb. Arguments may be deleted or 'moved', NPs may become PPs or vice-versa, and some PPs may be introduced by a new preposition. Alternations may also be restricted by means of constraints on their arguments.

Beth Levin has defined 79 alternations for English. They basically describe 'transformations' from a 'basic' form. However, these alternations have *a priori* little to do with the assumptions of Government and Binding theory and Movement theory, in spite of some similitudes. The form assumed to be basic usually corresponds to the direct realization of the argument structure, although this point of view may clearly be subject to debate. Here are now a few types of alternations, among the most common ones.

The *Transitivity alternations* introduce a change in the verb's transitivity. In a number of these alternations the subject NP is deleted and one of the objects becomes the subject, which must be realized in English. The *Middle alternation* is typical of this change:

(17) *John cuts the cake* $\rightarrow$ *The cake cuts easily.*

As can be noticed, it is often necessary to add an adverb to make the sentence acceptable. The *Causative/inchoative alternation* concerns a different set of verbs:
Edith broke the window $\rightarrow$ *The window broke.*
Verbs undergoing this alternation can roughly be characterized as verbs of change of state or position.

Under the transitivity alternations fall also alternations where an object is unexpressed. This is the case of the *Unexpressed object alternation* where the object1 is not realized. A number of verbs undergo this alternation. In most cases, the 'typical' object is somewhat 'implicit' or 'incorporated' into the verb, or deducible from the subject and the verb. This is the case, e.g., for the *Characteristic property of agent alternation*:

(18) *This dog bites people* $\rightarrow$ *This dog bites.*

We also find alternations that change the object NP into a PP, as in the *conative alternation*:

(19) *Edith cuts the bread* $\rightarrow$ *Edith cuts at the bread.*

Other sets of alternations include the introduction of oblique complements, reflexives, passives, there-insertion, different forms of inversions and the introduction of specific words such as the way-construction.

It is clear that these alternations are specific to English. They are not universal, even though some are shared by several languages (e.g. the passive alternation). Every language has its own alternation system, and has a more or less important number of alternations. The characteristics of the language, such as case marking, are also an important factor of variation of the form, the status and the number of alternations. English seems to have a quite large number of alternations, this is also the case e.g. for ancient languages such as Greek. French and Romance languages in general have much fewer alternations, their syntax is, in a certain way, more rigid. The number of alternations also depends on the way they are defined, in particular the degree of generality via constraints imposed on context elements is a major factor of variation.

Construction of verb semantic classes Verb semantic classes are then constructed from verbs, modulo exceptions, which undergo a certain number of alternations. From this classification, a set of verb semantic classes is organized. We have, for example, the classes of verbs of putting, which include Put verbs, Funnel Verbs, Verbs of putting in a specified direction, Pour verbs, Coil verbs, etc. Other sets of classes include Verbs of removing, Verbs of Carrying and Sending, Verbs of Throwing, Hold and Keep verbs, Verbs of contact by impact, Image creation verbs, Verbs of creation and transformation, Verbs with predicative complements, Verbs of perception, Verbs of desire, Verbs of communication, Verbs of social interaction, etc. As can be noticed, these classes only partially overlap with the classification adopted in WordNet. This is not surprising since the classification criteria are very different.

Let us now look in more depth at a few classes and somewhat evaluate the use of such classes for natural language applications (note that several research projects make an intensive use of B. Levin's classes). Note that, w.r.t. WordNet, the classes obtained via alternations are much less hierarchically structured, which shows that the two approaches are really orthogonal.

There are other aspects which may weaken the practical use of this approach, in spite of its obvious high linguistic interest, from both theoretical and practical viewpoints. The first point is that the semantic definition of some classes is somewhat fuzzy and does not really summarize the semantics of the verbs it contains. An alternative would be to characterize a class by a set of features, shared to various extents by the verbs they are composed of. Next, w.r.t. the semantic characterization of the class, there are some verbs which seem to be really outside the class. Also, as illustrated below, a set of classes (such as movement verbs) does not include all the 'natural' types of classes one may expect (but 'completeness' or exhaustiveness has never been claimed to be one of the objectives of this research). This may explain the unexpected presence of some verbs in a class. Finally, distinctions between classes are sometimes hard to make, and this is reinforced by the fact that classes may unexpectedly have several verbs in common. Let us illustrate these observations on two very representative sets of classes: verbs of motion and verbs of transfer of possession (notice that a few other classes of transfer of possession, e.g. deprivation, are in the set of classes of Remove verbs).

Verbs of *Motion* include 9 classes:

- Inherently directed motion (arrive, go,...),
- Leave verbs,

- Manner of motion:
  - Roll verbs (bounce, float, move, ...),
  - Run verbs (bounce, float, jump, ...),
- Manner of motion using a vehicle:
  - Vehicle name verbs (bike, ...),
  - Verbs not associated with vehicle names (fly,..),
- Waltz verbs (boogie, polka, ...),
- Chase verbs (follow, pursue, ...),
- Accompany verbs.

Note that the labels 'Roll' and 'Run' do not totally cover the semantics of the verbs in the corresponding class. Also, the difference between the two classes is not very clear. Waltz and chase verbs are interesting examples of very specific classes which can be constructed from alternations. However, few domains are represented, and major ones are missing or under-represented (e.g. type of movement, medium of movement, manner of motion, etc.).

Verbs of *transfer of possession* include 9 classes:

- Give verbs (feed, give, lease, ...),
- Contribute verbs (distribute, donate, submit, ...),
- Providing:
  - Fulfilling verbs (credit, provide, ...),
  - Equip verbs (arm, invest, ...),
- Obtaining:
  - Get (book, buy, call, cash, order, phone, ...),
  - Obtain (accept, accumulate, seize, ...),
- Future having verbs (advance, assign, ...),
- Exchange verbs,
- Berry verbs (nest, clam, ...).

In this example, the difficulty of defining the semantics of a class is evident, e.g.: fulfilling, future having: these terms do not exactly characterize the class. Note also the Get class is very large and contains very diverse verbs. Domain descriptions (family, education, law, etc.) as well as moral judgements on the transfer (legal, illegal, robbery) are not accounted for in this classification.

About the underlying semantics of alternations It is of much interest to analyze in depth the set of verbs which undergo an alternation. It is also interesting to analyze exceptions, i.e. verbs not associated with an alternation but which are closely related to verbs which are associated with it, in order to narrow down the semantic characterization of this alternation.

Besides the theoretical interest, the underlying semantics conveyed by syntactic construction plays an important role in semantic composition and in the formation of lexicalization patterns (§2.5.1).

There is, first, a principle of non-synonymy of grammatical forms: 'a difference in syntactic form always spells a difference in meaning' which is commonly assumed. We have, for example, the following syntactic forms with their associated Lexical Semantic Template (Goldberg 94):

- Ditransitive: X CAUSES Y to RECEIVE Z,
- Caused Motion: X CAUSES Y to MOVE Z,
- Resultative: X CAUSES Y to BECOME Z,
- Intransitive Motion: X MOVES Y,
- Conative: X DIRECTS action AT Y.

From these general observations, we see that form and meaning cannot be considered apart. From the point of view of the principle of compositionality, the meaning of a sentence should not only be derived from the meaning of its components, but it should also include the implicit, partial semantics associated with the syntactic construction. Let us now consider several examples.

About the identification of relevant meaning components The problem addressed here is the identification in verbs of those meaning components which determine the fact that a verb undergoes or not a certain alternation. ([Pin89], pp. 104 and following), explains that in the conative construction, where the transitive verb takes an oblique object introduced by the preposition *at* instead of a direct NP, there is the idea that the subject is attempting to affect the oblique object, but may not succeed. But the conative alternation applies to much narrower sets of verbs than those whose actions could be just attempted and not realized. For example, verbs of cutting and verbs of hitting all undergo the alternation, but verbs of touching and verbs of breaking do not.

It turns out, in fact, that verbs accepting the conative construction describe a certain type of motion and a certain type of contact.

The same situation occurs for the *Part-possessor ascension alternation* (*Ann cuts John's arm* ↔ *Ann cuts John on the arm*) which is also accepted by verbs of motion followed by contact. Here verbs of breaking do not participate in that alternation whereas verbs of hitting and touching do.

Finally, the *Middle alternation*, which specifies the ease with which an action can be performed on a theme, is accepted only by verbs that entail a real effect, regardless of whether they involve motion or contact. Therefore, verbs of beaking and of cutting undergo this alternation whereas verbs of touching do not.

As can be seen from these examples, a common set of elements of meaning can be defined for a set of alternations, such as motion, contact and effect, which contributes to differentiating the semantics conveyed by alternations, and therefore to characterizing quite precisely verbs which potentially undergo and alternation or not. Therefore, membership of a verb to a class depends on some aspects of meaning that the semantic representation of the verb constrains. These aspects may moreover be surprisingly subtle and refined, and difficult to identify and to describe in a formal system. These observations reinforces the arguments in favor of a *certain autonomy of lexical semantics.*

The dative alternation The dative alternation applies to a number of verbs of transfer of possession, but the semantic components which account for the difference between verbs which do accept it and those which do not are really subtle. This alternation conveys the idea of X CAUSE Y to HAVE Z. However, as noted by [Pin89], while the class of verb of *instantaneous imparting of force causing a ballistic motion* (throw, flip, slap) allow the dative alternation, the *verbs of continuous imparting of force in some manner causing accompanied motion* do not (pull, push, lift).

Similarly, verbs where "X commits himself that Y will get Z in the future" allow the dative alternation (offer, promise, allocate, allot, assign). There are also verb classes which accept either one or the other form of the dative alternation (with or without the preposition *to*). Verbs of 'long-distance' communication (fax, telephone) also accept this alternation.

From these examples, it is possible to deduce that the dative alternation is accepted by verbs where the actor acts on a recipient (or a destination) in such a way that causes him to possess something. This is opposed to the actor acting on an object so that it causes it to go to someone. For example, in verbs like *push*, the actor has not *a priori* in mind the destination, but just the object being pushed. On the contrary, *ask* accepts the dative alternation because when someone is asking something he has (first) in mind the way the listener will react, the 'physical' transfer of the information is in general less important.

The location alternations The location alternations (a family of alternation which involve a permutation of object1 and object2 and a preposition change) are also of much interest. The participation to certain of these alternations allows one to predict the type of motion and the nature of the end state. Verbs which focus only either on the motion (e.g. pour) or on the resulting state (e.g. fill) do not alternate. Verbs that alternate constrain in some manner both motion and end state. Let us now specify in more depth these constraints, since in fact quite few verbs do alternate.

For example, let us consider the *into/with* alternation. [Pin89] differentiates among verbs which more naturally accept the *into* form as their basic form and which alternate with a *with* form. Their general form is:

Verb NP(+theme) onto NP(+destination), and they alternate in:
Verb NP(+destination) with NP(+theme).

These verbs naturally take the theme as object (e.g. pile). Other verbs more naturally take the location/container as object (e.g. stuff), their basic form is more naturally:

Verb NP(location) with NP(+theme), and alternate in:
Verb NP(+theme) onto NP(+destination).

For these two types of constructions, only a very few verbs require the obligatory presence of the two objects.

If we now consider the first set of verbs, those whose basic form is more naturally the 'into/onto' form, then verbs which have one of the following properties alternate: simultaneous forceful contact and motion of a mass against a surface (brush, spread, ...), vertical arrangement on a horizontal surface (heap, pile, stack), force is imparted to a mass, causing ballistic motion along a certain trajectory (inject, spray, spatter), etc. Those which do not alternate have for example one of the following properties: a mass is enabled to move via gravity (spill, drip, spill), flexible object extended in one direction put around another object (coil, spin, twist, wind), mass is expelled from inside an entity (emit, expectorate, vomit). As can be seen here, the properties at stake are very precise and their identification is not totally trivial, especially for verbs which can be used in a variety of utterances, with some slight meaning variations.

These properties are derived from the observation of syntactic behaviors. While some properties seem to have a clear ontological status, others seem to be much more difficult to characterize. They seem to be a conglomeration of some form of more basic properties.

Semantics of the verb and semantics of the construction Let us now consider the combination of a verb, with its own semantics, with a syntactic construction. The *Construction Grammar* approach [Gol94] sheds a particularly clear and insightful light on this interaction; let us present here some of its aspects, relevant to the verb semantic class system. The first point concerns the nature of the verb semantics, the nature of the semantics of a construction and the characterization of the interactions between these two elements. The second point concerns the meaning relations between constructions. These elements are of much importance for lexicalization and the construction of propositions (see §2.5.1).

Verbs usually have a central use, characterized by a specific syntactic form, but they may also be used with a large variety of other syntactic forms. In this case, the meaning of the proposition may be quite remote from the initial meaning of the verb. Let us consider a few illustrative cases. In the sentence *Edith baked Mary a cake.* the initial sense of *bake* becomes somewhat marginal, in favor of a more global meaning, e.g. 'Edith INTENDS to CAUSE Mary TO HAVE cake'. There is not here a special sense of bake which is used, but *bake* describes a kind of 'manner' of giving Mary a cake.

Consider now the case of slide, suggested by B. Levin. From the two following sentences, one may conclude that there are two senses for slide (probably very close). The first sense would constrain the goal to be animate while the second would have no constraint.

(20) a. *Edith slid Susan/*the door the present.*

b. *Edith slid the present to Susan/to the door.*

Now, if we impose, in the ditransitive construction, that the goal must be animate, then we can postulate just one sense for *slide*, which is intuitively more conceptually appropriate. We then need to posit constraints in the alternations on the nature of the arguments which would then allow only those verbs which meet the constraints to undergo that alternation. As noticed very early by Lakoff, a verb alone (and its associated lexical semantics) cannot be used to determine whether a construction is acceptable, it is necessary to take into account the semantics of the arguments.

Depending on the construction and on the verb, the verb may either play an important part in the elaboration of the semantics of the proposition or may simply express the means, the manner, the circumstances or the result of the action, while the construction describes the 'central' meaning. In fact, the meanings of verbs and of constructions often interact in very subtle ways. One might conclude then that there is no longer a clear separation between lexical rules and syntactic rules.

The difficulty is then to identify and describe the syntactically relevant aspects of verb meaning, i.e. those aspects which are relevant for determining the syntactic expression of arguments, via linking rules. Pinker notes that these aspects should exist in small number, since they resemble characteristics of closed-classes. This is not very surprising, since syntactic alternations form a set of closed elements.

Classification of verbs w.r.t. semantic properties relevant for describing thematic relations

Having dealt with alternations, let's turn to thematic relations and their role in classification of verbs. Thematic relations express generalizations on the types of lexical functions that are established between the verb and its arguments in the predication. There is a consensus among researchers that assignment of thematic roles to the arguments of the predicate imposes a classification on the verbs of the language. Since the type of thematic roles and their number are determined by the meaning of the verb, the lexical decomposition of verb meanings seems to be a prerequisite for semantic classification of verbs. The close affinity between the compositional and relational lexical meanings plays a central role in the classifications of verbs outlined in this subsection (see §2.4.1, 2.5.2 2.6.2).

Verb classifications that are surveyed below were developed within the frameworks of Case Grammar and Role and Reference Grammar (RRG). Works of Chafe [Cha70], Cook [Coo79] and Longacre [Lon76] address the issues of verb classification with regard to thematic roles within the framework of the Case Grammar model. RRG, a structural-functionalist theory of grammar, is presented in works of Foley and Van Valin [Fol84] and Van Valin [Van93]. Characteristic for RRG is that it accounts for a detailed treatment of lexical representation that proves to be instrumental in describing the thematic relations in typologically different languages. It also incorporates the insights of Dowty's and Jackendoff's theories. There is, however, an important difference in the treatment of thematic relations within those two frameworks. In Case Grammar, they have a double function, namely, they serve as a partial semantic representation of the lexical meaning and also as an input to the syntactic operations such as for example subjectivization, objectivization and raising. In the latter, the RRG model, thematic relations have only the second function.

This difference highlights the problem of selection of semantic content in NLP lexicons. The following arguments might be posed in favour of the variant which includes the information on a partial semantic representation in the lexicon: (i) significant generalizations about lexical meaning of verbs are accounted for in an explicit way in the verb entry; (ii) such information can support disambiguation of senses in case of polysemy; (iii) as a consequence of (ii), the semantic correctness of verb classifications increases, which in turn can improve the results of syntactic and semantic rules that operate on verb classes; (iv) it can also contribute to the integration of semantic and syntactic content in the lexicon.

There is no doubt that the model of semantic roles from the seventies, and in particular its repertory of roles and definitions, has to be substituted with a more stringent semantic model to suit the needs of NLP. The combination of the Dowty [Dow89] model of protoroles with the model of thematic sorts proposed by Poznansky and Sanfilippo [San92a] and elaborated in Sanfilippo [San93b] seems to be a very interesting proposal or solution (see §2.6.2 for description of these models).

Let's finish these general remarks with a quotation from [Lon76], which captures the essentials of verb classification w.r.t. semantic roles. "An understanding of the function of cases or roles is insightful for the understanding of language. Even more insightful, however, is the grouping of these roles with verb types with which they characteristically occur. To do this we must specify features, which distinguish one set of the verbs from another set of verbs, then we must specify the roles that occur with verbs characterised by these features. The result will be sets of verbs with characteristic constellations of accompanying substantives in given role...Such a set of verbs with characteristic accompanying nouns in particular roles is called a *case frame*. To assemble and compare the case frames of a language is to evolve a semantic typology or classification of its verbs... As soon as one begins to assemble a number of case frames, similarities in sets of case frames begin to be evident. This leads to the feeling that case frames should constitute some sort of system, i.e. that they are not merely list or inventory, but a system with intersecting parameters."

Chafe's basic verb types Characteristic for Chafe's approach is the position that a sentence is build around a predicative element, usually the verb. The nature of the verb determines what nouns will accompany it, what the relation of these nouns to it will be and how these nouns will be semantically specified. For example if the verb is specified as an action, as in *The men laughed*, such a verb dictates that the noun to be related is *agent* which might be further specified as animate.

Chafe distinguished four basic verb types: states, processes, actions and action processes. State verbs describe the state or condition of a single argument (*The elephant is dead*) and they associate with Patient. Non-state verbs are subdivided into three subclasses: processes, action and action-processes. Processes express a change of condition or state in its argument (*The elephant died*). They cooccur with Patients. Actions describe something that verb argument does or performs (*Harriet sang*), hence Agent is associated with action verbs. Action-processes account for both actions and processes. They have two arguments, the performer of the action, Agent, and the thing undergoing the process, Patient (*The tiger killed the elephant*). Chafe offers a number of semantico-syntactic tests that are indicative in distinguishing the basic verb types. The following relations are discussed in Chafe's model: Agent, Patient, Experiencer, Beneficiary, Complement, Locative and Instrument. He also draws attention to the "derivational" relations between these four basic verb types, which enable these to be established by postulating features like inchoative, causative, decausative, resultative. Thus, for example, the feature "inchoative" when added to a state gives a process. These derivational features, which often can be manifested morphologically, reflect the compositionality of verb meaning.

Cook's verb classification Cook's case grammar matrix is a system based on two parameters. The vertical parameter has four values: state verbs, process verbs, action verbs and action processes, taken from Chafe [Cha70]. The other parameter has also four values: either with no further nuclear role added (e.g. only agent (A) and/or patient (P)), or with experiencer (E), benefactive (B) or locative (L) added as further nuclear elements. The content in Cook's matrix is presented below. (After Cook [Coo79] pp. 63-65.)

- **A. The four basic verb types**

- **A.1 State verbs**
  - Case frame [__Os], where Os=stative Object
  - examples: *broken, dry, dead, tight*
- **A.2 Process verbs**
  - Case frame [__O]
  - examples: *die, dry* (iv.), *break*(iv.), *tighten*(iv.)
- **A.3 Action verbs**
  - Case frame [__A]
  - examples: *dance, laugh, play, sing*
- **A.4 Action-process verbs**
  - Case frame [__A, O]
  - examples: *break* (tv.), *kill, dry* (tv.), *tighten* (tv.)

- **B. The four experiential verb types**
  - **B.1 State experiential verbs**
    - Case frame [__E, Os], where Os=stative Object
    - examples: *doubt, know, like, want*
  - **B.2 Process experiential verbs**
    - Case frame [__E, O]/ + psych movement
    - examples: *amuse, annoy, frighten, please* (when used with inanimate subjects)
  - **B.3 Action experiential verbs**
    - Case frame [__A, E]/ (derived frame)
    - examples: *amuse, annoy, frighten, please* (with A=O) (when used with animate subjects)
    - examples: *answer, congratulate, praise, question* (with lexicalised O or with O which is coreferential with A)
  - **B.4 Action-process experiential verbs**
    - Case frame [__A, E, O]
    - examples: *ask, say, speak, tell*

- **C. The four benefactive verb types**
  - **C.1 State benefactive verbs**
    - Case frame [__B, Os], where Os=stative Object
    - examples; *have, have got, own, belong to*
  - **C.2 Process benefactive verbs**
    - Case frame [__B, O]
    - examples: *acquire, gain, lose, win*
  - **C.3 Action benefactive verbs**
    - Case frame [__A, B] (derived frame)

    * examples: *arm, bribe, help, supply* (with lexicalized O)
  - **C.4 Action-process benefactive verbs**
    * Case frame [___A, B, O](derived frame)
    * examples: *give, accept, buy, sell*

- **D. The four locative verb types**
  - **D.1 State locative verbs**
    * Case frame [___Os, L], where Os=stative Object
    * examples: *dwell, stay,(be) in, (be) at*
  - **D.2 Process locative verbs**
    * Case frame [___O, L] (derived frame)
    * examples: *come, go, move, shift* (with inanimate subjects)
  - **D.3 Action locative verbs**
    * Case frame [___A, L] (derived frame)
    * examples: *come, go* (with animate subjects, where A=O), *run, walk* (where A=O)
  - **D.4 Action-process locative verbs**
    * Case frame [___A, O, L] (derived frame)
    * examples: *bring, put, place, take*

Longacre's verb classification Longacre extended the number of nuclear cases to 10, which resulted in a considerable enlargement of the number of verb classes. His scheme of case frames reminds one of that of the periodic chart of the chemical elements. The horizontal parameter accounts for the four basic verb types. The vertical parameter covers verb classes specified below and marked with letters (a) to (h). The following thematic roles were posed as nuclear by Longacre: Experiencer (E), Patient (P), Agent (A), Range (R), Measure (M), Instrument (I), Locative (L), Source (S), Goal (G), Path (Pa), Time. Manner and Accompaniment were considered peripheral roles. The verb classes specified in Longacre's scheme include following (for the complete exemplification (see [Lon76] pp. 44-9):

- (a) ambient verbs: *It's hot. It's cooling off.*
- (b) ambient-experiential verbs: *It hailed on me. The patient is hot. John got cold.* (E)
- (c) emotive, psych and impingement verbs *John is discouraged (about his work).* E (I), *John hit Bill (with his fist).* A E (I)
- (c') factual knowledge verbs *know, learn, teach, study.* They are illustrated below for the four basic verb types: state *Susan knows algebra.* (E R); process *Susan has learned a lot of algebra.* (E R); action *Susan is studying algebra.* (A/E R)
- (d) verbs expressing desire or cognition: state *Mary wants a Cadillac.* (E G); process: *Mary fell in love with Tom.* (E G); action-process *I introduced Mary to Tom.* (A E G); action *John scorned Mary.* (A E/G or A G)

- (d') verbs of sensation, speech and attention like f.ex. *see, tell, listen*

- (e) physical verbs (These verbs correspond roughly to Chafe's and Cook's basic verb types.)

- (f) measure verbs: state *The statue weights one ton.* (P M); process *My bonds went down 10%.*(P M); action-process *I shortened it two inches.* (A P M); action *The cowboy gained five yards.* (A M)

- (g) locative verbs which are static in nature and combine with Locative *The knife is in the box.* (P L),*They placed the book by the phone.*(A P L)

- (g') motion, propulsion and locomotion verbs which occur with cases like Source, Path, Goal*Don fell from the chair.*

- (h) verbs of possession, acquisition, transfer and "grab" *have, obtain, give, grab*, all but the last of which involve the idea of property *Dick has a new book.* (P G), *T. gave B. a book.* (A/S P G).

- (h') is similar to (h) but adds a semantic component of transitoriness (by having the feature MOTION instead direction.) *T. gave B. a book for Sue.* (A/S P Path G).

In Longacre's frame scheme there are 45 filled cells with the total of 48 case frames. It is characterised by an overall regularity with some spots of local irregularity. Longacre observes that rows (a-d') may have Experiencer but not Patient while rows (e-h') can have Patient but not Experiencer. This plus the distribution of Agent in the columns describing action-processes and actions show some major verb classes in the case frame approach.

Role and Reference Grammar (RRG) The Role and Reference Grammar (RRG) has some characteristics in common with the models discussed above. The theory of verb classes occupies a central position in the system of lexical representation in RRG. The verb is also assumed to be a determining element in the nucleus predication. RRG starts with the Vendler [Ven68] classification of verbs into states (e.g. *have, know, believe*), achievements (e.g. *die, realise, learn*), accomplishments (e.g. *give, teach, kill*) and activities (e.g. *swim, walk, talk*). It utilises a modified version of the representational scheme proposed in Dowty [Dow89] to capture the distinctions between these verb classes.

Dowty explains the differences between the verb classes in terms of lexical decomposition system in which stative predicates (e.g. *know, be, have*) are taken as basic and other classes are derived from them. Thus achievements which are inchoative semantically are treated as states plus a BECOME operator, e.g. BECOME **know'** "learn". Accomplishments which are inherently causative are represented by the operator CAUSE linked to the achievements operator BECOME, e.g. CAUSE [BECOME **know'**] "teach". Activities are marked by the operator DO for agentive verbs. These decompositional forms are termed Logical Structures (LS) by Dowty. In RRG they are interpreted in the following way:

Verb Class	Logical Structure
STATE	**predicate'**(x) or (x,y)
ACHIEVEMENT	BECOME **predicate'** (x) or (x,y)
ACTIVITY (+/- Agentive)	(DO (x) [**predicate'** (x) or (x,y)])
ACCOMPLISHMENT	ϕ CAUSE ψ, where ϕ is normally an activity predicate and ψ an achievement predicate.

(Adapted from [Van93], p. 36)

These LSs are starting point for the interpretation of the thematic relations in RRG. Thematic relations are defined in terms of argument positions in the decomposed LS representations, following Jackendoff [Jac90]. Their repository and distribution among the verb classes is presented in the table below:

I. STATE VERBS		
A. Locational	**be-at'** (x,y)	x=locative, y=theme
B. Non-Locational		
1. State or condition	**broken'**(x)	x=patient
2. Perception	**see'**(x,y)	x=experiencer, y=theme
3. Cognition	**believe'**(x,y)	x=experiencer, y=theme
4. Possession	**have'**(x,y)	x=locative, y=theme
5. Equational	**be'**(x,y)	x=locative, y=theme
II ACTIVITY VERBS		
A. Uncontrolled		
1. Non-motion	**cry'**(x,y)	x=effector, y=locative
2. Motion	**roll'**(x)	x=theme
B. Controlled	DO (x, [**cry'**(x)])	x=agent

(Adapted from [Van93]. p.39)

Patient is associated with the single argument of a single-argument stative verb of state or condition, as in *The watch is broken.* Theme is the second argument of two place stative verbs, e.g. *the magazine* in *The magazine is on the desk. the desk* is a locative, the first argument of two-place locational stative verbs. Experiencer is the first argument with a two place stative perception verbs. The single argument of a motion activity verb is a theme, as it undergoes a change of location: *The ball rolled.* The first argument of a non-motion activity verb is an effector, the participant which does some action and which is unmarked for volition and control as in *The door squeaks.* Such interpretation of thematic roles leads to the conclusion that the thematic roles in RRG are independently motivated.

LS and thematic roles are part of the semantic representation in RRG. Thematic roles function as one of the links between the semantic and syntactic representation. Semantic macroroles, Actor and Undergoer are the other link. Macroroles conceptually parallel the grammatical notions of arguments in a transitive predication. Being an immediate link to the level of Syntactic Functions, they control the assignment of syntactic markers to the arguments of the verb. It should be noted that the rich delineation of the lexical representations in the RRG model is well suited for the description of typologically different languages.

The classes of verbs in the table above cover different cognitive dimensions of language. The main cognitive distinction is drawn between two conceptual categories such as State and

Activity. State verbs are subclassified into two major classes comprising locational and non-locational verbs. Among the non-locational verbs, the following subclasses are distinguished: state or condition, perception, cognition, possession and equational verbs. Activity verbs are subdivided with respect to the control component. Uncontrolled verbs are further subclassified with respect to the motion component.

Comparing approaches in Case Grammar oriented models and RRG A common characteristic for the approaches to the classifications of verbs sketched in this subsection is that they search for a subset of recurrent semantic components and semantic roles that are relevant for the description of thematic relations. The two approaches reveal some interesting parallels concerning the decompositional analysis of verb meanings with regard to the subclassification of verbs into more or less equivalent types; thus states = states, activity = action, achievement = process, accomplishment = action-process. Since the LSs in the RRG model correspond to the thematic relations that other theories associate with a verb in their lexical entry, there is some partial similarity that the classifications of verbs within the two frameworks share. These two frameworks also show some overlap as far as the semantic affinity of the major subclasses of verbs is concerned.

The frameworks differ with regard to (a) the choice of description model, e.g. hierarchy vs. matrix model, (b) the level of semantic granularity in the subclassification of verbs, (c) the function that thematic relations play in the semantic representation in the respective frameworks.

The issues addressed within these frameworks turn attention, in the first place, to some basic linguistic questions that have to be answered when approaching the description and formalization of the lexical meaning in lexicons designed for both general and NLP purposes. The classification of verbs w.r.t. thematic relations should be seen as a preparatory stage that aims at a partial semantic representation of the lexical meaning of verbs. It has to be well adjusted to the chosen model of the semantic representation, which in turn has to be integrated with the model of syntactic representation.

LCS-Based Verb Classification

Let us now introduce the Lexical Conceptual Structure (LCS), which is an elaborated form of semantic representation, with a strong cognitive dimension. The LCS came in part from the Lexical Semantics Templates (see above) and from a large number of observations such as those of [Gru67]. The present form of the LCS, under which it gained its popularity, is due to Jackendoff [Jac83], [Jac90]. The LCS was designed within a linguistic and cognitive perspective, it has some similarities, but also major differences, with approaches closer to Artificial Intelligence such as semantic nets or conceptual graphs. The LCS is basically designed to represent the meaning of predicative elements and the semantics of propositions, it is therefore substantially different from frames and scripts, which describe situations in the world like going to a restaurant or been cured from a desease. It is not specifically oriented toward communication acts or toward the representation of abstract objects of the world (by means of e.g. state of affairs, complex indeterminates), represented as objects, as in Situation Semantics (e.g. a complex indeterminate to model a person who utters an sentence).

Main Principles and characteristics The LCS is mainly organized around the notion of *motion*, other semantic/cognitive fields being derived from motion by analogy (e.g. change of possession, change of property). This analogy is fine in a number of cases, as shall be seen below, but turns out to be unnatural in a number of others. From that point of view, the LCS should be considered both as a semantic model providing a representational framework and a language of primitives on the one hand, and as a methodology on the other hand, allowing for the introduction of new primitives to the language, whenever justified. Another important characteristics of the LCS is the close relations it has with syntax, allowing the implementation of a comprehensive system of semantic composition rules. From that point of view one often compares the LCS with a kind of X-bar semantics.

Let us now introduce the different elements of the LCS language. They are mainly: conceptual categories, semantic fields and primitives. Other elements are conceptual variables, semantic features (similar to selectional restrictions, e.g. such as eatable entity, liquid), constants (representing non-decomposable concepts like e.g. money, butter) and lexical functions (which play minor roles). [Pin89] also introduces relations between events (e.g. means, effect). These latter elements are not presented formally, but by means of examples.

Conceptual Categories [Jac83] introduces the notion of conceptual constituent defined from a small set of ontological categories (also called conceptual parts of speech), among which the most important are: *thing, event, state, place, path, property, purpose, manner, amount, time.* These categories may subsume more specific ones, e.g. the category *thing* subsumes: *human, animal, object.* These categories may be viewed as the roots of a selectional restriction system.

The assignment of a conceptual category to a lexical item often depends on its context of utterance, for example the noun *meeting* is assigned the category *time* in (21) while it is assigned the category *event* in (22).

(21) after the meeting

(22) the meeting will be held at noon in room B34

There are constraints on the types of conceptual categories which can be assigned to a lexical item. For example, a color will never be assigned categories such as event or distance.

Conceptual categories are represented as an indice to a bracketed structure as shown in (23) where the contents of that structure has the type denoted by the semantic category.

(23) $[_{<semantic\ category>}\ ]$

Here are a few syntactic realizations of conceptual categories:

(24) [$_{thing}$ Mozart] is [$_{property}$ famous].

(25) He composed [$_{amount}$ many [$_{thing}$ symphonies]].

(26) [$_{event}$ The meeting] starts at [$_{time}$ 2 PM].

(27) Ann switched off the electricity [$_{purpose}$ to prevent a fire].

(28) The light is [$_{state}$ red].

(29) [$_{event}$ Edith begins a new program].

Conceptual primitives The LCS is based on a small number of conceptual primitives. The main ones are BE, which represents a state, and GO, which represents any event. Other primitives include: STAY (a BE with an idea of duration), CAUSE (for expressing causality), INCH (for inchoative interpretations of events), EXT (spatial extension along something), REACT, EXCH (exchange), ORIENT (orientation of an object), etc. Their number remains small, while covering a quite large number of concepts. A second set of primitives, slightly larger (about 50) describes prepositions: AT, IN, ON, TOWARD, FROM, TO, BEHIND, UNDER, VIA, etc. These primitives are 'lower' in the primitive hierarchy and their number is *a priori* fixed once for all.

The LCS uses some principles put forward in [Gru67], namely that the primitives used to represent concepts of localization and movement can be transposed to other fields by analogy, and generalized. The main fields considered in the LCS are the following: localization (+loc), time (+temp), possession (+poss) and expression of characteristics of an entity, its properties (+char,+ident) or its material composition (+char,+comp). [Pin89] introduces additional fields such as: epistemic (+epist) and psychological (+psy).

Primitives can then be specialized to a field, e.g. GO_{+loc} describes a change of location, GO_{+temp} a change of time, GO_{+poss} a change of possession, and $GO_{+char,+ident}$ a change in the value of a property (e.g. weight, color).

Verb classes based on LCS patterns The LCS allows us to classify verbs at different levels of granularity and according to different conceptual dimensions:

1. the state / event distinction
2. the primitives they are built from
   (a) primitive root of the representation: GO, BE, CAUSE,
   (b) the semantic fields marking that primitive,
   (c) the different arguments of the verb and their semantic fields.

It is clear that classifications must be based on generic LCS patterns. A verb belongs to the class associated with a pattern if and only if its representation is subsumed by that LCS pattern (LCS patterns can be viewed as types). This classification method, based on a conceptual representation has some similarities with classifications based on semantic roles, since it has been shown that LCS patterns have strong relations (and are even more precise) than semantic roles. It also allows us to introduce prepositions into verb meaning and to somewhat characterize in extension the lexical patterns associated with PPs.

Here are now a few examples, constructed for French verbs. As can be noted the classes formed from LCS patterns are substantially different in nature and form from those obtained from syntactic of thematic criteria:

Verbs of spatial motion:
LCS pattern: $[_{event}$ GO $_{+loc}$ $([_{thing}$ $], [_{path}$ $])$
examples:
{aller, venir, partir, sortir, entrer, arriver, amener, déplacer, se rendre, s'amener, marcher, commander, livrer, approacher, avancer, mettre, apparaitre, survenir, quitter, bouger, poser, dissiper, extraire, monter, descendre, pénétrer ...}

Verbs of transfer of possession:
LCS pattern: $[_{event}$ $GO_{+poss}([_{thing}$ $], [_{path}$ $])]$ (a CAUSE may be added) examples:
{donner, prendre, voler, adresser, acquérir, alimenter, apprendre, cambrioler, allouer, offrir,

prodiguer, retenir, consacrer, acheter, vendre, céder, fournir, livrer, échanger, troquer, abandonner, sacrifier, confier, procurer, apporter, remettre, porter, distribuer, rendre, octroyer, abandonner, dilapider, perdre, ...}

Verbs of temporal 'motion':
LCS pattern: $[_{event}\ \mathrm{GO}\ _{+temp}\ ([_{thing}\ ], [_{path}\ ])]$
example:
{retarder, déplacer, avancer, ajourner, remettre, reporter, attarder, différer, repousser, anticiper, ... }

Verbs of change of state:
LCS pattern: $[_{event}\ \mathrm{GO}_{+char,+ident}([_{thing}\ ], [_{path}\ ])]$ (a CAUSE may also be added to that pattern).
examples:
{devenir, changer, évoluer, régresser, se modifier, se transformer, progresser, varier, diversifier, casser, altérer, aliéner, détériorer, détruire, construire ...}

Verbs of persistance of a state:
LCS pattern: $[\ _{event}\ \mathrm{STAY}_{+char,+ident}\ ([_{thing}\ ], [_{place}\ ])]$
examples:
{maintenir, rester, persister, laisser, fixer, arrêter, immobiliser, entretenir, stabiliser, geler, figer, paralyser, pétrifier, ...}

Verbs of possession (state):
LCS pattern: $[_{state}\ \mathrm{BE}_{+poss}\ ([_{thing}\ ], [_{place}\ ])]$
examples:
{avoir, posséder, détenir, bénéficier, jouir, disposer, ...}

Verbs of spatial extension:
LCS pattern: $[_{state}\ \mathrm{EXT}_{+loc}([_{thing}\ ], [_{place}\ ])]$
examples:
{s'étendre, longer, côtoyer, raser, border, s'étaler, caboter, entourer, prolonger, ... }

Verbs of temporal extension:
LCS pattern: $[_{state}\ \mathrm{EXT}_{+temp}([_{thing}\ ], [_{place}\ ])]$
examples: { durer, immortaliser, éterniser, perpétuer, prolonger, ... }

2.6.3 Comparisons between approaches

It is quite difficult to compare the three above approaches. They are based on very different assumptions. We can however indicate that classes constructed on syntactic criteria are of much interest from a theoretical point of view, in the study of the cooperation between syntax and semantics. They are certainly less useful in LKB design since they are far from complete and include many forms of exceptions.

The approach based on general semantic criteria is much more concrete and applicable. However, the classes which can be formed on this basis remain very general. Classes formed using the ontological criteria of WordNet, from that point of view, are more fine-grained, and they should be preferred (see §3.6). The LCS-based classification is also fine-grained, its main advantage is to base the classification on semantic frames, which can then be used for semantic representation. Frames may be felt to be more arbitrary, but, in fact, they share a lot with the previous classification method. LCS representations are more formal, they allow thus more precise classifications, but they may also be felt to be incomplete (because of their formalism based entirely on predicates and functions) and to be difficult to establish.

2.6.4 Relations with other areas of lexical semantics

It is clear that verbs is one of the most central syntactic category in language. They have deep relations with the other categories: nouns because they select arguments which are often nominals, adverbs because adverbs modify verbs, prepositions, since they introduce PPs. Verbs assign thematic roles to their arguments and to prepositions, which, in turn assign thematic roles to NPs. Verbs associated with adverbs permit the computation of aspect.

2.6.5 Verb semantic classes in LKB

For the same reasons as above, verbs have a privilege position in LKBs. The following resources have developed a quite extensive description of verbs, described in the next chapter of this report: §3.8, 3.11, 3.12, 3.12.2.

2.6.6 Verb semantic classes in Applications

Information about verb semantic classes and semantic verb frames is central in many applications, in particular Machine Translation (§4.1). It is now starting to acquire relevance for information systems since it provides the opportunity for enforcing a indexing procedure based on conceptual structures rather than token strings as in traditional m keyword-based methods (see §4.3, 4.2).

2.7 Nouns

2.7.1 Introduction

In this section we address the basic requirements for encoding the lexical properties of nominals and nominalizations. The approaches that we have surveyed present important differences in the aims that are pursued and hence on the type of linguistic data that are analysed. Thus, the representational choices that are made concerning the semantics of nominal forms depend on the task at hand as well as on the way in which theoretical questions are framed. Our survey of contemporary work on the semantics of nominals and nominalization indicates that Generative Lexicon theory is the only framework that has argued extensively for a rich representation for nominals. A reason for this is that the semantics of nominal forms can be exploited to explain fairly naturally a number of compositional processes involving verbal aspect, the interpretation of compound expression, meaning shifts, etc.

2.7.2 Description of different approaches

The studies on the linguistic properties of nouns and nominalizations have been addressed in a number of frameworks and approaches that can be broadly distinguished in terms of four major models:

(1) Taxonomic/phychological Models

(2) Decompositional Models

(3) Formal Models

(4) Generative Models

Taxonomic models reflect the contributions of both research in artificial intelligence and psychology [Qui94], [Mil76]. These models have been very influential for work on computational lexicons, but they cannot be properly characterized as lexical semantics since the relation between underlying representation and syntactic behavior is not addressed. The focus here is on defining a set of relations (*is-a*, *has*, etc.) and a set of lexical primitives (expressed as features) that connect words or concepts in a network or that drive inferences. The approaches that would appear under this heading include conventional inheritance models [Bra79], [Bra85b], [Bra85a], descriptive approaches to the study of semantic universals [Wie72], [Wie89b], [Dix91] and finally models aimed at describing interlexical relations, such as Wordnet, [Mil85a], [Mil90b], [Cru86].
Decompositional models, unlike *taxonomic models*, pay special attention to syntactic behavior as the basis for classifying words. Under this heading we include work ¡influenced by contemporary linguistics [Dow79], [Jac90], [Gri90]. In these models, the properties of nominals have been analyzed insofar as they bear directly on issues of argument structure and thematic roles. Thus, the main interest is with nominalizations, while non-derived nominals are taken to be atomic entities with no internal constitution. The distinction between semantic entities does not play a role and thus there is no explicit discussion of issues of inheritance.
Formal Models describe the semantics of terms as referring constructions and the properties of nouns as heads of these terms [Bun81], [Lin83], [Lin88], [Sch81], [Lan89], [Lan87], [Pel86], [Kri87]. Nouns are differentiated for their inherent nominal aspect (set, mass, collective, group) which determines the way terms refer to entities: as discrete or non-discrete entities. Discreteness, in turn, determines the way in which terms can be quantified, how agreement takes places in certain languages, and how predications are distributed to the entities (collective, reciprocal, distributive). *Formal Models* do not describe the conceptual and denotational properies of nominals, and are fully complementary to the *Taxonomic Models*.
Generative Models merge the above views, in that they assume a richer structure for nominals. These involve the rich notion of semantic information from primitive-based inferencing techniques in AI [Wil75], as well as from abductive models such as [Hob93]. However, these models adopt the methodology of decompositional models for determining the information in the lexical representation, which ought to be linguistically motivated. Under this heading we include Generative Lexicon theory [Pus95a], [Bus96a], [Bou97], the model of lexical description adopted in Aquilex [Cop91], [Cop92], [Vos95a], as well as the most recent views outlined in [Jac96].

2.7.3 Comparison intended to identify basic notions

The theoretical vocabulary which is adopted in each of the approaches has important repercussion on representational issues and thus on the descriptive and explanatory adequacy of the model. The main differences can be summarized as follows:

(1) Is there internal structure to nominal forms?

(2) What are the primitives in the system?

(3) What are the relations between different lexical items?

The semantic characterization of lexical items in both *taxonomic* and *decompositional* models can be braodly viewed in terms of lexical primitives, although there are some distinctions to be made. In the case of Wordnet, the notion of lexical primitive is not appropriate since lexical semantics is expressed only by the relations that words bear to each other. If there are primitives it is in definition of the relations. For *decompositional* models this is only true for verbs. As said above, nominals are atomic units that carry only part of speech information, with the exception of nominalizations that may contain variables indicating argument and event positions. The primitives within generative and formal models are theoretical primitives which reflect the underlying structure of lexical knowledge (qualia) combined with typing information from an ontology of basic types. The latter reflects, in part, the interlexical relations of taxonomic models. The table below reflects the general views concerning questions (1) and (2) above:

	nouns		**nominalizations**	
	internal-structure	interlexical-rels	internal-structure	interlexical-rels
taxonomic	X	X		
decompositional			X	
formal	X			
generative	X	X	X	X

The table above shows that taxonomic approaches have not dealt with nominalizations, and this is a natural result of the fact that syntax does not play a crucial role. The opposite situation is true with decompositional models where morphologically derived nouns have received special attention since the relation between a verb and the derived noun is transparent. Formal models only account for the syntactic properties of nouns as far as they reflect compositional properties of nouns as heads of referring terms.

Representations

In this section we briefly describe and make explicit the representations in different frameworks. The main distinction will be drawn between a generative model and a model that combines part of speech information with semantic features that place a word in a semantic network. This latter model reflects the structure of what [Pus95a] has termed Sense Enumeration Lexicons (SELs). Consider the representation of the nominals *knife* and *window* within an SEL:

(30) $\begin{bmatrix} \textbf{knife} \\ \textsc{cat} = \textbf{count-noun} \\ \textsc{genus} = \textbf{phys\_object} \end{bmatrix}$

(31) $\begin{bmatrix} \textbf{window}_1 \\ \textsc{cat} = \textbf{count-noun} \\ \textsc{genus} = \textbf{phys\_object} \end{bmatrix}$

(32) $\begin{bmatrix} \textbf{window}_2 \\ \textsc{cat} = \textbf{count-noun} \\ \textsc{genus} = \textbf{aperture} \end{bmatrix}$

The nominal *window* requires two entries in order to account for its polysemy as reflected in the following alternation: *The boy fell through the window* (aperture); *the boy broke the window* (physical object). In addition to categorial information, the Generative Lexicon model relies on qualia structure and the structuring of information therein. A simplified description of a nominal form, α, involves information concerning argument structure, event structure and qualia structure:

(33)
$$\begin{bmatrix} \alpha \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x \\ \ldots \end{bmatrix} \\ \text{EVENSTR} = \begin{bmatrix} \text{D-E1} = e_1 \\ \text{D-E2} = e_2 \\ \ldots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \textbf{what } x \textbf{ is} \\ \text{CONST} = \textbf{what } x \textbf{ is made of} \\ \text{TELIC} = e_2 \textbf{ as the characterizing event of } x \\ \text{AGENTIVE} = e_1 \textbf{ as the coming into being of } x \end{bmatrix} \end{bmatrix}$$

The nominals *knife* and *window* are shown below:

(34)
$$\begin{bmatrix} \textbf{knife} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x : artifact\_tool \\ \ldots \end{bmatrix} \\ \text{EVENSTR} = \begin{bmatrix} \text{D-E1} = e_1 \\ \text{D-E2} = e_2 \\ \ldots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \textbf{artifact\_tool(x)} \\ \text{CONST} = \textbf{part\_of(x,blade),part\_of(x,handle)} \\ \text{TELIC} = \textbf{cut}(e_2\textbf{,w,y,x)} \\ \text{AGENTIVE} = \textbf{inherit(artifact)} \end{bmatrix} \end{bmatrix}$$

(35)
$$\begin{bmatrix} \textbf{window} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x : phys\_object \\ \text{ARG2} = y : aperture \\ \ldots \end{bmatrix} \\ \text{EVENSTR} = \begin{bmatrix} \text{D-E1} = e_1 \\ \text{D-E2} = e_2 \\ \ldots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORMAL} = \textbf{hold(y,x)} \\ \text{CONST} = \textbf{part\_of(x,y),made\_of(x,z:glass),..} \\ \text{TELIC} = \textbf{see\_through(e,person,x)} \\ \text{AGENTIVE} = \textbf{inherit(phys\_object)} \end{bmatrix} \end{bmatrix}$$

Notice that there is only one lexical representation for *window* which accounts for the polysemous behavior. This is achieved by assuming an underlying relational structure of the type, which reflect the logical relation between the two senses, whereby for *widow* the aperture holds the physical aspect of the object.

Advantages and Disadvantages

The advantage of the generative model is that it avoids multiple listing of lexical entries. There is a very serious theoretical effort involved in reducing the number of senses to an underspecified representation and stating the compositional operations. However, if the aim is that of achieving explanatory adequacy, then this appears to be a necessary solution given the potentially large number of sense extensions for a particular lexical item. Consider, for instance, the nominal *violinist*:

(36) a. John is a violinist for the Boston Symphony orchestra. (*occupation*)

b. The first violinist is sick. (*role*)

c. Mary is a good violinist. (*ability*)

d. The violinist on the stage broke a string. (*stage- level*)

e. The violinist of the orchestra are on strike. (*member*)

With respect to nominals such as violinist, there is nothing guaranteeing that the number of senses that it may acquire in context are limited to the ones mentioned above, an enumeration approach will certainly fail in light of new examples.

2.7.4 Relation to other areas of lexical semantics

The semantics of nominals bears directly on a number of areas in semantics and lexical semantics:

- the composition with adjectives;
- the interpretation of nominal compounds;
- their role in lexicalization processes;
- their relation with aspect;
- their interaction with quantification.

Adjectives

Adjectives are sensitive to the internal structure of the head noun that they modify. Assuming internal structure for nouns appears to yield nice generalizations for the polysemous behavior of adjectives.

(37) a. a fast car
(e.g., a car that can be driven fast)

b. a fast highway
(e.g., a highway where one can drive fast)

c. a fast dancer
(e.g., a dancer who dances fast)

d. a fast waltz
(e.g., a waltz that is danced fast)

In addition, certain meaning components appear to be involved in licensing adjectives with different nominals:

(38) a. A natural/talented violinist.

b. *A natural/talented smoker.

(39) a. *A frequent violinist.

b. A frequent smoker.

Nouns and Ontologies

The theoretical issues discussed here are to a large extent complementary to the work carried out by the ANSI Ad Hoc Ontology Standards Group (ANSI X3T2), which operates in the US. The ANSI Committee meets since March 1996 and has brought together researchers and developers of ontologies from a variety of disciplines and institutes. The final aim is the development of a standardized Reference Ontology of general and fundamental 10,000 concepts, based on the major concepts of a variety of existing resources, such as CYC (§3.9.2), Pangloss-Sensus (§3.9.5), Pennman 3.9.4, MikroKosmos (§3.9.3), EDR (§3.8), and WordNet1.5 (§3.6.2). A description of the work done is given in [HovFC]. The main focus of their effort is not the lexicon but ontologies as a separate object. Nevertheless, many notions and issues discussed here are relevant to both lexical and non-lexical semantic resources.

Compounds

Complex nominals, and in particular noun-noun compounds, have received a great deal of attention in linguistic research, ([Berg91], [Jes42], [Mar69], [Lee70], [Dow77], [Levi78], [War87]) and in computational linguistics, where their analysis has presented a serious challenge for natural language processing systems ([Fin80], [McD82], [Isa84], Als87, [Hob93], [Boui92], [Job95], [Bus96b],[Cop97]). The reason for such interest is that the interpretation of Noun-Noun compounds poses serious problems. The strategy for forming English Noun-Noun compound involves combining a modifier noun and a head noun, without making explicit the semantic relation between compound members. The main difficulty liese in recovering the implicit semantic information holding by the nouns, which varies greatly:

(40) a. coffee cup
b. coffee producer
c. coffee grinder
d. coffee creamer
e. coffee machine

When looking at cross-linguistic data complex nominals require systematic solutions. In Italian, for example, complex nominals are generally comprised of a head followed by a preposition and a modifying noun. Consider the correspondences below in (41).

(41)	a.	bread knife	coltello <u>da</u> pane
	b.	wine glass	bicchiere <u>da</u>vino
	c.	lemon juice	succo <u>di</u> limone
	d.	glass door	porta<u>a</u> vetri

The analysis in ([Levi78], [Job95]), relies on the enumeration of possible semantic relations observed in the language. A number of proposal exploiting generative models are basing the composition on the qualia structure, which provides the 'glue' which links together the semantic contributions of modifying nouns and the head noun in the compound ([Joh95], [Bus96b], [Joh98], [Fab96]).

Lexicalization

Lexicalization is concerned with the linguistic components that are responsible for the constructions allowed by a word itself. Consider an example of communicating that there was a shooting event. This can be done in many ways as shown below:

(42) a. The man shot the president twice.
b. The killer shot the president twice.
c. The man shot the president dead.
d. ??The killer shot the president dead.

The sentences above communicate different outcomes of the event of shooting the president. Whereas the sentence in (42b) strongly conveys the information that the president died, the one in (42a) does not. What is clearly responsible for the different entailments is the lexical content differentiating *man* from *killer*.

Events

The role of events in the representation of nominals constitutes a rather controversial topic. In general, the only nouns that are taken to contain events in their representation are those that express an event directly. For instance, nouns such as *war*, *destruction*, and so on. Within generative models, and more specifically Generative Lexicon theory, all lexical items carry event based information which is crucially involved in the compositional processes involving: type coercion, adjectival modification, event nominalizations and, among others, also agentive nominals.

(43) a. TYPE COERCION John enjoyed the book.
"John enjoyed <u>some event</u> that can be reconstructed from the lexical item 'book'."
Minimally: *read*, *write*.
b. John began the movie.
"John began <u>some event</u> that can be reconstructed from the lexical item 'movie'."
Minimally: *show*, *produce*.

(44) a. ADJECTIVAL MODIFICATION an occasional cup of coffee/cigarette.
"a cup of coffee that someone <u>drinks</u> occasionally."
b. a quick beer.
"a beer to be <u>drunk</u> quickly."
c. a long record.
"a record that <u>plays</u> for a long time."

(45) a. EVENT NOMINALIZATIONS The <u>building</u> of this house took many years. (*process*)
b. The <u>building</u> next to ours is really ugly. (*result*)

(46) a. AGENTIVE NOMINALS The <u>customer</u> left without paying.
b. The <u>rebels</u> are still holding key positions.

Finally, as mentioned in the introduction, certain nominals affect the aspectual interpretation of the verb:

(47) a. John painted a beautiful portrait. (*create*)

b. John painted the outside wall. (*change of state*)

Quantification

Quantification within the semantics of nominals is a very broad and difficult topic. Within Formal Models the type of entity referred to determines a range of quantificational phenomena:

1. quantification

(48) a. a policeman/ two/some policemen (*singleton and multiform set of objects*)

b. some/two police (*multiform set of objects*)

c. a squadron/ two/some squadrons (*singleton and multiform set of groups*)

d. some/much traffic (*non-discrete collection of objects*)

e. some/ much water (*non-discrete amount of substance*)

f. three medicines(*the number of medicine-types is counted*)

Ordinary count nouns can take discrete quantifiers to form multiform sets, see (48a). In the case of group nouns, (48c), groups are counted and not the members of the group. In the case of collectives, e.g. (48d), the number of elements is measured and not counted. Genuine *substances* as in (48e) represent measurements both as referential entities and as conceptual units. Finally, we see a case of quantification of entity types in (48f). The diversity is quantified, nothing is said about the amount of medicine.

2. predication distribution

(49) a. the police/squadron/policemen carried new equipment (*distributive*)

b. The policemen carried the piano upstairs (*collective*)

c. The family loves each other (*reciprocal*)

d. The committee meets (*multiplicity of subject*)

e. two big healthy families/boys (*different property distributions*)

In (49a) we see that the multiplicity (set, group or collective) of the subject and the object explains that we can have distributive, collective and cumulative interpretations: each persons carries a single set of equipment, they all carry one set of equipment, any combination is possible. In (49b) the fact that *a piano* is a single ad big individuated object suggests that the subject carries collectively. In the next two examples we see that the multiplicity of the subject is a necessary selection criterion for the reciprocal interpretation of the predicate. The final example shows that group nouns exhibit two different entity levels for quantification and property distribution.

3. agreement

(50) a. The band played/plays well

b. The committee met/meets

The conditions determining agreement vary a lot across languages. In the English examples, we see that quantificational semantic properties can play a role. Here we see that the singular subjects can govern a plural verb when interpreted as a multiform group noun. See also [Cop95] for a discussion of group nouns.

Semantic Roles

Nominal of all kinds carry argument structure:

(51) a. the baker of the cake;

b. the owner of the red car;

(52) a. the book of poetry;

b. the door to the kitchen;

(53) a. the war between England and France;

b. the meeting between the diplomats;

2.7.5 How is information encoded in lexical databases

There is a fairly large number of treatments which will be surveyed by the next meeting. These include 3.9, *Ontos* (Nirenburg), *Penman* (§3.9.4).

Acquilex

Nouns in Acquilex 3.12.3 are represented using Typed Feature Structures in an HPSG- tradition, encoding the following types of information:

CAT: morpho-syntactic properties such as number, gender, person, countability

SEM: formal semantic properties such as set; mass

QUALIA: conceptual semantic properties

The QUALIA specification roughly corresponds to the Generative lexicon approach. The SEM specification incorporates formal semantic notions on entity types. These are correlated with the QUALIA properties. These correlations predict collective and distributive property distributions for the different multiform nouns (plural, group, collective) as discussed above.

2.7.6 How is the information used in LE applications

In principle, basic semantic networks and taxonomies such as wordnet can directly be used in various applications. There is however a direct correlation between the amount of detail in semantic information that is encoded and the level of interpretation and information processing that can be achieved. There are a number of important reasons for devoting a substantial research effort into a lexical model that accounts for the syntactic and semantic properties of nominal types. Some of the reasons are as follows:

(A.) Certain type of nouns appear to impose restrictions on subsequent discourse;

(B.) nouns restrict the possible interpretations of ellipsed predicates, which bears directly on the determination of what a portion of text is about;

(C.) both people and systems make extensive use of nominal forms when searching through an index or making a query;

(D.) in absence of specific information about the participants in a given event a nominalization is used rather than a verbal form, since its syntactic expression does not require saturation of all the argument positions.

Lexical Representations for Natural Language Generation

Natural language generation (§4.5) faces the serious problem of connecting a discourse plan with the linguistic module that is responsible for the surface realization. Thus chosing a given lexical item is crucial for packaging the information which is best suited for communicating a given meaning. A number of approaches to generation have argued in favor of a strong lexicalist position where lexical choice determines grammatical realization (cf. [McD91], [Wan94]). Take for example the nominals *runner* and *skipper*. The former is productively associated with the event of running, the latter however may have the transparent sense of a productive agentive nominalization where it can refer to an individual who has skipped (e.g. the skipper of the class). Conversely, it can have an opaque sense as a person who works on a sail boat (cf. [Met92]). The lexical choice, then can be either determined by a productive process of nominalization of a particular type of event, or else it is a matter of chosing a lexically encoded lexical form, according to exactly the same rules that are involved in choosing *knife* to refer to a sharp instrument for cutting.

Lexical Representations for Multilingual Tasks

Machine translation (§4.1) poses similar problems in that generation is a crucial step for mapping an utterance in the target language, where the set of syntactic realizations associated with a lexical item might vary as the result of how a given language treats a particular semantic type. To this end the study of the lexicon cannot be separated from a cross- linguistic study of a given phenomenon, since multilingual tasks require parametric variations between languages to be stated explicitly. One such example is illustrated by the difficulty of translating English compound forms into other languages.

Lexical Representations for Information Extraction

Information extraction (§4.2) requires "shallow understanding" of real newspaper texts (articles, reports, newswires, and so on), in order to fill a set of templates, similar to semantic frames, with information derived from the text. Consider, for instance, the following article from the MUC-3 corpus (cf. [Sun91], where compounds are underlined and agentive nominals are indicated in bold face:

> [...] **INSURGENTS** FROM THE LEFTIST FARABUNDO MARTI NATIONAL LIBERATION FRONT (FMLN) STILL HELD KEY POSITIONS AROUND THE CAPITAL, FENDING OFF FIERCE SALVADORAN AIR AND ARTILLERY ATTACKS. THE POWER OUTAGE, WHICH HAD AFFECTED CERTAIN ZONES OF THE CITY AND SURROUNDING SUBURBS BUT NOT THE ENTIRE CITY OVER THE PAST WEEK OF FIGHTING, CAME

> AMID UNCONFIRMED REPORTS THAT THE **ARMY** WAS PREPARING A MAJOR COUNTER-OFFENSIVE TO DRIVE THE **REBELS** FROM THE SAN SALVADOR METROPOLITAN AREA. [...] **INSURGENTS** WHO THURSDAY HAD EVACUATED THE DISTRICT OF ZACAMIL – DEVASTATED BY THE **ARMED FORCES**' AERIAL FIREBOMBINGS – REGROUPED IN AYUTUXTEPEQUE, AN AREA NORTH OF THE CAPITAL, ACCORDING TO PRESS REPORTS. [...] IN THE CITY OF SAN MIGUEL, 140 KILOMETERS (87 MILES) EAST OF THE CAPITAL, THE **REBELS** SHOT DOWN A GOVERNMENT A-37 COMBAT PLANE, **JOURNALISTS** THERE SAID. [...] (Article from MUC-3 Corpus)

The MUC-3 task required extracting information concerning terrorist incidents and indentifying the perpetrators and the victims. Reporting this type of information from the above text involves recognizing the salient events, e.g., *rebel offensive*, *attacks*, *counter-offensive*, and the participants to those events, e.g., *rebels*, *insurgents*, *army*, *armed forces*. To this end, a system needs a basis for distinguishing the different profiles of the agentive nominals appearing in the text, e.g. *rebels* versus *army* (cf. [Hob91]), as well as, relative to those profiles which nominals are relevant for reporting the required information, e.g., *rebels*, *insurgents*, *government soldiers*, and which nouns are not, e.g., *journalists*, *press reports*. The relation between events and participants is in turn given by the properties of those events. For instance, event nominals such as *attack* and *offensive* have a rich semantics which can be roughly characterized as a violent act against individuals for the purpose of hurting them. In the wording of this description it is already implicit that a qualia-based representation can provide substantial knowledge. In discussing some of the higher-rated mistakes in information extraction based on abductive inference — another knowledge-intensive method of NLP — [Hob91] points to cases of missing axioms (or lexical knowledge) such as knowing that a *bishop* is a profession. Thus, event type and quantificational information, such as *occupations* are crucial form drawing the correct inferences concerning individuals involved in events. This indicates also that linguistically motivated lexical knowledge provides a ground for stating axioms within an inference-based system.

2.8 Adjectives

2.8.1 Introduction

Adjectives have not much been studied in traditional lexical semantics compared to the large amount of work devoted to verbs and to nouns. They nevertheless present an interesting polymorphic behaviour. Syntactically, they can appear in different positions in the sentence, as modifier of the noun (*a sad book*) or complement of a copular verb like *be* (*this book is sad*). Semantically, adjectives, more than other categories, are able to take different meanings depending on their context (for example *difficult* in *a difficult child*, *a difficult book*, *a difficult exam* or *fast* in *a fast car*, *a fast motorway*, *a fast procedure*, etc.) ([Mar83], [Lah89], etc.). As senses are only restricted by the number of possible contexts, it is questionable if they can be enumerated. Even more, richer representations and compositional mechanisms seem to be required ([Bou97].

In the following, we give first a general overview of this polymorphism. We then examine how they are represented in two models: relational (*WordNet*, §3.6) and generative (Generative Lexicon, §2.7.2).

2.8.2 Classification of adjectival polymorphism

Adjectives differ in many ways. We will examine their polymorphism from three different perspectives: syntactic, semantic and ontological.

Syntactic classes

Syntactically, adjectives can be classified with respect to three features: function, complementation and alternation.

1. **Function:** adjectives can appear in *attributive* position, as noun modifiers inside NP as shown in (54), or in *predicative* position as a complement of a verb like *be, seem, consider*, etc. as shown in (55).

   (54) A happy person

   (55) They are happy, they consider him happy, he seems happy, etc.

   This criteria allows us to distinguish three different types of adjectives: *predicative-only* (56) and *attributive-only* (57) for adjectives that are only used in one position, or *central* for those used both predicatively and attributively, as *tall* in (58) (Quirk et al., 1994, for example).

   (56) *Afraid people, people are afraid

   (57) The atomic scientist, *the scientist is atomic

   (58) The tall man, the man is tall

   When attributive, adjectives can be both *postnominal* and *prenominal* (59a), even if the postnominal position is less common in English [Sad93]. But there are restrictions: an adjective like *former* cannot be postnominal (59b) and some adjectives appear only after the noun (59c).

   (59) a. navigable rivers, rivers navigable
   b. former architect, *architect former
   c. *aforethought malice, malice aforethought

2. **Complementation:** As verbs, adjectives differ in terms of their argument structure. Many adjectives accept no complement at all (for example, *Belgian*, *red*, etc.). Those that accept complements can be subclassified as follows:

   - the type of the complement they subcategorize, i.e. prepositional as in (60a), or clausal (verb phrase as in (60b) or sentence as in (60c));

     (60) a. I'm proud of you
     b. I'm sad to leave
     c. It is possible that they leave

- whether the complement is optional or not (*I'm desirous to leave, *I'm desirous* vs *I'm sad to leave, I'm sad*);
- for prepositional complements, the preposition that marks the NP (for example *capable of, skillful at, absent from*, etc.).

3. **Alternations:** As for verbs, adjectives enter into alternations.

- Among adjectives that take clausal complement, two subclasses can be distinguished, according to whether this complement may or may not be realized as the subject. It is common to call *object embedding* adjectives whose complements cannot be realized as subject (61) and *subject embedding* those which can take a clausal subject (62) ([Pic78], [Ven68], [Arn89], [Dix91], [Sil77].

(61) a. i. I'm sad to leave
ii. To leave is sad for me

b. i. They are certain to leave
ii. To leave is certain for them

c. i. Sam was brave to leave
ii. To leave was brave of Sam

(62) a. i. I'm eager to come
ii. *To come is eager

The adjectives that allow their subject to be clausal can be further classified depending on what preposition marks the accompanying nominal complement, as in (63) [Arn89], [Sil77]. This allows [Arn89] to distinguish four classes of adjectives: *S-only* and *S+toNP* adjectives indicate some perception of the modality of the clausal clause; *S+ofNP* adjectives give an evaluation of an individual, on the presupposition that he is responsible for the state of affairs described in the proposition. Finally, *S+forNP* adjectives characterize directly a state of affaires, but may also indicate an experiencer.

(63) a. That they left is possible (S-only)

b. That they left is clear to Sam (S+toNP)

c. That he left was brave of him (S+ofNP)

d. To ignore pollution is dangerous for the country (S+forNP)

- *It-extraposition*: All adjectives that are suject-embedding allow the it-extraposition (64).

(64) a. It is possible that they left

b. It is clear to Sam that hey left

c. It was brave of him that he left

d. It is dangerous for the country to ignore pollution

- *Tough-construction*: some *s+forNP* adjectives allow the tough-construction (65b), where the subject of the adjective is understood as filling a non-subject position in the complement of the adjective (see for example [Arn89] for an overview).

(65) a. It is difficult to support this proposal

b. This proposal is difficult to support

Logical classes

Semantically, adjectives can belong to three different classes, which differ in their logical behaviour in the following way ([Chi90]; [Arn89]; [Par90], pp. 43-44).

1. An adjectives (ADJ) is said *absolute* (or *intersective*, *predicative*, etc.) if (66a) implies (66b) and (66c). These adjectives are standardly analyzed as predicates: they denote properties and the denotation of the adjectif-noun construction is the intersection of the denotation of the ADJ and the N.

   (66) a. this is a red (ADJ) table (N)

   b. $\rightarrow$ this individual is a N

   c. $\rightarrow$ this individual is ADJ

   Typical examples of this category are adjectives wich denote:

   i. a shape: *hexagonal*

   ii. a social group or a nationality: *communist, Belgian*, etc.

   iii. a color

2. An adjective is *property-modifying* (or *non-intersective*, *operators*, etc.) if (67a) does not imply (67b), nor often (67c): *a former architect* is not *an architect*, nor *former*. These adjectives have been analyzed as operators: they denote a function from properties to properties. Thus, for example, *former* in (67a) takes the denotation of *architect* to the set of individuals who used to be architect.

   (67) a. this is a former (ADJ) architect (N)

   b. $/\rightarrow$ this individual is a N

   c. $/\rightarrow$ this individual is ADJ

   Property-modifying adjectives include: nominal (or relational) adjectives (*polar bear*, *atomic scientist*, etc. Cf. [Lev78], manner (or adverbial) adjectives (*a poor liar*, *a fast car*), emotive (*a poor man*) and modals, i.e. all adjectives which are related to adverbs, quantifiers or determiners (*a feeble excuse*, *the specific reason*, *a fake nose*, etc.).

3. An adjective is *relative* (or *scalar*) if (68a) implies (68b), but not (68c); *a big mouse*, for example, is not *a big animal*. As absolute adjectives, they characterize the individual described by the noun (68b), but, unlike them, it is relative to some norm or standard of comparison: *a big mouse* is *big for an F*, where F is supplied by the context (68d). As they share properties with absolute and relative adjectives, they have been analyzed both as predicates and operators ([Par90], p. 44): on the predicative treatment, *x is a clever N* for example, means therefore *x is an N & x is clever for an F* and on the operator treatment, it means *clever(x is an N that is F)*.

(68) a. this is a big mouse
b. $\rightarrow$ this individual is a N
c. $/\!\!\rightarrow$ this individual is Adj
d. $\rightarrow$ this individual is Adj for a F

Other semantic classes

The adjectives can also be classified with respect to other semantic features [Qui94], as:

1. **Aspect:** an adjective can be *stage-level* (if it expresses a temporary or accidental property) as in (69) or *individual-level* (in case of a generic, permanent or inherent property), see (70) [Car77], [Car95].

   (69) *drunk, available* , etc.

   (70) *clever, tall*, etc.

2. **Gradation:** an adjective can be gradable or not.

Adjective taxonomies

Adjective taxonomies classify adjectives in the different semantic categories they can express (see [Ras95] for a good introduction). [Dix91] is one of the most representative. He classifies adjectives as the following:

(71)
1. DIMENSION: *big, short*, etc.
2. PHYSICAL PROPERTY: *strong, ill*, etc.
3. SPEED: *fast, quick*, etc.
4. AGE: *new, old*, etc.
5. COLOR: *red, black*, etc.
6. VALUE: *good, bad*, etc.
7. DIFFICULTY: *easy, difficult*, etc.
8. QUALIFICATION: DEFINITE (*probable*), POSSIBLE (*possible*), USUAL (*usual*), LIKELY (*likely*), SURE (*sure*), CORRECT (*appropriate*)
9. HUMAN PROPENSITY: FOND (*fond*), ANGRY (*angry*, *jealous*), HAPPY (*anxious*, *happy*), UNSURE (*certain*), EAGER (*eager*, *ready*), CLEVER (*clever*, *stupid*, *generous*)
10. SIMILARITY: *similar*, *different*, etc.

For each class, Dixon specifies the syntactic behaviour of each adjective, as follows: "EAGER takes an NP or a THAT or MODAL (FOR) TO complement, e.g. *I'm eager for the fray*, *I'm eager that Mary should go*, *I'm eager (for Mary) to go*. *Ready* may only take an NP or a Modal (FOR) TO clause (not a THAT complement) while *willing* must take a THAT or Modal (FOR) TO clause, i.e. it cannot be followed by proposition plus NP." ([Dix91], p. 83).

These classifications have a descriptive value, but none of these types of taxonomies have been applied in practical applications ([Ras95], p. 10). They pose two main problems: first, they reveal little about the functional and relational properties of the adjective; secondly, they don't explain why adjectives share or do not share some syntactic behaviour.

2.8.3 Representation of the properties

Introduction

The former section identified the different properties of adjectives. This one examines how they are represented and generalized in two major models: *WordNet* and *Generative Lexicon*.

Semantic network, *WordNet*

In *WordNet* (§3.6), adjectives are divided in two main classes which are said to account for the majority of adjectives: *ascriptive* wich are considered to ascribe a value of an attribute to a noun and *nonascriptive* which are similar to nouns used as modifiers ([Gro90]).

Ascriptive adjectives are organized in terms of antonymy and synonymy, as in (72):

(72) {DRY, WET1,! anhydrous,& arid,& ...}
{anhydrous, dry1, & }
{arid, waterless,dry1, & }
{dehydrated, desiccated, parched, dry1, & }
{dried, dry1, & }
...

Non-ascriptive ones are considered as stylistic variants of modifying nouns and are cross-referenced to the noun files. For example, the entry (73) indicates that *astral* and *stellar* have the meaning of *pertaining to a star or stars*

(73) {star | astral, stellar}

Gradation is not indicated in *WordNet* (§3.6) because it is not often lexicalized in English. Restrictions on syntactic position (for **p**renominal-only and **p**ostnominal-only adjectives) are encoded for specific word senses, as they cannot be generalized over entire sets of synonyms.

WordNet does not say anything about the way senses are related: adjectives have as much senses as synsets. Moreover, it does not provide the means to predict grammatical properties from the representation (complementation, alternations, selective restriction). These two features distinguish the relational approach of *WordNet* from a Generative approach, like Generative Lexicon.

Generative Lexicon

The Generative Lexicon [Pus95a] focusses on the two aspects neglected in *WordNet*:

1. how the different adjectival senses are related and how they can be derived compositionaly from the representations of the noun and the adjective [Bou97], and

2. the syntax-semantics interface. In this theory, the adjectival polymorphism is explained by richer representations of adjectives and nouns (the qualia structure) and the way they combine together.

Take as an example the ambiguity of the French adjective *vieux* (*old*) in *un vieux marin* (*an old sailor*) which can be both relative (*aged*) and property modifying (with the meaning *who has this capacity for a long time*, without being necessarily aged). It can be explained in the following way (cf. [Bou97] for more details).

1. Semantics of the noun: nouns to which adjectives apply have complex representations, which define the different predicates necessary for the definition of the word ([Bus96a]); a sailor, for example, is defined as *somebody who has the capacity of sailing*, i.e. as the conjunction of three types: human, state (*have the capacity to*) and event (*to sail*).

2. Semantics of the adjective: the semantics of *vieux* indicates that this adjective has two functional types: it can apply to an individual (*un vieil homme*) or a state (*une vieille amitié*).

3. Composition: As *vieux* has two different functional types and as*marin* is defined by the conjunction of these types, it can apply to both of them, giving rise to the ambiguity of the adjective-noun construction: in one sense, the adjective applies to the type *human* denoted by *marin* (*un vieux marin* is then understood as *an old individual who has the capacity of sailing*); in the other, it applies to the state (i.e. *have a capacity*) (*un vieux marin* is then interpreted as *somebody who has had this capacity for a long time*).

With this kind of treatment, adjectives which belong to different logical classes are not considered as homonymous. The different senses can be derived from the semantics of the noun and the adjective.

2.9 Prepositions

2.9.1 Introduction

Prepositions have not been much studied in traditional lexical semantics compared to the large amount of work devoted to verbs and to nouns. They nevertheless play a very important role in semantics and have many connections with most of the other syntactic categories. They have been, however, extensively studied in psycho-linguistics, in particular spatial prepositions. Previous work includes [Zel93], [Van94] and [Lin97].

2.9.2 Main organization of prepositions

Prepositions occur in the following three main syntactic situations:

- as heads of prepositional phrases,

- as an element in verbal compounds (e.g. turn out, give up, aufsteigen, anrufen, etc. or a more complex form as: sich fressen gegenseitig auf), where they largely contribute to the semantics of the expression. This compound is subject to various more or less strong constraints, in particular in Germanic languages,

- some prepositions function s case markers as in: *donner qqchose à Marie* where [à Marie] can possibly be analysed not as a PP but as an NP[prep=a] or a NP[case=dative].

- as a 'connector' in nominal compounds (planche à voile, plancha a vela, plan de vuelo),

- as introducing small clauses in a number of languages,
- preposition may be incorporated [Bak88] (to climb (up) versus to climb down, where down is obligatory whereas up may be incorporated into the verb).

When dealing with prepositions, a clear distinction appears between **semantically full** and **semantically empty** prepositions.

- **Semantically empty** prepositions are prepositions which are strongly bound by a predicative head, like *on*, in the example: *John relies on his father* where *on* cannot be substituted by a semantically close preposition like *in* or *over*. They are considered to be **case markers** by some approaches (as suggested by [Pol87] and developped by [Bad96]). These prepositions are lexically predicted by the predicative head and, therefore, do not contribute to the semantics of the sentence. In a Machine Translation system like EUROTRA (§3.11.1), they are featurised (o elevated) in the process of parsing: that is, they disappear as lexical nodes and their lexical value is encoded as an attribute of the predicative head. This implies that they are not translated, ut are generated in the target language from the information encoded in the entry of the predicative head.
- **Full prepositions**, on the contrary, are heads of prepositional phrases (see §2.9.2) . They always subcategorize for (only) one argument, which in most cases is an NP, but can also be a finite or no inite clause, an adverb or even a PP. Full prepositional phrases may function as:
  - **adjuncts** (weakly bound by a predicative governor), like *on*, in: The book lies on the table where *on* can be substituted by a semantically close preposition, because the predicate demands a locative complement and not a specific preposition.
  - or **modifiers** (like *for*, in: John works for his father where the prepositional phrase is not particularly predicted by the verb, and the semantics it conveys (benefactive) may apply to other verbs as well: John (dances /bakes a cake / fixes the TV set) for his father

Let us now examine a simple semantic classification for prepositions. Note first that prepositions are highly polysemous; they are also subject to many metaphorical extensions, for example spatial prepositions have for most of them a metaphorical extension to abstract entities, viewed as locations (e.g. against a wall → against some principles). We can however tentatively propose the following general classification, where some items may overlap on different ontological domains (e.g. time and space):

- temporal prepositions:
  - expressing anteriority (before, until),
  - expressing duration (durant, pendant, while),
  - expressing posteriority (after, despuès, as soon as).
- spatial prepositions:
  - expressing source (from, à partir de, depuis, since),

  - expressing position, either absolute (en, chez) or relative (under, in, infront of),
  - goal, reached, (à),
  - direction, not necessarily reached (via, toward, to, vers, hasta).

- prepositions expressing cause (because of, lors de, par, suite à),
- prepositions expressing consequences (in order to, de façon à),
- prepositions expressing goals or purpose (for, with the intention of, afin de),
- prepositions expressing conditions (à condition de),
- prepositions expressing means or manner:
  - manner (en selon),
  - accompaniement (with),
  - instrumental means (avec, by means of par).

Prepositions across languages

The use of prepositions can, in general, be specified quite unambiguously. For example, for an event occurring after a certain date, the prepositions *after, nach(dem), na(dat) après, despuès* will be respectively used for English, German, Dutch, French and Spanish, if the action takes place before the date, then we respectively have *before, (be)vor, voor(dat), avant, antes.*

Prepositions do not have often direct correspondences across languages. For example if S is an event, we have the following preposition lexicalizations depending on the duration and the type of event:

1. simple point: *at, bei, bij, à* (English, German, Dutch, French),
2. S has a duration, and happened once in the past: *when/as, als, toen, lors de*; als becomes wie in case of present tense in German,
3. otherwise: *when/as, wenn, als/wanneer, lors de.*

From a different perspective *around* translates in German as *gegen, um*, *at* translates as *um, zu, bei, an, in* and *by, until* translate as *bis.*

Similarly, verbs may select very different types of prepositions in two different languages:
dream about someone,
réver de/à quelqu'un,
soñar con alguien. (Spanish: to dream with someone)
Whereas *about* and *de* have about equivalent meanings, *con* is completely different. There are many divergences of this type which make machine translation quite difficult.

In the context of a multilingual system, the need for assigning some sort of semantics to **modifying prepositional phrases** may be illustrated by these two facts

- **structural disambiguation**: i.e. ambiguous PP attachment, as *on* in: John read the report on new technologies where *on* may be attached to the verb *read* or to the noun *report.*
- **lexical selection** of the target preposition in transfer:

```
        Italian                       Spanish
per (benefactive)     =>                para
per (cause)           =>                por
```

Italian *per* incorporates both the meanings of Spanish *para* and *por*: without any kind of semantic information it is not possible to decide which translation must be chosen, as may be seen in the following examples. The first set corresponds to benefactive *per* and the second to causative *per*:

1. Ha scritto una lettera *per* la sua fidanzata
Ha escrito una carta *para* su novia
(He) has written a letter *for* her girlfriend
2. Ha scritto una lettera *per* paura
Ha escrito una carta *por* miedo
(He) has written a letter *out of* fear.

2.9.3 Prepositions and Semantic Relations for Modifiers

An experiment on Semantic Relations for Modifiers was carried out in the context of the Eurotra (§3.11.1) project by some language groups during 1989 and was afterwards adopted as legislation by the rest of the groups.

From the examples above, it seems clear that a mere lexical typification of the preposition (like *benefactive* or *causative*) is not sufficient to perform the disambiguation, due to the polysemous nature of most prepositions. A calculation of the semantic value (or relation) of the whole **prepositional phrase** is needed. Thus, the semantic relation (or *modsr* value) of a PP of the form: P + NP has to be calculated on the basis of the lexical value of the preposition P combined with the lexical semantic value of the NP. Thus, for instance, the preposition *with* may have several semantic values, depending on the NP that follows it:

```
PP[ with + NP (sem=hum)] =>  PP[modsr=accompaniment]

PP[ with + NP (sem=non-hum)] => PP[modsr=instrument]
```

It appears that some set of lexical semantic features (LSF) for nouns is needed, but since the process of calculation is internal to each language module, each of them is free to use its own set, which may have different degrees of granularity. On the other hand, the set of modsr labels belongs to the Eurotra Interface Structure (IS) and thus is shared by all languages. Here follows the list of modsr values that was agreed upon. They are divided into two big groups:

- those that apply to PPs modifying predicative heads (verbs and predicative nouns)
- those that apply to PPs modifying nonpredicative heads (nonpredicative nouns)

This distinction is justified because, from a semantic point of view, the relation between the modifier and the head is of a different type in both cases, as sustained by [Pol87]. Also , from a syntactic point of view, the choice of the preposition for modifiers of nominal heads is much more restricted, especially in the Romance languages where the most frequent is *of* (*e* in French, Spanish, Portuguese and Catalan, and *di* in Italian). We could say that nonpredicative (or nominal) prepositional modifiers behave like adjectives while predicative prepositional modifiers behave like adverbs. Predicative nouns are a bit special because they combine a verbal and a nominal nature.

- Modifiers of predicative heads:
  1. Place-position: 'The report is ON THE TABLE'. It is further subdivided in positioninon
  2. Place-goal: 'The boy is looking TOWARDS THE BEACH'
  3. Place-origin: 'FROM THIS WINDOW, the views are magnificent'
  4. Place-path: 'He ran THROUGH THE FIELDS'
  5. Cause: 'He died BECAUSE OF THE COLD'
  6. Aim: 'He did it TO SAVE HIMSELF'
  7. Concern: 'She told me ABOUT HER FRIEND'
  8. Accompaniment: 'The soldiers came WITH THEIR GIRLFRIENDS'
  9. Instrument: 'They bombed Baghdad WITH MISSILES'
  10. Benefactive: 'He turned on the heater FOR LINDA'
  11. Substitute: 'He appeared on the news, speaking FOR THE STATE DEPARTMENT'
  12. Manner: 'The government reacted WITH CAUTION'
  13. Function: 'I tell you this AS A FRIEND'
  14. Measure: 'Fabrics are sold BY THE METRE'
  15. Comparison: 'Columbia was a young city COMPARED TO VENERABLE CHARLESTON'
  16. Time: 'He only sees her AT CHRISTMAS AND EASTERN'
- Modifiers of non-predicative heads:
  1. Quality-place: it may be further subdivided in:
     (a) place = position: 'a brick IN THE WALL'
     (b) place = goal: 'the train TO LONDON'
     (c) place = origin:'oranges FROM SPAIN'
     (d) place = path: 'a road THROUGH THE DESERT'
  2. Quality-whole: 'the other side OF THE SQUARE'
  3. Quality-stuff: 'loose-fitting garments OF LINEN'
  4. Quality-quality: it may be further subdivided in:
     (a) quality = inherent: 'a car WITH FIVE DOORS'
     (b) quality = state: 'a car WITH A BROKEN GLASS'
  5. Quality-concern: 'a report ON THE IMPLEMENTATION'
  6. Quality-aim: 'a room FOR RENT'
  7. Quality-specification: 'the island OF CUBA'
  8. Quality-function: 'Clinton, AS PRESIDENT OF THE US'
  9. Quality-measure: 'a project OF SEVERAL MILLIONS'

2.9.4 Relations with other areas of lexical semantics

The relations that prepositions have with other elements of lexical semantics are the following:

- Prepositions head prepositional phrases, as such they impose selectional restrictions, e.g. a spatial proposition expects an NP of type location. Prepositions can then be associated with a kind of subcategorization frame.
- Prepositions assign thematic roles to the NP they head, they can therefore be associated with a thematic grid, similarly to verbs.
- We may also consider that they assign a thematic role to other NPs involved in the relation described by the preposition.
- Prepositions play an important role in the definition of verb syntactic alternations (see §2.6.2), e.g. the into/onto, on/with and conative alternations [Lev93].
- Preposition play a crucial role for identifying the semantics of modifiers, in particular they indicate means, manner, purpose, time, location, accompaniment and amount (§2.6.2).

2.9.5 Prepositions in LKB

Prepositions being a closed set of words are usaully described in depth in most lexical knowledge bases. They can be grouped into families as suggested above in §2.9.1. They usually share a number of properties, but are not in general included into major taxonomies.

2.9.6 Prepositions in Applications

As already shown above, the translation of prepositions provides a strong challenge for MT applications (§4.1). Lexical semantics is of much use to help disambiguate and select the right prepositions (see [Tru92] and related discussions in §3.11, 3.12, 3.9 and 5.3).

Another important role of prepositions is to contribute to solving ambiguities in PP-attachment. This is also a complex task, but the semantics of prepositions can be used to decide whether the PP is a verb complement (argument or modifier) or a noun complement.

Finally, prepositions convey a quite important semantic load, in spite of their high polysemic nature. In particular, they are important in identifying the semantics of modifiers. From this point of view, the information they carry may turn out to be useful in information retrieval/extraction systems (§4.3, 4.2).

Part II

Lexical Semantic Resources

Chapter 3

Lexical Semantic Resources

3.1 Introduction

In this section an overview is given of a variety of resources that contain lexical semantic information. There are many more resources available but here we only describe a selection illustrating the different types of resources, and focusing on the use of these resources for NLP. The selection covers: traditional machine readable dictionaries, wordnets, taxonomies, dictionaries with feature classifications, MT-lexicons, formalized lexical knowledge bases for NLP, resources for expert systems, traditional bilingual dictionaries. It is not easy to divide the lexical resources in more overall groups because each resource often contains different mixtures of data. For example, it seems obvious to group the Princeton WordNet1.5 with the EuroWordNet data as semantic networks but EuroWordNet also contains equivalence relations, making it a multilingual resource, and it incorporates a formalized ontology, making it partly a conceptual knowledge base. Similarly, we could make a division between traditional resources such as the Longman Dictionary of Contemporary English (LDOCE) and lexical resources developed for NLP purposes, but LDOCE includes a feature system (both syntactic and semantic) making it very useful as an NLP resource, as often has been shown. We therefore provide some grouping when it is obvious but also discuss resources individually, without however suggesting a difference in relevance or detail.

For each of these resources we give a description of the semantic content, both in terms of quantity and kind of data. Where possible we will describe the most frequent semantic values. This may either be the top-levels of hierarchies, definitions of features and concepts, or definitions of relations. Furthermore, we will provide references to the lexical semantic notions described in the previous section and the usage of the resources in NLP.

3.2 The Longman Dictionary and Thesaurus

3.2.1 Introduction

The Longman Dictionary and the Longman Lexicon of Contemporary English have extensively been used in the pioneer work to extract NLP-lexicons from Machine-Readable Dictionaries. Many of the insights for building large-scale NLP lexicons have been based on studies of these resources. Because of their age, their organization and structuring is still based on the traditional practice of making dictionaries, but certain features have made them particularly

	Entries	Senses	Polysemy
Nouns	23800	37500	1.6
Verbs	7921	15831	1.9
Adjectives	6922	11371	1.6
Total	38643	64702	1.7

Table 3.1: Number of Entries and Senses in LDOCE

suitable for deriving NLP-lexicons.

3.2.2 The Longman Dictionary of Contemporary English

The Longman Dictionary of Contemporary English [Pro78] is a middle-size learner's dictionary: 45,000 entries and 65,000 word senses. Entries are distinguished as homographs on the basis of the historic origin of words and their part-of-speech, where each entry may have one or more meanings. The entry-sense distributions for the major parts of speech are as shown in Table 3.1.

The information provided in entries comprises:

- Definitions using a limited set of 2000 Controlled Vocabulary Words and 3000 derived words.
- Examples.
- Grammatical information on the constituent structure of complementation of the words. LDOCE is mostly known for its high-quality grammatical coding, however, since the focus is here on semantics these are not further specified here.
- Usage labels in the form of codes and comments, covering register , style (11 codes), dialect (20 codes) and region constraints (9 codes).
- Subject Field codes and comments indicating the domain of interest to which a meaning is related.
- Semantic codes either classifying nominal meanings or expressing selectional restrictions for the complementation of verbal and adjectival meanings.

Most of the information is stored in textual form. However, the usage codes, the subject-field code and the semantic codes are stored in the form of a unique code system.

There are 100 main Subject Field codes which can be subdivided as follows:

MD medical

MDZA medical anatomy

ON occupation

VH vehicles

The Subject Field Codes have been stored for 30% of the verb senses and 59% of the noun senses. There are 100 main fields and 246 subdivisions. Two main fields can also be combined, MDON represents both *medical* and *occupation.*

In total, there are 32 different semantic codes in LDOCE. A distinction can be made between basic codes (19 codes) and codes that represent a combination of a basic code (13 combinations):

A Animal

B Female Animal

C Concrete

D Male Animal

E Solid or Liquid (not gas): S + L

F Female Human

G Gas

H Human

I Inanimate Concrete

J Movable Solid

K Male Animal or Human = D + M

L Liquid

M Male Human

N Not Movable Solid

O Animal or Human = A + H

P Plant

Q Animate

R Female = B + F

S Solid

T Abstract

U Collective Animal or Human = (Collective + O)

V Plant or Animal = (P + A)

W Inanimate Concrete or Abstract = (T + I)

X Abstract or Human = (T + H)

Y Abstract or Animate = (T + H)

Z Unmarked

1 Human or Solid = (H + S)

2 Abstract or Solid = (T + S)

4 Abstract Physical

5 Organic Material

6 Liquid or Abstract = (L + T)

7 Gas or Liquid = (G + L)

The basic codes are organized into the hierarchy shown in Figure 3.1 Most noun senses have

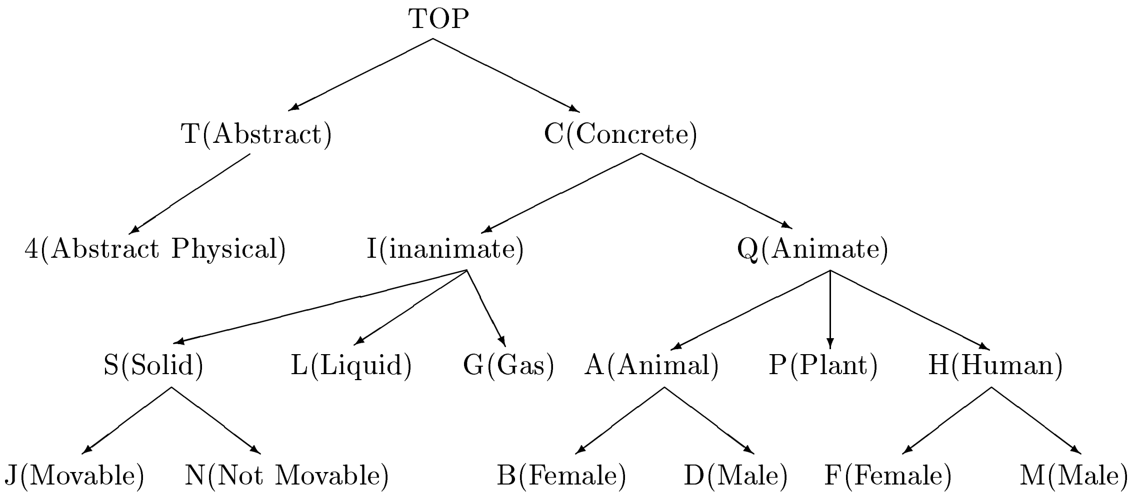


Figure 3.1: Hierachy of semantic codes in LDOCE

a semantic code. In the case of nouns these codes can be seen as a basic classification of the meaning. In the case of verbs and adjectives however the codes indicate selection restrictions of their arguments. These selection restrictions can also be inferred from their definitions in which constituents corresponding with the complements of the defined verbs or adjectives have been put between brackets.

3.2.3 The Longman Lexicon of Contemporary English

LLOCE, the *Longman Lexicon of Contemporary English*, is a small size learner style dictionary largely derived from LDOCE and organized along semantic principles. A quantitative profile of the information provided is given in the table below.

Number of entries	16,000	
Numer of senses	25,000	
Semantic fields	Major codes	14
	Group codes	127
	Set codes	2441
Grammar codes	same as LDOCE	
Selectional restrictions	same as LDOCE	
Domain & register Labels	same as LDOCE	

Semantic classification in LLOCE is articulated in 3 tiers of increasingly specific concepts represented as major, group and set codes, e.g.

```
<MAJOR: A> Life and living things
                  |
<GROUP: A50-61> Animals/Mammals
                  |
<SET: A53> The cat and similar animals:
              cat, leopard, lion, tiger,...
```

Each entry is associated with a set code, e.g.

```
<SET: A53> nouns The cat and similar animals
--------------------------------------------
cat 1 a small domestic [=> A36] animal ...
    2 any animal of a group ...
...
panther [Wn1] 1 a leopard ...
              2 AmE cougar.
...

<SET: A53> nouns The dog and similar animals
--------------------------------------------
dog a domestic animal with a coat of hair ...
```

Relations of semantic similarity between codes not expressed hierarchically are crossreferenced, e.g.

```
<SET: A53> nouns The cat and similar animals
--------------------------------------------
cat 1 a small domestic [=> A36] animal ...
                       ^^^^^^^^


<SET: A36> Man breeding living things
-------------------------------------
....
```

There are 14 major codes, 127 group codes and 2441 set codes. The list of major codes below provides a general idea of the semantic areas covered:

```
<A> Life and living things
<B> The body, its functions and welfare
<C> People and the family
<D> Buildings, houses, the home, clothes, belongings, and personal care
<E> Food, drink, and farming
<F> Feelings, emotions, attitudes, and sensations
<G> Thought and communication, language and grammar
<H> H Substances, materials, objects, and equipment
<I> Arts and crafts, sciences and technology, industry and education
<J> Numbers, measurement, money, and commerce
<K> Entertainment, sports, and games
<L> Space and time
<M> Movement, location, travel, and transport
<N> General and abstract terms
```

The list of group and set codes for the `M` domain (`Movement, location, travel, and transport`) given in Table 3.2 provides an example of the degree of details used in semantic classification.

3.2.4 Comparison with Other Lexical Databases

LDOCE is a traditional Machine-Readable Dictionary. However, because of its controlled-vocabulary, the systematic coding of the information and the elaborate use of codes it has been a very useful starting point for deriving basic NLP lexicons. [Bri89] give an extensive description of the possibilities for elaboration. Except for the semantic features, LDOCE does not contain complete semantic hierarchies as in WordNet, EDR or other ontologies.

The bottom level of word sense clustering in LLOCE consists of sets of semantically related words which need not be synonyms. For example, the set D172 (`baths and showers`) contain nouns such as *bath, shave, shower*. This contrasts with lexical databases such as WordNet where synsets are meant to contain synonymous word senses.

A further difference with WordNet regards taxonomic organization. In Wordnet, hierarchical relations are mainly encoded as hyp(er)onymic links forming chains of synsets whose length can vary considerably. In LLOCE there are only three tiers and considerable crossreferencing. Moreover, only the terminal leaves of the LLOCE taxonomy correspond to actual

Moving, coming, and going
- M 1 moving, coming, and going
- M 2 (of a person or object) not moving
- M 3 stopping (a person or object) from moving
- M 4 leaving and setting out
- M 5 arriving, reaching, and entering
- M 6 letting in and out
- M 7 welcoming and meeting
- M 8 getting off, down, and out
- M 9 climbing and getting on
- M 10 movement and motion
- M 11 staying and stopping
- M 12 passages, arrivals, and departures
- M 13 climbing, ascending, and descending
- M 14 moving
- M 15 not moving
- M 16 moving quickly
- M 17 not moving quickly
- M 18 speed
- M 19 particular ways of moving
- M 20 walking unevenly, unsteadily, etc
- M 21 walking gently, etc
- M 22 walking strongly, etc
- M 23 walking long and far, etc
- M 24 running and moving quickly, etc
- M 25 running and moving lightly and quickly, etc
- M 26 crawling and creeping, etc
- M 27 loitering and lingering, etc
- M 28 flying in various ways
- M 29 driving and steering, etc
- M 30 going on a bicycle, etc
- M 31 moving faster and slower
- M 32 coming to a stop, moving away, etc
- M 33 hurrying and rushing
- M 34 following, chasing, and hunting
- M 35 escaping, etc
- M 36 things and persons chased, etc
- M 37 avoiding and dodging
- M 38 leaving and deserting
- M 39 moving forward, etc
- M 40 turning, twisting, and bending
- M 41 flowing
- M 42 coasting and drifting
- M 43 bouncing and bobbing

Putting and taking, pulling and pushing
- M 50 putting and placing
- M 51 carrying, taking, and bringing
- M 52 sending and transporting
- M 53 taking, leading, and escorting
- M 54 sending and taking away
- M 55 showing and directing
- M 56 pulling
- M 57 pulling out
- M 58 pushing
- M 59 throwing
- M 60 throwing things and sending things out
- M 61 extracting and withdrawing
- M 62 sticking and wedging
- M 63 closing, shutting, and sealing
- M 64 fastening and locking
- M 65 opening and unlocking
- M 66 open and not open
- M 67 openings

Travel and visiting
- M 70 visiting
- M 71 inviting and summoning people
- M 72 Meeting people and things
- M 73 visiting and inviting
- M 74 travelling
- M 75 travelling
- M 76 people visiting and travelling
- M 77 people guiding and taking
- M 78 travel businesses
- M 79 hotels, etc
- M 80 in hotels, etc
- M 81 people in hotels, etc
- M 82 in hotels, travelling, etc
- M 83 in hotels, travelling, etc

Vehicles and transport on land
- M 90 transport
- M 91 vehicles generally
- M 92 special, usu older, kinds of vehicles
- M 93 lighter motor vehicles, etc
- M 94 heavier motor vehicles
- M 95 buses, etc
- M 96 bicycles and motorcycles, etc
- M 97 persons driving vehicles, etc
- M 98 smaller special vehicles, etc
- M 99 vehicles for living in
- M 100 parts of vehicles outside
- M 101 parts of vehicles inside
- M 102 the chassis and the engine
- M 103 parts of a bicycle
- M 104 related to motocycles
- M 105 garages and servicing
- M 106 trams
- M 107 railways
- M 108 trains
- M 109 places relating to railways, travel, etc
- M 110 persons working on railways, etc
- M 111 driving and travelling by car, etc
- M 112 crashes and accidents

Places
- M 120 places and positions
- M 121 space
- M 122 edges, boundaries, and borders
- M 123 neighbourhoods and environments
- M 124 at home and abroad
- M 125 roads and routes
- M 126 special roads and streets in towns
- M 127 special roads and streets in the country
- M 128 special streets in towns
- M 129 very large modern roads
- M 130 no-entries and cul-de-sacs
- M 131 paths and tracks
- M 132 parts of roads, etc
- M 133 lights on roads, etc
- M 134 bends and bumps, etc
- M 135 intersections and bypasses
- M 136 bridges and tunnels

Shipping
- M 150 boats
- M 151 boats in general
- M 152 smaller kinds of boats
- M 153 larger kinds of sailing boats
- M 154 powered ships
- M 155 ships with special uses
- M 156 merchant ships, etc
- M 157 parts of ships
- M 158 positions on ships, etc
- M 159 harbours and yards
- M 160 quays and docks
- M 161 lighthouses, buoys, etc
- M 162 crews
- M 163 sailors, etc
- M 164 ship's officers, etc
- M 165 mooring and docking
- M 166 setting sail
- M 167 oars and paddles
- M 168 floating and sinking, etc
- M 169 wrecking and marooning, etc

Aircraft
- M 180 aircraft and aviation
- M 181 jet aeroplanes
- M 182 balloons, etc
- M 183 helicopters
- M 184 spaceships
- M 185 airports
- M 186 parts of aircraft
- M 187 landing and taking off
- M 188 landing and taking off
- M 189 people working on and with aeroplanes

Location and direction
- M 200 surfaces and edges
- M 201 higher and lower positions in objects, space, etc
- M 202 front, back, and sides
- M 203 about and around, etc
- M 204 in, into, at, etc
- M 205 out, from, etc
- M 206 here and not here
- M 207 across, through, etc
- M 208 against
- M 209 near
- M 210 far
- M 211 between and among
- M 212 away and apart
- M 213 back and aside
- M 214 to and towards
- M 215 from place to place
- M 216 on and upon
- M 217 off
- M 218 below, beneath, and under
- M 219 above and over
- M 220 after and behind
- M 221 in front, before, and ahead
- M 222 through and via
- M 223 past and beyond
- M 224 up
- M 225 down

Table 3.2: Set codes for the domain of *Movement, location, travel and trasport* in LLOCE.

word senses; the labels associated with intermediate levels (major, group and set codes) are abstractions over sets of semantically related word senses, just like the *intermediate concepts* used in the EDR (see §3.8).

3.2.5 Relations to Notions of Lexical Semantics

The semantic codes for nouns in LDOCE represents very minimal and shallow. The LLOCE classification is more elaborated but is still not very deep. This classification information is similar to the taxonomic models described in §2.7.

LLOCE in addition combines the entry format of LDOCE, which provides detailed syntactic information (in the form of grammar codes) with the semantic structure of a thesaurus. This combination is particularly well suited for relating syntactic and semantic properties of words, and in particular to individuate dependencies between semantic predicates classes and subcategorization frames as described in §2.4.

3.2.6 LE Uses

LDOCE has been most useful as a syntactic lexicon for parsing. The usage of LDOCE as a semantic resource is not as wide-spread as one would expect. This is mainly due to its restricted availability and the fact that it still requires considerable processing to derive a full-covarge NLP lexicon from it. [Bri89] give an overview of the different kind of NLP lexicons that can be derived from it. [Vos95b] give a description how a richly encoded semantic lexicon with weighted features can be derived which is used in an information retrieval task.

[San92a] and [San93b] use LLOCE to derive verb entries with detailed semantic frame information. [Poz96] describe a system which uses LLOCE to assign semantic tags to verbs in bracketed corpora to elicit dependencies between semantic verb classes and their admissible subcategorization frames.

3.3 Cambridge International Dictionary of English

The Cambridge International Dictionary of English (published 1995) is a large corpus-based advanced learner's dictionary, the text of which is available as an SGML file. This file includes approximately 80,000 sense entries (including some undefined derived runons as senses) and 110,000 example sentences. A typical sense entry includes:

- definitions (written in a 2,000 word defining vocabulary);
- guide words (distinguishing different senses of the same word);
- grammatical coding;
- register and variety coding, and
- example sentences with highlighted collocates.

There are also a number of encoding schemes that did not appear in the printed book but which were entered with both NLP purposes and future lexicographic development in mind, including:

- subject domain coding (each sense/example classified into 900 categories hierarchically structured in 4 levels)
- selectional preference coding (each noun sense classified into 40 categories; these categories then used to classify the possible fillers for each verb/adjective/preposition sense/example)
- multi-word unit tagging (marking grammatical function and inflectability of each component word)

This data, developed partly in conjunction with Cambridge Language Services Ltd under the UK government DTI-SALT funded project Integrated Language Database, is available as part of an SGML file package termed CIDE+ (see http://www.cup.cam.ac.uk/elt/reference/data.htm).

Subject domain coding The initial purpose of the subject domain coding scheme was to enable entries to be selected for revision by subject specialists, thus increasing the accuracy of particularly the definitions and examples and also ensuring that the most important specialist words were included in the dictionary. This coding was done not just at the entry level but also at the example level, for instances where either the examples illustrated shades of meaning specific to a particular subject domain, or where the examples were in a particular subject domain context (not necessarily to do with the word being exemplified).

There are 900 subject code categories hierarchically structured in 4 levels. The top level is as follows:

```
A   Agriculture
B   Building & Civil Engineering
C   Arts & Crafts
D   Food & Drink
E   Earth Science (include Outer Space)
F   Finance & Business
G   Sports (include Games & Pastimes
H   History
I   Media & Publishing
J   Crime & the Law
K   Language & Literature
L   Life Science
M   Maths, Physics & Chemistry
N   Entertainment & Music
O   Medicine, Health & Psychology
P   Names of Plants & Animals
Q   Education
R   Religion
S   Society
T   Technology
U   Politics, Diplomacy & Government
V   Travel & Transport
W   War & the Military
X   General Language
```

```
Y   Mythology, Occult & Philosophy
Z   Clothes, Fashion & Beauticulture
```

As the coding is for subject domain rather than semantic similarity, the coding was inappropriate for abstract vocabulary which was typically encoded under the category X; this accounted for approximately 30% of the senses in CIDE+.

The fine detail of the coding can be illustrated by the following sub-section:

```
F - FINANCE AND BUSINESS
    1. Non-technical words (double code if necessary) eg fee, thrifty,
     pay, fare
    2. Informal financial terms eg tick, flog, hard up
  BZ - BUSINESS AND COMMERCE
       1. Technical business terms eg consultant, tender
       2. Industrial terms eg organisational, manufacturing, business
       3. Meetings, conferences
     ACC - ACCOUNTING AND BOOKKEEPING
           1. More technical, business words (informal in F) eg books,
              expenses, receipt
           2. Tech words from accounts eg balance sheet, P and L
              account, cashflow etc
     LAB - STAFF AND THE WORKFORCE  (INC LABOUR RELATIONS)
           1. 'People' aspect of work
           2. Internal work hierarchy eg grades, ranks, titles,
              foreman, director
           3. Interaction, industrial relations eg dismissal,
              employment
           4. Interview terms
     MAR - MARKETING AND MERCHANDISING
           1. Buying, selling,
           2. Distribution, logistics eg freight, cargo
         MERG - Mergers, Monopolies, Takeovers, Joint Ventures
         PUBL - Public Relations
         SPON - Sponsorship
         RETA - Retailing
                1. Shops (double code with subj area of the business
                   eg greengrocer [FO])
                2. Shopping
     OFF - OFFICE PRACTICES AND EQUIPMENT
            1. See also PAPP Paper and Stationery
     STP - POSTAL SYSTEM  AND POSTAL TERMS (EXCL STATIONERY [PAPP])
           1. inc philately
  EC - ECONOMICS AND FINANCE
     CUR - CURRENCIES AND MODERN COINS (MONETARY UNITS)
           1. Currencies eg yen franc cent and incl old ones eg penny,
              guinea
           2. Physical coins and production of physical currency eg
              coin, mint, note
```

```
    DEV - ECONOMIC DEVELOPMENT & GROWTH
    ECY - ECONOMIC CONDITIONS & FORECASTS
    FIN - COMPANY FINANCE - BORROWING, SHARES ISSUES, BANKRUPTCIES
    INF - INFLATION, PRICES & COSTS
    PEN - PENSIONS
    PFE - BANKING AND PERSONAL FINANCE
          1. Banking terms eg credit, cheque, withdraw
          2. Building societies, savings accounts etc
          3. Personal income
          4. International banking
    TRA - INTERNATIONAL TRADE - IMPORTS AND EXPORTS
 IN - INSURANCE
 IV - INVESTMENT AND STOCK EXCHANGE
    MOR - MORTGAGES AND REAL ESTATE
 MF - MANUFACTURING
    DYE - PAINTS, DYES, AND PIGMENTS
    FOU - FOUNDRY (CASTING)
    GLA - GLASS
    HOR - WATCHES AND CLOCKS
          1. Manufacture and repair
          2. Parts
    PLC - PLASTICS
    PLT - PLATING
    RUB - RUBBER
    SME - SMELTING
    TAN - TANNING AND LEATHER
    TEX - TEXTILES
 TA - TAXATION
    ITX - INCOME TAX
    PTA - LOCAL GOVERNMENT TAXATION: POLL TAX, COUNCIL TAX
    VAT - VALUE ADDED TAX, PURCHASE TAXES, EXCISE DUTIES
```

Selectional preference In order for any sentence or phrase to be well-formed, both syntactic and semantic criteria must be met. Certain of these criteria are imposed internally, by the words within the sentence. For example, the verb put must always be followed by both an object and a prepositional or adverbial phrase, hence the ill-formedness of: **John put the book.*

Syntactic restrictions imposed by words as verbs and adjectives have been well researched and defined. What are less easily classifiable are the semantic restrictions which content words place on those words with which they interact in a sentence. Consider the verb *assassinate*: in order for the verb's sense to be complete, certain things must be true of its subject and object. The subject must be human; the object must also be human, and must furthermore be a person of great political or social importance (one cannot assassinate one's mother-in-law, for example).

These semantic restrictions imposed by lexical items on the words with which they interact in a clause (what we shall call their arguments) are what we call selectional preferences.

Clearly, the vast majority of words carry selectional preferences. We have chosen to concentrate on what are arguably the most easily definable and classifiable of these: the restrictions which verbs, and also adjectives and prepositions, place on their noun arguments. We include as verb arguments their subjects and objects and the noun phrases involved in any prepositional phrase complements. Adjective arguments are the noun which the adjective modifies and, again, the noun phrase in any prepositional phrase attached to the verb, such as interested in. In all cases, we aim to restrict the argument to an easily identifiable subset of all the nouns in the language - what we term a selectional class.

Clearly, such information is intrinsic to the sense and use of the word and, for a learner of English, achieving a clear grasp of these restrictions is an important part of understanding the correct use of the language. Furthermore, the correct interpretation of these restrictions is a key factor in disambiguating a sentence, and, we hope, will be a significant resource in word sense disambiguation as well as other NLP tasks.

Our aims, therefore, in undertaking the encoding of selectional preferences are far- reaching in two important areas of our overall work. Firstly, to create an important tool for lexicographers for the refining and enhancement of definition text, and, secondly, to build a further source of information for the in-house sense tagger to use when assigning a CIDE sense to each word in the corpus.

The semantic classes used for selectional preference coding were as follows:

```
Anything
  Physical
     Animate
          Human (+ gods etc.)
          Group (+institutions)
          Animal (+dragons etc.)
          Plant
          Body Part (+plant part)
      Inanimate
          Object
               Device
               Instrument (measuring, musical etc.)
               Money
               Vehicle
               Container
          Places
               Building (all constructions)
          Substance
          Solid
          Liquid
               Drink
          Gas
          Food
          Clothing
          Energy
     Abstract
          Time
```

```
Measurement (+Amounts/levels)
Sound
Activity (actions, movements)
Quality (+ability, colours)
Sensations (+emotions, feelings)
Event
Communication (+Language, symbols, communicated things etc.)
     Text
Process
Belief (+knowledge, systems, science, disciplines, styles)
State (+conditions, diseases)
```

3.3.1 LE Uses

The CIDE+ subject domain coding could be useful in a number of NLP tasks, and has already been used in document summarisation ([San98]) and word sense disambiguation ([Har97]).

3.4 Cambridge International Dictionary of English

The Cambridge International Dictionary of English (published 1995) is a large corpus-based advanced learner's dictionary, the text of which is available as an SGML file. This file includes approximately 80,000 sense entries (including some undefined derived runons as senses) and 110,000 example sentences. A typical sense entry includes:

- definitions (written in a 2,000 word defining vocabulary);
- guide words (distinguishing different senses of the same word);
- grammatical coding;
- register and variety coding, and
- example sentences with highlighted collocates.

There are also a number of encoding schemes that did not appear in the printed book but which were entered with both NLP purposes and future lexicographic development in mind, including:

- subject domain coding (each sense/example classified into 900 categories hierarchically structured in 4 levels)
- selectional preference coding (each noun sense classified into 40 categories; these categories then used to classify the possible fillers for each verb/adjective/preposition sense/example)
- multi-word unit tagging (marking grammatical function and inflectability of each component word)

This data, developed partly in conjunction with Cambridge Language Services Ltd under the UK government DTI-SALT funded project Integrated Language Database, is available as part of an SGML file package termed CIDE+ (see http://www.cup.cam.ac.uk/elt/reference/data.htm).

Subject domain coding The initial purpose of the subject domain coding scheme was to enable entries to be selected for revision by subject specialists, thus increasing the accuracy of particularly the definitions and examples and also ensuring that the most important specialist words were included in the dictionary. This coding was done not just at the entry level but also at the example level, for instances where either the examples illustrated shades of meaning specific to a particular subject domain, or where the examples were in a particular subject domain context (not necessarily to do with the word being exemplified).

There are 900 subject code categories hierarchically structured in 4 levels. The top level is as follows:

```
A   Agriculture
B   Building & Civil Engineering
C   Arts & Crafts
D   Food & Drink
E   Earth Science (include Outer Space)
F   Finance & Business
G   Sports (include Games & Pastimes
H   History
I   Media & Publishing
J   Crime & the Law
K   Language & Literature
L   Life Science
M   Maths, Physics & Chemistry
N   Entertainment & Music
O   Medicine, Health & Psychology
P   Names of Plants & Animals
Q   Education
R   Religion
S   Society
T   Technology
U   Politics, Diplomacy & Government
V   Travel & Transport
W   War & the Military
X   General Language
Y   Mythology, Occult & Philosophy
Z   Clothes, Fashion & Beauticulture
```

As the coding is for subject domain rather than semantic similarity, the coding was inappropriate for abstract vocabulary which was typically encoded under the category X; this accounted for approximately 30% of the senses in CIDE+.

The fine detail of the coding can be illustrated by the following sub-section:

```
F - FINANCE AND BUSINESS
    1. Non-technical words (double code if necessary) eg fee, thrifty,
     pay, fare
    2. Informal financial terms eg tick, flog, hard up
  BZ - BUSINESS AND COMMERCE
       1. Technical business terms eg consultant, tender
```

```
  2. Industrial terms eg organisational, manufacturing, business
  3. Meetings, conferences
ACC - ACCOUNTING AND BOOKKEEPING
      1. More technical, business words (informal in F) eg books,
         expenses, receipt
      2. Tech words from accounts eg balance sheet, P and L
         account, cashflow etc
LAB - STAFF AND THE WORKFORCE  (INC LABOUR RELATIONS)
      1. 'People' aspect of work
      2. Internal work hierarchy eg grades, ranks, titles,
         foreman, director
      3. Interaction, industrial relations eg dismissal,
         employment
      4. Interview terms
MAR - MARKETING AND MERCHANDISING
      1. Buying, selling,
      2. Distribution, logistics eg freight, cargo
    MERG - Mergers, Monopolies, Takeovers, Joint Ventures
    PUBL - Public Relations
    SPON - Sponsorship
    RETA - Retailing
           1. Shops (double code with subj area of the business
              eg greengrocer [FO])
           2. Shopping
OFF - OFFICE PRACTICES AND EQUIPMENT
       1. See also PAPP Paper and Stationery
STP - POSTAL SYSTEM  AND POSTAL TERMS (EXCL STATIONERY [PAPP])
      1. inc philately
EC - ECONOMICS AND FINANCE
CUR - CURRENCIES AND MODERN COINS (MONETARY UNITS)
      1. Currencies eg yen franc cent and incl old ones eg penny,
         guinea
      2. Physical coins and production of physical currency eg
         coin, mint, note
DEV - ECONOMIC DEVELOPMENT & GROWTH
ECY - ECONOMIC CONDITIONS & FORECASTS
FIN - COMPANY FINANCE - BORROWING, SHARES ISSUES, BANKRUPTCIES
INF - INFLATION, PRICES & COSTS
PEN - PENSIONS
PFE - BANKING AND PERSONAL FINANCE
      1. Banking terms eg credit, cheque, withdraw
      2. Building societies, savings accounts etc
      3. Personal income
      4. International banking
TRA - INTERNATIONAL TRADE - IMPORTS AND EXPORTS
IN - INSURANCE
IV - INVESTMENT AND STOCK EXCHANGE
```

```
    MOR - MORTGAGES AND REAL ESTATE
  MF - MANUFACTURING
    DYE - PAINTS, DYES, AND PIGMENTS
    FOU - FOUNDRY (CASTING)
    GLA - GLASS
    HOR - WATCHES AND CLOCKS
          1. Manufacture and repair
          2. Parts
    PLC - PLASTICS
    PLT - PLATING
    RUB - RUBBER
    SME - SMELTING
    TAN - TANNING AND LEATHER
    TEX - TEXTILES
  TA - TAXATION
    ITX - INCOME TAX
    PTA - LOCAL GOVERNMENT TAXATION: POLL TAX, COUNCIL TAX
    VAT - VALUE ADDED TAX, PURCHASE TAXES, EXCISE DUTIES
```

Selectional preference In order for any sentence or phrase to be well-formed, both syntactic and semantic criteria must be met. Certain of these criteria are imposed internally, by the words within the sentence. For example, the verb put must always be followed by both an object and a prepositional or adverbial phrase, hence the ill-formedness of: **John put the book*.

Syntactic restrictions imposed by words as verbs and adjectives have been well researched and defined. What are less easily classifiable are the semantic restrictions which content words place on those words with which they interact in a sentence. Consider the verb *assassinate*: in order for the verb's sense to be complete, certain things must be true of its subject and object. The subject must be human; the object must also be human, and must furthermore be a person of great political or social importance (one cannot assassinate one's mother-in-law, for example).

These semantic restrictions imposed by lexical items on the words with which they interact in a clause (what we shall call their arguments) are what we call selectional preferences. Clearly, the vast majority of words carry selectional preferences. We have chosen to concentrate on what are arguably the most easily definable and classifiable of these: the restrictions which verbs, and also adjectives and prepositions, place on their noun arguments. We include as verb arguments their subjects and objects and the noun phrases involved in any prepositional phrase complements. Adjective arguments are the noun which the adjective modifies and, again, the noun phrase in any prepositional phrase attached to the verb, such as interested in. In all cases, we aim to restrict the argument to an easily identifiable subset of all the nouns in the language - what we term a selectional class.

Clearly, such information is intrinsic to the sense and use of the word and, for a learner of English, achieving a clear grasp of these restrictions is an important part of understanding the correct use of the language. Furthermore, the correct interpretation of these restrictions is a key factor in disambiguating a sentence, and, we hope, will be a significant resource in word sense disambiguation as well as other NLP tasks.

Our aims, therefore, in undertaking the encoding of selectional preferences are far- reaching in two important areas of our overall work. Firstly, to create an important tool for lexicographers for the refining and enhancement of definition text, and, secondly, to build a further source of information for the in-house sense tagger to use when assigning a CIDE sense to each word in the corpus.

The semantic classes used for selectional preference coding were as follows:

```
Anything
  Physical
     Animate
          Human (+ gods etc.)
          Group (+institutions)
          Animal (+dragons etc.)
          Plant
          Body Part (+plant part)
      Inanimate
          Object
               Device
               Instrument (measuring, musical etc.)
               Money
               Vehicle
               Container
          Places
               Building (all constructions)
          Substance
          Solid
          Liquid
               Drink
          Gas
          Food
          Clothing
          Energy
     Abstract
          Time
          Measurement (+Amounts/levels)
          Sound
          Activity (actions, movements)
          Quality (+ability, colours)
          Sensations (+emotions, feelings)
          Event
          Communication (+Language, symbols, communicated things etc.)
               Text
          Process
          Belief (+knowledge, systems, science, disciplines, styles)
          State (+conditions, diseases)
```

	All PoS	Nouns	Verbs	Adjectives	Adverbs	Other
Number of Entries	61050	41810	7641	9296	1169	1134
Number of Senses	67785	45446	9752	10184	1323	1080
Senses/Entry	1.11	1.09	1.28	1.10	1.13	0.95
Morpho-Syntax			Yes	Yes		
Synsets	Yes					
- Number of Synsets	34201	196303	5895	7280	821	572
Sense Indicators	No					
Semantic Network	No					
Semantic Features	No					
Multilingual Relations	NO					
Argument Structure	19082	7406	9739	1929	1	7
Domain Labels	Yes					
- Domain Types	95					
- Domain Tokens	85971	58297	8102	16823	756	2173
- Domains/Sense	1.27	1.28	0.83	1.65	0.43	2.01

Table 3.3: Numbers and figures for GLDB:NEO

3.4.1 LE Uses

The CIDE+ subject domain coding could be useful in a number of NLP tasks, and has already been used in document summarisation ([San98]) and word sense disambiguation ([Har97]).

3.5 GLDB - The Göteborg Lexical DataBase

3.5.1 Introduction

The work on the GLDB started in 1977 by professor Sture Allén and his research group at Språkdata (today the department of Swedish), Göteborg University, Sweden. The underlying linguistic model is the lemma-lexeme model where, in short, the lemma stands for the canonical form of a word and all its formal data, whilst the lexeme stands for the semantic division of a lemma into one or more senses. The GLDB has the advantage of covering the 'whole' language and is not just a testbase comprising a small subset. Two major printed Swedish monolingual dictionaries, [Sve86] and [Nat95] have been generated from the GLDB. Both are also available on CD-rom.

The numbers and figures in Table 3.3 describe a subset of the GLDB, named NEO, that has been utilized in the printed Swedish monolingual dictionaries, [Sve86] and [Nat95].

3.5.2 Description

The information in the database is centered around two main information blocks, the lemma and the lexeme. The lemma comprises formal data: technical stem, spelling variation, part of speech, inflection(s), pronunciation(s), stress, morpheme division, compound boundary and element, abbreviated form, verbal nouns (for verbs). The lexemes are in their turn divided into two main categories, a compulsory kernel sense and a non-compulsory set of one or more sub-senses, called the cycles. Both categories comprise the variables definition, definition extension, formal (mainly grammatical) comment, main comment, references, morphological examples and syntactical examples. The kernels are in addition marked with terminological

Freq	Code	Domain Type
8418	admin.	administration
1064	anat.	anatomy
781	arb.	work
76	arkeol.	archaeology
478	arkit.	architecture
13	astrol.	astrology
307	astron.	astronomy
44	bergsvet.	science of minining
343	biol.	biology
1033	bok.	the world of books
9	bokf—r.	bookkeeping
1803	bot.	botany
278	byggn.tekn.	building
135	dans.	dancing
166	databehandl	dataprocessing, computers
42	dipl.	diplomati
1911	ekon.	ekonomi
221	eltekn.	electrotechnology
375	fil.	philosophy
134	film.	film, cinematography
88	flygtekn.	aviation, aeronautics
195	form.	designing
188	foto.	photography

Freq	Code	Domain Type
772	fys.	physics
536	fysiol	physiology
263	frg.	colour terms
818	geogr.	geography
344	geol.	geology
11	geom.	geometry
247	handarb.	needlework,embroidery
513	handel.	commerce
478	heminr.	interior decoration
36	herald.	heraldry
126	historia.	history
835	hush.	housekeeping
358	hyg.	hygiene
275	instr.	instrument
353	jakt.	hunting
846	jordbr.	agriculture
1221	jur.	law
608	kem.	chemistry
1194	kld.	clothes
2548	kokk.	cookery
4859	komm.	communication
685	konstvet.	art
14	lantmt.	surveying
555	litt.vet.	literature

Table 3.4: Terminological domains in GLDB:NEO (part 1)

domain(s), and the cycles have additional information about area of usage and type of subsense.

The Aristotelian type of definition with genus proximum and differentia specifica focusing relevant semantic concepts of the kernel sense is the main source of semantic information. A great deal of the definitions are extended with additional semantic information of a non-kernel character which has a separate place in the database, the definition extension. For instance, selection restrictions of external and/or internal arguments of verbs and adjectives are often specified here. (The information on selection restrictions is not strictly formalized and therefore not included in the above table.)

Semantic relations like hyperonomy, cohyponomy, hyponomy, synonymy and semantic opposition are linked for a substantial number of the kernels and cycles

There are 95 types of terminological domains in the GLDB:NEO database that are linked to kernel senses. Their types and frequency are listed in Tables 3.4-3.5.

3.5.3 Comparison with Other Lexical Databases

The GLDB is a sense-oriented full scale lexical database. The explicit information on the semantic relations such as hyperonymy, synonymy, cohyponymy and semantic opposition are comparable to Wordnet1.5 but, in addition it has traditional information.

Freq	Code	Domain Type	Freq	Code	Domain Type
238	maskin.	mechanical engineering	248	skogsbr.	forestry
573	mat.	mathematics	511	slkt.	kinship, family
279	matrl.	material	498	sociol.	sociology
2635	med.	medicine	742	spel.	game, play
80	metallurg.	metallurgy	1765	sport.	sport
513	meteorol.	meteorology	1281	sprkvet.	linguistics
1739	mil.	military	77	statist.	statistics
160	mineral.	mineralogy	1741	tekn.	technology
387	m-med.	mass media	5	teol.	theology
1160	mus.	music	258	textil.	textiles
241	mtt	measure	2656	tid.	time
130	numism.	numismatics	1431	trafik.	traffic
183	optik.	optics	18	tryck.tekn.	printing technique
998	pedag.	pedagogy	95	trdg.	gardening
734	pol.	politics	608	utstr.	extension
5654	psykol.	psychology	456	verkt.	tools
248	radiotekn.	radio engineering	475	vetenskapl.	science
1702	relig.	religion	41	veter.	veterinary
1292	rum.	room, space	5745	yrk.	proffesion, people
316	sag.	the world of fairy-tale	2417	zool.	zoology
3968	samh.	society, community	445	mne.	matter, substance
706	scen.	dramatic art	147	allm. kult.	culture
290	serv.	service	182	allm. vrd.	valuation
1592	sj—.	navigation, shipping	1562	land.	countries, ethnic groups

Table 3.5: Terminological domains in GLDB:NEO (part 2)

3.5.4 Relation to Notions of Lexical Semantics

The GLDB is a valuable resource for building ontologies and semantic networks based on the lexical system of the Swedish language. The rich semantic content of GLDB is instrumental for such tasks, e.g. the explicit information on the terminological domains and semantic relations such as hyperonymy, synonymy, cohyponymy and semantic opposition as well as the retrievable information on the genus proximum and selection restictions in definitions and their extensions.

3.6 Wordnets

3.6.1 Introduction

WordNet1.5 is a generic lexical semantic network developed at Princeton University [Mil90a] structured around the notion of synsets. Each synset comprises one or more word senses with the same part of speech which are considered to be identical in meaning, further defined by a gloss, e.g.:

- Synset = file 2, data file 1
- Part of Speech = noun
- Gloss = a set of related records kept together.

A synset represents a concept and semantic relations are expressed mostly between concepts (except for antonymy and derivational relations). The relations are similar to the lexical semantic relations between word senses as described by [Cru86] (see §2.7).

EuroWordNet [Vos96] is a multilingual database containing several monolingual wordnets structured along the same lines as Princeton WordNet1.5: synsets with basic semantic relations. The languages covered in EuroWordNet are: English, Dutch, German, Spanish, French, Italian, Czech and Estonian. In addition to the relations between the synsets of the separate languages there is also an equivalence relation for each synset to the closest concept from an Inter-Lingual-Index (ILI). The ILI contains all WordNet1.5 synsets extended with any other concept needed to establish precise equivalence relation across synsets. Via the ILI it is possible to match synsets from one wordnet to another wordnet (including WordNet1.5). Such mapping may be useful for cross-language Information-Retrieval, for transfer of information and for comparing lexical semantic structures across wordnets.

3.6.2 The Princeton WordNet1.5

General information on the size and coverage of WordNet1.5 is given in Table 3.6, taken from [Die96b]:

The following relations are distinguished between synsets:

Synonyms: members of the synset which are equal or very close in meaning, e.g.
{man 1, adult male}

Antonyms: synsets which are opposite in meaning, e.g.
{man, adult male} ⇒ {woman, adult female}

	All PoS	Nouns	Verbs	Adjectives	Adverbs	Other
Number of Entries	126520	87642	14727	19101	5050	0
Number of Senses	168217	107484	25768	28762	6203	0
Senses/Entry	1.33	1.23	1.75	1.51	1.23	
Morpho-Syntax			Yes			
Synsets	yes					
- Number of Synsets	91591	60557	11363	16428	3243	0
- Synonyms/Synset	1.84	1.77	2.27	1.75	1.91	
Sense Indicators	Yes	yes	yes	yes	yes	
- Indicator Types	1	1	1	1	1	
- Indicator Tokens	76705	51253	8847	13460	3145	
- Indicators/Sense	0.46	0.48	0.34	0.47	0.51	
Semantic Network						
- Relation Types*	13	10	6	5	2	
- Relation Tokens	128313	80735	13321	30659	3598	
- Relations/Sense**	1.400935	1.333207	1.172314	1.866265	1.109467	
- Number of Tops	584	11	573			
Semantic Features	No					
Multilingual Relations	No					
Argument Structure	Yes					
- Semantic Roles	No					
Semantic Frames	Yes					
- Frame Types	35		35			
Selection Restrictions	Yes					
- Restriction Types	2		2			
Domain Labels	No					
Register Labels	No					

Table 3.6: Numbers and figures for WordNet1.5 (* the synonymy relation is included in the notion of synset and is not counted here; ** the relations/sense is here calculated for synsets, because most relations apply to the synsets as a whole)

Hyperonyms: synsets which are the more general class of a synset, e.g.
{man, adult male} ⇒ {male, male person}

Hyponyms: synsets which are particular kinds of a synset, e.g.

{weather, atmospheric condition, elements} ⇒ {cold weather, cold snap, cold wave, cold spell}
{fair weather, sunshine, temperateness}
{hot weather, heat wave, hot spell}

Holonyms: synsets which are whole of which a synset is a part.

[Part of] e.g.,
{flower, bloom, blossom} PART OF: {angiosperm, flowering plant}

[Member of] e.g.,
{homo, man, human being, human} MEMBER OF: {genus Homo}

[Substance of] e.g.,
{glass} SUBSTANCE OF: {glassware, glasswork}
{plate glass, sheet of glass}

Meronyms: synsets which are parts of another synset.

[Has Part] e.g.,

{flower, bloom, blossom} HAS PART: {stamen}
{pistil, gynoecium}
{carpel}
{ovary}
{floral leaf}
{perianth, floral envelope}

[Has Member] e.g.,
{womankind} HAS MEMBER: {womanhood, woman}

[Has Substance] {glassware, glasswork} HAS SUBSTANCE: {glass}

Entailments: synsets which are entailed by the synset, e.g.
{walk, go on foot, foot, leg it, hoof, hoof it}⇒ {step, take a step}

Causes: synsets which are caused by the synset, e.g.
{kill} ⇒ {die, pip out, decease, perish, go, exit, pass away, expire}

Value of: (adjectival) synsets which represent a value for a (nominal) target concept. e.g.
poor VALUE OF: {financial condition, economic condition}

Has Value: (nominal) synsets which have (adjectival) concept as values, e.g.
size ⇒ {large, big}

Also see: Related synsets, e.g.
{cold} Also See→ {cool, frozen}

Similar to: Peripheral or Satellite adjective synset linked to the most central (adjectival) synset, e.g.
{damp, dampish, moist} SIMILAR TO: {wet}

Relation	Nouns	Verbs	Adjectives	Adverbs
Antonym	1713	1025	3748	704
Hyponym	61123	10817	0	0
Mero-member	11472	0	0	0
Mero-sub	366	0	0	0
Mero-part	5695	0	0	0
Entailment	0	435	0	0
Cause	0	204	0	0
Also-See	0	840	2686	0
Value of	1713	0	636	0
Similar to	0	0	20050	0
Derived from	0	0	3539	2894
Total	82082	13321	30659	3598

Table 3.7: Numbers and figures for WordNet1.5

Derived from: Morphological derivation relation with a synset, e.g.
{coldly, in cold blood, without emotion} Derived from adj → {cold}

In the description of WordNet1.5 ([Fel90]), troponymy is discussed as a separate relation. It is restricted to verbs referring to specific manners of change. However, in the database is it is represented in the same way as hyponymy. In the description given here verb-hyponymy is thus eqquivalent to troponymy.

Table 3.7 gives the distribution of the relations for each Part of Speech in terms of the number of synsets.

The semantic network is distributed over different files representing some major semantic clusters per parts-of-speech:

- *Noun Files in WordNet1.5*

  noun.act

  nouns denoting acts or actions

  noun.animal nouns denoting animals

  noun.artifact nouns denoting man-made objects

  noun.attribute nouns denoting attributes of people and objects

  noun.body nouns denoting body parts

  noun.cognition nouns denoting cognitive processes and contents

  noun.communication nouns denoting communicative processes and contents

  noun.event nouns denoting natural events

  noun.feeling nouns denoting feelings and emotions

  noun.food nouns denoting foods and drinks

  noun.group nouns denoting groupings of people or objects

  noun.location nouns denoting spatial position

  noun.motive nouns denoting goals

noun.object nouns denoting natural objects (not man-made)

noun.person nouns denoting people

noun.phenomenon nouns denoting natural phenomena

noun.plant nouns denoting plants

noun.possession nouns denoting possession and transfer of possession

noun.process nouns denoting natural processes

noun.quantity nouns denoting quantities and units of measure

noun.relation nouns denoting relations between people or things or ideas

noun.shape nouns denoting two and three dimensional shapes

noun.state nouns denoting stable states of affairs

noun.substance nouns denoting substances

noun.time nouns denoting time and temporal relations

- *Verb Files in WordNet1.5*

verb.body verbs of grooming, dressing and bodily care

verb.change verbs of size, temperature change, intensifying, etc.

verb.cognition verbs of thinking, judging, analyzing, doubting

verb.communication verbs of telling, asking, ordering, singing

verb.competition verbs of fighting, athletic activities

verb.consumption verbs of eating and drinking

verb.contact verbs of touching, hitting, tying, digging

verb.creation verbs of sewing, baking, painting, performing

verb.emotion verbs of feeling

verb.motion verbs of walking, flying, swimming

verb.perception verbs of seeing, hearing, feeling

verb.possession verbs of buying, selling, owning

verb.social verbs of political and social activities and events

verb.stative verbs of being, having, spatial relations

verb.weather verbs of raining, snowing, thawing, thundering

Within each of these files there may be one or more synsets which have no hyperonym and therefore represent the tops of the network. In the case of nouns there are only 11 tops or unique- beginners, in the case of verbs 573 tops.

Noun Tops in WordNet1.5

entity something having concrete existence; living or nonliving

psychological feature a feature of the mental life of a living organism

abstraction a concept formed by extracting common features from examples

location, space a point or extent in space

shape, form the spatial arrangement of something as distinct from its substance

state the way something is with respect to its main attributes; “the current state of knowledge”; “his state of health”; “in a weak financial state”]

event something that happens at a given place and time

act, humanaction, humanactivity something that people do or cause to happen

group, grouping any number of entities (members) considered as a unit

possession anything owned or possessed

phenomenon any state or process known through the senses rather than by intuition or reasoning

Whereas semantic relations such as hyponymy and synonymy are strictly paradigmatic relations, other relations such as meronymy and cause can be seen as syntagmatic relations imposing a preference relation between word senses:

Meronymy the head of a lion

Cause she died because he killed her

Finally, provisional argument-frames are stored for verbs. These frames provide the constituent structure of the complementation of a verb, where —-s represents the verb and the left and right strings the complementation pattern:
Verb-frames in WordNet1.5

1. Something —-s
2. Somebody —-s
3. It is —-ing
4. Something is —-ing PP
5. Something —-s something Adjective/Noun
6. Something —-s Adjective/Noun
7. Somebody —-s Adjective
8. Somebody —-s something
9. Somebody —-s somebody
10. Something —-s somebody
11. Something —-s something

12. Something —-s to somebody
13. Somebody —-s on something
14. Somebody —-s somebody something
15. Somebody —-s something to somebody
16. Somebody —-s something from somebody
17. Somebody —-s somebody with something
18. Somebody —-s somebody of something
19. Somebody —-s something on somebody
20. Somebody —-s somebody PP
21. Somebody —-s something PP
22. Somebody —-s PP
23. Somebody's (body part) —-s
24. Somebody —-s somebody to INFINITIVE
25. Somebody —-s somebody INFINITIVE
26. Somebody —-s that CLAUSE
27. Somebody —-s to somebody
28. Somebody —-s to INFINITIVE
29. Somebody —-s whether INFINITIVE
30. Somebody —-s somebody into V-ing something
31. Somebody —-s something with something
32. Somebody —-s INFINITIVE
33. Somebody —-s VERB-ing
34. It —-s that CLAUSE
35. Something —-s INFINITIVE

The distinction between human (Somebody) and non-human (Something) fillers of the frame-slots represents a shallow type of selection restriction.

The data in WordNet1.5. is stored in two separate files for each part of speech. The data file contains all the information for the synsets, where a file-offset position identifies the synset in the file. In the next example, the synset for *entity* is given:

```
00002403 03 n 01 entity 0 013
~ 00002728 n 0000
~ 00003711 n 0000
~ 00004473 n 0000
~ 00009469 n 0000
~ 01958400 n 0000
~ 01959683 n 0000
~ 02985352 n 0000
~ 05650230 n 0000
~ 05650477 n 0000
~ 05763289 n 0000
~ 05763845 n 0000
~ 05764087 n 0000
~ 05764262 n 0000
| something having concrete existence; living or nonliving
```

The first line in this example starts with the file-offset number, which uniquely identifies a synset within a part-of-speech file. It is followed by a reference to the global semantic cluster (03 = noun.animal), the part-of-speech, the size of the synset, a verbal synset name, a sense number of the verbal synset name, and the number of relations. On the next lines the related synsets are given where the symbol ˜indicates the type of relation, which is followed by a file-offset identifying the target synset, its part-of-speech and a number code for relations holding between synset members only. The final line contains the gloss. Verbal synsets may have an additional number for the verb-frames attached to it. A separate index-file then contains a list of lemmas with references to the synsets in which they occur:

```
abacus n 2 1 @ 2 02038006 02037873
abandoned\_infant n 0 1 @ 1 06098838
abandoned\_person n 0 2 @ ~ 1 05912614
abandoned\_ship n 0 1 @ 1 02038160
abandonment n 0 2 @ ~ 3 00116137 00027945 00051049
```

Here, each word lemma is followed by a part-of-speech code, the polysemy rate (0 if only 1 sense), a number indicating the number of different relations a word has, a list of the relation types (@ ˜) and a number indicating the number of synsets. Finally, the actual synsets are listed as file-off-set positions.

3.6.3 EuroWordNet

As indicated in Table 3.8, the size of the wordnets in *EuroWordNet* will be (as it is still in development) between 15,000-30,000 synsets and 30,000-50,000 word senses per language. The vocabulary is limited to general language but some subvocabulary is included for demonstration purposes. The information is limited to nouns and verbs. Adjectives and adverbs are only included in so far they are related to the nouns and verbs. Since the wordnets are still under development we cannot complete quantitative data.

The data in EuroWordNet is divided into separate modules:

- The Language Modules

	All PoS
Number of Entries	20.000
Number of Senses	50.000
Senses/Entry	2.5
Morpho-Syntax	no
Synsets	
- Number of Synsets	30.000
- Synonyms/Synset	1.7
Sense Indicators	
- Indicator Types	
Semantic Network	yes
- Relation Types	46
- Number of Tops	1
Semantic Features	yes
- Feature Types	63
- Feature Tokens	1024
Multilingual Relations	yes
- Relation Types	17
Argument Structure	yes
- Semantic Roles	yes
- Role Types	8
Semantic Frames	no
Selection Restrictions	no
Domain Labels	yes
Register Labels	yes

Table 3.8: Numbers and figures for EuroWordNet

- The Language-independent Modules
  - The Top Ontology
  - The Domain Ontology
  - The Inter-Lingual-Index

The Language Modules

The following information is then stored for each synset in the language-specific wordnets (the Language Modules):

Part of Speech Noun, Verb, Adjective or Adverb

Synset Set of synonymous word meanings (synset members)

Language-internal relations to one or more target synsets

Language-external relations to one or more ILI-records

Each of the synset-members represents a word sense for which further information can be specified:

- Usage Labels
  - Register
  - Style
  - Dialect
  - Region
- Frequency
- Morpho-syntactic information
- Definition
- Gloss

Most of this information for the synset-members or variants is optional.

The language-internal relations The most basic semantic relations, such as synonymy, hyponymy and meronymy, have been taken over from WordNet1.5. Some relations have been added to capture less-clear cases of synonymy, to be able to relate equivalences across parts-of-speech (so-called XPOS- relations), to deal with meronymy-relations between events (SUBEVENT), and to express role-relations between nouns and verbs (ROLE/INVOLVED relations):

Paradigmatic relations in EuroWordNet

- Synonymy
  - Synset-membership
  - Near_synonym, e.g. *machine*, *apparatus*, *tool*, *instrument*
  - XPOS_near_synonym
  - adorn V XPOS_NEAR_SYNONYM adornment N

- Hyponymy

  Has_Hyperonym/ Has_Hyponym

  Has_XPOS_hyperonym/ Has_XPOS_hyponym

  arrivo HAS_XPOS_HYPERONYM andare

  andare HAS_XPOS_HYPONYM arrivo

- Antonymy

  Antonym

  Near_antonym

  XPOS_near_antonym,

  dead XPOS_near_antonym live

Syntagmatic relations in EuroWordNet

- Role/Involved-relations

  Role/Involved

  Role_Agent/Involved_Agent

  watch-dog ROLE_AGENT to guard

  Role_Patient/ Involved_Patient

  to teach INVOLVED_PATIENT learner

  Role_Instrument/ Involved_Instrument

  hammer ROLE_INSTRUMENT to hammer

  Role_Location/ Involved_Location

  school ROLE_LOCATION to teach

  Role_Direction/ Involved_Direction

  Role_Source_Direction/ Involved_Source_Direction

  to emigrate INVOLVED_SOURCE_DIRECTION one's country

  Role_Target_Direction/ Involved_Target_Direction

  rincasare(to go back home) INVOLVED_TARGET_DIRECTION casa (home)

- Be_In_State/ State_of

  'the poor' are 'poor' (noun @ adjective)

- Cause/Caused_by

  'to kill' causes 'to die'

- Meronymy

  Has_Meronym/ Has_Holonym

  Has_Mero_Part/ Has_Holo_Part

  a whole and its constituent parts (e.g., hand' - finger)

Has_Mero_Member/ Has_Holo_Member

a set and its members (e.g., fleet - ship)

Has_Mero_Portion/ Has_Holo_Portion

a whole and a portion of it (e.g., metal - ingot)

Has_Mero_Madeof/ Has_Holo_Madeof

a thing and the substance it is made-of (e.g., book - paper).

Has_Holo_Location/ Has_Holo_Location

a place and location included within it (e.g., desert - oasis)

- Has_Subevent/ Is_Subevent_of

  'to buy' has subevent 'to pay'

As indicated in the table, the language-internal relations can be differentiated into paradigmatic and syntagmatic relations. The syntagmatic relations can be seen as specification of a potential semantic context for a word, where especially the role-relations may coincide with grammatical contexts as well. Furthermore, relations can be augmented with specific features to differentiate the precise semantic implication expressed:

- Conjunction or disjunction of multiple relations of the same type

  airplane

  HAS_MERO_PART *door* conjunctive

  HAS_MERO_PART *engine* conjunctive

  door

  HAS_HOLO_PART *car* disjunctive

  HAS_HOLO_PART *room* disjunctive

  HAS_HOLO_PART *airplane* disjunctive

- Factivity of causal relations

  kill CAUSES *die* factive

  search CAUSES *find* non-factive

- Reverseness of relations

  paper-clip HAS_MERO_MADEOF *metal*

  metal HAS_HOLO_MADEOF *paper-clip* reversed

- Negation of implications expressed by relations

  monkey HAS_MERO_PART *tail*

  ape HAS_MERO_PART *tail* not

The language external relations The equivalence relations are used to link the language-specific synset to the Inter-Lingual-Index or ILI. The relations parallel the language-internal relations:

- EQ_Synonym
- EQ_Near_Synonym
- HAS_EQ_Hyperonym and HAS_EQ_Hyponym
- HAS_EQ_Holonym and HAS_EQ_Meronym
- EQ_Involved and EQ_Role
- EQ_Causes and EQ_Is_Caused_By
- EQ_HAS_Subevent and EQ_IS_Subevent_Of
- EQ_Be_In_State and EQ_Is_State_Of

Eq_synonym is the most important relation to encode direct equivalences. However, when there is no direct equivalence the synset is linked to the most informative and closest concept using one of the complex equivalence relations. Eq_near_synonym is used when a single synset links to multiple but very similar senses of the same target word (this may be the result inconsistent sense-differentiation across resources). Has_eq_hyperonym and has_eq_hyponym are typically used for gaps, when the closest target synsets are too narrow or too broad. The other relations are only used when the closest target concept cannot be related by one of the previous relations.

Below is an example of the Dutch synset (*aanraking*; *beroering*: touch as a Noun) in the EuroWordNet database import format:

```
0 WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "aanraking"
      3 SENSE 1
      3 DEFINITION "het aanraken"
        4 FEATURE "Register"
          5 FEATURE_VALUE "Usual"
      3 EXTERNAL_INFO
        4 CORPUS_ID 1
          5 FREQUENCY 1026
        4 SOURCE_ID 1
          5 NUMBER_KEY 1336
    2 LITERAL "beroering"
      3 SENSE 2
        4 FEATURE "Date"
          5 FEATURE_VALUE "Old-fashioned"
      3 EXTERNAL_INFO
        4 CORPUS_ID 1
```

```
        5 FREQUENCY 238
      4 SOURCE_ID 1
        5 NUMBER_KEY 401472
1 INTERNAL_LINKS
  2 RELATION "XPOS_NEAR_SYNONYM"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "v"
      4 LITERAL "aanraken"
        5 SENSE 1
    3 SOURCE_ID 1001
  2 RELATION "HAS_HYPERONYM"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "n"
      4 LITERAL "beweging"
        5 SENSE 1
    3 SOURCE_ID 1001
  2 RELATION "CAUSES"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "n"
      4 LITERAL "contact"
        5 SENSE 1
    3 SOURCE_ID 1001
  2 RELATION "XPOS_NEAR_SYNONYM"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "v"
      4 LITERAL "raken"
        5 SENSE 2
    3 SOURCE_ID 1001
1 EQ_LINKS
  2 EQ_RELATION "EQ_SYNONYM"
    3 TARGET_ILI
      4 PART_OF_SPEECH "n"
      4 FILE_OFFSET 69655
    3 SOURCE_ID 1002
```

The Inter-Lingual-Index

The ILI is not internally structured: no lexical semantic relations are expressed between the ILI-records. In this respect it should not be seen as a language-neutral ontology but only as a linking-index between wordnets[1]. The Inter-Lingual-Index is thus basically a list of ILI-records, with the only purpose to provide a matching across wordnets. In addition, it also provides access to the language-neutral modules by listing the Top Concepts and Domain labels that may apply to it.

Simple ILI-records contain the following information fields:

[1]Note that it is possible to derive the hierarchical structuring of any wordnet, including WordNet1.5, for a particular set of ILI-records using a so-called Projection function in the EuroWordNet database

- variants: word forms, sense number
- part-of-speech
- gloss
- 1 or more relations to Top Concepts
- 1 or more relations to Domains
- 1 or more relations to synsets in the language-specific concepts

Most information is optional. The Top Concepts and Domains linked to an ILI-record can be transferred to the synsets in the local wordnets that are linked to the same ILI-record, as is illustrated in the next schema:

```
ES wordnet language-specific-word-meaning
   |___
    ___> eq_synonym-> ILI-record -has_top_concept-> Top Concept
   |
IT wordnet: language-specific-word-meaning
```

In addition to the Simple ILI-records, there are Complex ILI-records which group closely related meanings. These groupings are based on systematic polysemy relations between meanings, such as specialization of more general meanings, metonymy (§2.7, 3.12.2) and diathesis alternations (§2.6.2). Complex ILI-records are needed to provide a better linking between the wordnets. Inconsistent sense-differentiation across resources often makes it very difficult to find exact equivalences across the resources. By linking different meaning realizations (e.g. *university* as a *building* and as the *institute*) to same complex ILI-records it is still possible to find the closely related meanings.

Below is an example of a complex ILI-record in which specific meanings of *car* are grouped by a new generalized meaning:

```
0 ILI_RECORD
  1 PART_OF_SPEECH "n"
  1 NEW_ILI_ID 1234
  1 GLOSS "a 4-wheeled vehicle"
  1 VARIANTS
    2 LITERAL "car"
      3 SENSE 2
    2 LITERAL "automobile"
      3 SENSE 1
  1 EQ_RELATION "eq_generalization"
    2 ILI_RECORD
      3 FILE_OFFSET 54321
    2 ILI_RECORD
      3 NEW_ILI_ID 9876
```

Here, *eq_generalization* expresses the relation that holds with two more specific ILI-records, identified by FILE_OFFSET and NEW_ILI_ID respectively. The former indicates that it originates from WordNet1.5, the latter that it has been added as a new concept in EuroWordNet. These sense-groupings apply cross-linguistically, although the lexicalization of these meanings can differ from language to language.

Top Ontology

The EuroWordNet top ontology contains 63 concepts It is developed to classify a set of so-called Base Concepts extracted from the Dutch, English, Spanish and Italian wordnets that are being developed. These Base Concepts have most relations and occupy high positions in the separate wordnets, as such making up the core of the semantic networks. The Base Concepts are specified in terms of WordNet1.5 synsets in the ILI.

The top-ontology incorporates the top-levels of WordNet1.5, ontologies developed in EC-projects Acquilex (BRA 3030, 7315), and Sift (LE-62030)[Vos96], Qualia-structure [Pus95a] and Aktions-Art distinctions [Ven67], [Ver72], [Ver89], [Pus91b], entity orders [Lyo77]. Furthermore, the ontology has been adapted to group the Base Concepts into coherent semantic clusters. The ontology combines notions described in §2.2, 2.7, and 2.5. Important characteristics of the ontology are:

- semantic distinctions applying to situations cut across the parts of speech: i.e. they apply to both nouns, verbs and adjectives. This is necessary because words from different parts of speech can be related in the language-specific wordnets via a xpos_synonymy relation, and the ILI-records can be related to any part-of-speech.
- the Top Concepts are hierarchically ordered by means of a subsumption relation but there can only be one super-type linked to each Top Concept: multiple inheritance between Top Concepts is not allowed.
- in addition to the subsumption relation Top Concepts can have an opposition-relation to indicate that certain distinctions are disjunct, whereas others may overlap.
- there may be multiple relations from ILI-records to Top Concepts. This means that the Base Concepts can be cross-classified in terms of multiple Top Concepts (as long as these have no opposition-relation between them): i.e. multiple inheritance from Top Concept to Base Concept is allowed.

The Top Concepts are more like semantic features than like common conceptual classes. We typically find Top Concepts for Living and for Part but we do not find a Top Concept Bodypart, even though this may be more appealing to a non-expert. BCs representing body parts are now cross-classified by two feature-like Top Concepts Living and Part. The main reason for this is that a more flexible system of features is needed to deal with the diversity of the Base Concepts.

The top-concepts are structured according to the hierarchy shown in Fig 3.2. Following [Lyo77], the first level the ontology is differentiated into 1stOrderEntity, 2ndOrderEntity, 3rdOrderEntity. According to Lyons, 1stOrderEntities are publicly observable individual persons, animals and more or less discrete physical objects and physical substances. They can be located at any point in time and in, what is at least psychologically, a three-dimensional space. The 2ndOrderEntities are events, processes, states-of-affairs or situations which can be

```
    ____________________Top______________________
   /                     |                        \
 1stOrderEntity       2ndOrderEntity      3rdOrderEntity
 ..Origin             ..SituationType
 ....Natural          ....Dynamic
 ......Living         ......BoundedEvent
 .......Plant         ......UnboundedEvent
 .......Human         ....Static
 .......Creature      ......Property
 ...........Animal    ......Relation
 ....Artifact         ..SituationComponent
 ..Form               ....Cause
 ....Substance        ......Agentive
 ......Solid          ......Phenomenal
 ......Liquid         ......Stimulating
 ......Gas            ....Communication
 ....Object           ....Condition
 ..Composition        ....Existence
 ....Part             ....Experience
 ....Group            ....Location
 ..Function           ....Manner
 ....Vehicle          ....Mental
 ....Symbol           ....Modal
 ......MoneySymbol    ....Physical
 ......LanguageSymbol ....Possession
 ......ImageSymbol    ....Purpose
 ....Software         ....Quantity
 ....Place            ....Social
 ....Occupation       ....Time
 ....Instrument       ....Usage
 ....Garment
 ....Furniture
 ....Covering
 ....Container
 ....Comestible
 ....Building
```

Figure 3.2: Hierachy of Top Concepts in EuroWordNet

located in time. Whereas 1stOrderEntities exist in time and space 2ndOrderEntities occur or take place, rather than exist. The 3rdOrderEntities are propositions, such as ideas, thoughts, theories, hypotheses, that exist outside space and time and which are unobservable. They function as objects of propositional attitudes, and they cannot be said to occur or be located either in space or time. Furthermore, they can be predicated as true or false rather than real, they can be asserted or denied, remembered or forgotten, they may be reasons but not causes.

List of Top Ontology concepts in EuroWordNet with definitions

Top all

1stOrderEntity Any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimensional space.

2ndOrderEntity Any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heart, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist; e.g. *continue, occur, apply.*

3rdOrderEntity An unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten, e.g. *idea, thought, information, theory, plan.*

Origin Considering the way concrete entities are created or come into existence.

Natural Anything produced by nature and physical forces as opposed to artifacts.

Living Anything living and dying including *objects, organic parts or tissue, bodily fluids; e.g. cells; skin; hair, organism, organs.*

Human e.g. *person, someone.*

Creature Imaginary creatures; e.g. *god, Faust, E.T.*.

Animal e.g. *animal, dog.*

Plant e.g. *plant, rice.*

Artifact Anything manufactured by people as natural.

Form Considering the shape of concrete entities, fixed as an object or a-morf as a substance

Substance all stuff without boundary or fixed shape, considered from a conceptual point of view not from a linguistic point of view; e.g. *mass, material, water, sand, air.*

Solid Substance which can fall, does not feel wet and you cannot inhale it; e.g. *stone, dust, plastic, ice, metal*

Liquid Substance which can fall, feels wet and can flow on the ground; e.g. *water, soup, rain.*

Gas Substance which cannot fall, you can inhale it and it floats above the ground; e.g. *air, ozone.*

Object Any conceptually-countable concrete entity with an outer limit; e.g. *book, car, person, brick.*

Composition Considering the composition of concrete entities in terms of parts, groups and larger constructs

Part Any concrete entity which is contained in an object, substance or a group; *head, juice, nose, limb, blood, finger, wheel, brick, door.*

Group Any concrete entity consisting of multiple discrete objects (either homogeneous or heterogeneous sets), typically *people, animals, vehicles; e.g. traffic, people, army, herd, fleet.*

Function Considering the purpose, role or main activity of a concrete entity. Typically it can be used for nouns that can refer to any substance, object which is involved in a certain way in some event or process; e.g. *remains, product, threat.*

Vehicle e.g. *car, ship, boat.*

Software e.g. *computer programs and databases.*

Representation Any concrete entity used for conveying a message; e.g. *traffic sign, word, money.*

Place Concrete entities functioning as the location for something else; e.g. *place, spot, centre, North, South.*

Occupation e.g. *doctor, researcher, journalist, manager.*

Instrument e.g. *tool, machine, weapon*

Garment e.g. *jacket, trousers, shawl*

Furniture e.g. *table, chair, lamp.*

Covering e.g. *skin, cloth, shield.*

Container e.g. *bag, tube, box.*

Comestible *food* and *drinks*, including substances, liquids and objects.

Building e.g. *house, hotel, church, office.*

MoneyRepresentation Physical Representations of value, or money; e.g. *share, coin.*

LanguageRepresentation Physical Representations conveyed in language (spoken, written or sign language); e.g. *text, word, utterance, sentence, poem.*

ImageRepresentation Physical Representations conveyed in a visual medium; e.g. *sign language, traffic sign, light signal.*

SituationType Considering the predicate-inherent Aktionsart properties of Situations: dynamicity and boundedness in time. Subclasses are disjoint, every 2ndOrderEntity has only 1 SituationType.

Static Situations (properties, relations and states) in which there is no transition from one eventuality or situation to another: non- dynamic; e.g. *state, property, be.*

Relation Static Situation which applies to a pair of concrete entities or abstract Situations, and which cannot exist by itself without either one of the involved entities; e.g. *relation, kinship, distance, space.*

Property Static Situation which applies to a single concrete entity or abstract Situation; e.g. *colour, speed, age, length, size, shape, weight.*

Dynamic Situations implying either a specific transition from a state to another (Bounded in time) or a continuous transition perceived as an ongoing temporally unbounded process; e.g. *event, act, action, become, happen, take place, process, habit, change, activity.*

UnboundedEvent Dynamic Situations occurring during a period of time and composed of a sequence of (micro-)changes of state, which are not perceived as relevant for characterizing the Situation as a whole; e.g. *grow, change, move around, live, breath, activity, hobby, sport, education, work, performance, fight, love, caring, management.*

BoundedEvent Dynamic Situations in which a specific transition from one Situation to another is implied; Bounded in time and directed to a result; e.g. *to do, to cause to change, to make, to create.*

SituationComponent Considering the conceptual components that play a role in Situations. Situations can be cross-classified by any number of Situation Components

Cause Situations involving causation of Situations (both Static and Dynamic); *result, effect, cause, prevent.*

Stimulating Situations in which something elicits or arouses a perception or provides the motivation for some event, e.g. sounds, such as *song, bang, beep, rattle, snore*, views, such as *smell, appetizing, motivation.*

Phenomenal Situations that occur in nature controlled or uncontrolled or considered as a force; e.g. *weather, chance.*

Agentive Situations in which a controlling agent causes a dynamic change; e.g. *to kill, to do; to act.*

Usage Situations in which something (an instrument, substance, time, effort, force, money) is or can be used; e.g. *to use, to spent, to represent, to mean, to be about, to operate, to fly, drive, run, eat , drink, consume.*

Time Situations in which duration or time plays a significant role; Static e.g. *yesterday, day, pass, long, period*, Dynamic e.g. *begin, end, last, continue.*

Social Situations related to society and social interaction of people: Static e.g. *employment, poor, rich*, Dynamic e.g. *work, management, recreation, religion, science.*

Quantity Situations involving quantity and measure ; Static e.g. *weight, heaviness, lightness*; changes of the quantity of first order entities; Dynamic e.g. *to lessen, increase, decrease.*

	Nouns	Verb	Total
1stOrderEntities	491		491
2ndOrderEntities	272	228	500
3rdOrderEntities	33		33
Total	796	228	1024

Table 3.9: Distribution of Base Concepts in EuroWordNet

Purpose Situations which are intended to have some effect.

Possession Situations involving possession; Static e.g. *have, possess, possession, contain, consist of, own*; Dynamic changes in possession, often to be combined which changes in location as well; e.g. *sell, buy, give, donate, steal, take, receive, send.*

Physical Situations involving perceptual and measurable properties of first order entities; either Static e.g. *health, a colour, a shape, a smell*; or Dynamic changes and perceptions of the physical properties of first order entities; e.g. *redden, thicken, widen, enlarge, crush, form, shape, fold, wrap, thicken, to see, hear, notice, smell.*

Modal Situations (only Static) involving the possibility or likelihood of other situations as actual situations; e.g. *abilities, power, force, strength.*

Mental Situations experienced in mind, including emotional and attitudinal situations; a mental state is changed; e.g. *invent, remember, learn, think, consider.*

Manner Situations in which the way or manner plays a role. This may be Manner incorporated in a dynamic situation, e.g. *ways of movement such as walk, swim, fly*, or the static Property itself: e.g. *manner, sloppy, strongly, way.*

Location Situations involving spatial relations; static e.g. *level, distance, separation, course, track, way, path*; something changes location, irrespective of the causation of the change; e.g. *move, put, fall, drop, drag, glide, fill, pour, empty, take out, enter.*

Experience Situations which involve an experiencer: either mental or perceptual through the senses.

Existence Situations involving the existence of objects and substances; Static states of existence e.g. *exist, be, be alive, life, live, death*; Dynamic changes in existence; e.g. *kill, produce, make, create, destroy, die, birth.*

Condition Situations involving an evaluative state of something: Static, e.g. *health, disease, success or Dynamic e.g. worsen, improve.*

Communication Situations involving communication, either Static, e.g. *be about* or Dynamic (Bounded and Unbounded); e.g. *speak, tell, listen, command, order, ask, state, statement, conversation, call.*

The Top Concepts have been applied to the 1024 Base Concepts, distributed as shown i Table 3.9. As suggested above the Base Concepts are typically classified in terms of several top concepts. Below are examples of top concept conjunctions for the Base Concepts (note that some classifications may seem odd because they apply to rather specific senses of words):

- 1stOrderEntities

  Container+Part+Solid+Living

  blood vessel; passage; tube; vas; vein

  Place+Part+Solid

  face; field; layer; parcel; space

  Place+Part+Solid+Natural

  dry land

  Place+Part+Liquid+Natural

  body of water

  Furniture+Object+Artifact

  article of furniture; chair; seat; table

  Furniture+Group+Artifact

  furnishings

- 2ndOrderEntities

  Experience + Stimulating + Dynamic+Condition (undifferentiated for Mental or Physical)

  Verbs: cause to feel unwell; cause pain

  Physical + Experience + SituationType (undifferentiated for Static/Dynamic)

  Verbs: look; feel; experience;

  Nouns: sense; sensation; perception;

  Mental + (BoundedEvent) Dynamic + Agentive

  Verbs: identify; form an opinion of; form a resolution about; decide; choose; understand; call back; ascertain; bump into; affirm; admit defeat

  Nouns: choice, selection

  Mental + Dynamic + Agentive

  Verbs: interpret; differentiate; devise; determine; cerebrate; analyze; arrange

  Nouns: higher cognitive process; cerebration; categorization; basic cognitive process; argumentation; abstract thought

  Mental + Experience + SituationType (undifferentiated for Static/Dynamic)

  Verbs: consider; desire; believe; experience

  Nouns: pleasance; motivation; humor; feeling; faith; emotion; disturbance; disposition; desire; attitude

  Relation+Physical+Location

  Verbs: go; be; stay in one place; adjoin

  Nouns: path;course; aim; blank space; degree; direction; spatial relation; elbow room; course; direction; distance; spacing; spatial property; space

- 3rdOrderEntities

  theory; idea; structure; evidence; procedure; doctrine; policy; data point; content; plan of action; concept; plan; communication; knowledge base; cognitive content; know-how; category; information; abstract; info;

To apply to many other meanings, a Base Concept necessarily has to be abstract and general. Because of that most Base Concepts only qualify for a few top concepts rather than for rich combinations.

3.6.4 Comparison with Other Lexical Databases

WordNet1.5 is fundamentally different from a traditional dictionary because the semantic information is mainly stored for synsets rather then for words or word senses. Synsets are considered as conceptual units, and the lexical index table gives a mapping of the words in a language on these units. In this respect, WordNet1.5 is also rather different from many NLP lexicons, which often use traditional sense-based units for storing semantic information. WordNet1.5 can best be characterized as somewhere in between a semantic network and a conceptual ontology. The synsets are conceptual units rather than lexical semantic units. The relations are better seen as semantic inferencing schemes than as lexicalization patterns. However, compared to conceptual ontologies such as CYC the coverage is large and the content is shallow. It further differs from formalized NLP lexicons and feature-based resources in that the network is completely relational: i.e. no relations are expressed to semantic values or features outside the synset system (although the references to the lexicographer's files, representing the global semantic clusters, can be seen as a form of shallow feature encoding). Finally, WordNet1.5 has a large coverage of entries and senses with limited information and limited supportive devices or formalization, which makes it similar to traditional machine readable dictionaries.

Obviously, EuroWordNet shows much resemblance with WordNet1.5. The main differences are the multilinguality, the top concepts and domain ontology. A more fundamental difference between WordNet1.5 and EuroWordNet is that the former includes many non-lexicalized and artificial classes in the hierarchy, whereas the wordnets in EuroWordNet are restricted to lexicalized units (words and expressions) of a language.

Compared to traditional bilingual dictionaries (§3.13) there are also some differences. The equivalence relations in EuroWordNet are encoded at the synset level rather than the word sense level, as is done in traditional bilingual dictionaries. This means that the equivalence relations abstract from stylistic, pragmatic and minor morpho-syntactic differences. Another difference is that the kind of equivalence is explicitly coded in EuroWordNet, which is often not clear in bilingual dictionaries. Furthermore, EuroWordNet combines monolingual with multilingual information, which is very useful from a translation or language-learning perspective.

Finally, EuroWordNet is different from AI-ontologies such as CYC or Sensus/Pangloss (§3.9) in that its focus is on the linguistically-motivated relations rather than the semantic inference schemes only. In this respect, it provides information on the exact semantic relation between the lexicalized words and expressions of languages, even though this may still be useful for making inferences as well. Nevertheless, the design of the database still makes it possible to relate the wordnets to other ontologies, which focus on the cognitive implications only. Such linking is facilitated by the EuroWordNet Top Ontology.

3.6.5 Relations to Notions of Lexical Semantics

WordNet1.5 is a relational and taxonomic semantic model. It incorporates information on lexicalizations patterns, semantic components and conceptual inferences.

Like WordNet1.5, EuroWordNet represents a taxonomic, psychological model of meaning as a semantic network (§2.7). Even stronger than WordNet1.5, it expresses the lexicalization patterns and the important semantic components of languages (§2.5, 2.6) and the mapping of meaning to cognitive concepts as described in Generative Models (§2.7). Although EuroWordNet does not contain specifications of the argument structures of verbs, it does include a system for encoding conceptual dependencies between concrete nouns and abstract nouns and verbs, in the form of semantic roles. When combined with syntactic frames, these roles could be used to derive richly annotated argument structures as described in §2.4. Finally, the EuroWordNet Top Ontology provides a basic classification which incorporates formal notions of lexical aspect (§2.2) and semantic components (§2.6) for verbs and higher-order-nouns and qualia-roles as defined in Generative models of the lexicon (§2.7).

3.6.6 LE Uses

With LDOCE, WordNet1.5 is one of the most widely used lexical resources. This is mainly due to its availability and its large coverage. Because only limited semantic information is given, the usage is limited to more shallow processing such as information-retrieval (Ric95, Sme95). Furthermore, WordNet1.5 is used for tasks in NLP, such as similarity measurement and semantic tagging (Kur94, LiA95, Fuj97, Res95, San97, Agi96), or automatic acquisition and information extraction (McC97, Gri92, Rib95, Cha97). More elaborate usage requires a further encoding of information or linking with other resources with for example syntacic information.

As EuroWordNet is still in development, there are no actual uses on which we can report. Within the project, the resource will be applied to cross-language text-retrieval, experiments on this usage are reported in [Gil97]. In addition, we foresee that it will be a very useful tool for language generation tasks or authoring tools, for machine-translation tools, language-learning tools and for summarizers.

3.7 Resources from MemoData

3.7.1 Introduction

The Memodata resources can be described as a package containing definition dictionary for French, Multilingual dictionaries for French, Italian, Spanish, English and German and an ontology for French. The French side has been more developed but there is on on-going work for resources on other languages as well. All the existing resources can be linked together. For example one can start with the French definition dictionary, link it to the French ontology, next, access the different translations corresponding to the different meanings. As for applications, the dictionary is meant for semantic disambiguation and to be used as a thesaurus, where you start from the idea and get to the words.

3.7.2 Description

Dictionnaire intégral is the core dictionary. It is meant to be used both for dictionary editing purpose and natural language processing. It exists in five languages: French (20.000 classes (links or sort of synonyms)) that give an ontological organization of around 180.000 words-meaning (130.000 words grammatical categories), English, Italian, Spanish and German (around 35.000 words for each languages roughly corresponding to 45.000 word meanings). They are 350 labels (level, origin, domain, ...) and more than 40.000 concepts that organize the meanings.

Dicologique It is the editorial version of the dictionary. It refers to the interface as well as to the content.

Lexidiom It is a tool meant for lexicographers to edit/ change/enrich the dictionary.

Semiographe It is the compiled version of the dictionary that can be used in applications. For instance a recent version of SEMIOGRAPH is used for disambiguation, genericity, indexing the French yellow pages. SEMIOGRAPH contains information and features that are not in DICOLOGIQUE.

How to integrate the different components . The multiligual dictionaries are part of LEXIDIOM, attached to the same concepts as the French equivalents and linked via a relation of the type "to be translated by" in that particular sense. There are two types of information in this dictionary.

- the conceptual network which associates words or expressions with sets of referents.
- the linguistic level (for choosing the translation for instance)

3.7.3 LE Uses

Semantic disambiguation. The Semiographe takes into account words context to choose between the different meanings of a word. For instance it can choose between the different meaning of the word "renard" :

A- Voilà! Jean a (obturé) le renard (du barrage)
B- Regarde Justine qui (porte) (a revtu) son magnifique renard
C- le (corbeau) et le renard
D- la France fait la (chasse) au renard pour lutter contre le développement de la (rage).

Where parentheses indicate :

- words that are used to perform the disambiguation task
- for each sentence containing two parentheses that one word is optional

The corresponding meanings are given as follows:

- A: fente accidentelle par où s'échappe l'eau d'une installation (barrage, coque de navire, écluse....) (This meaning is not the default one)
- B: pour fourrure de renard (not default, it is an ellispis)
- The C example has just one word in between parenthesis. This word is optional as well, because, at the end, "renard" has the default meaning.
- In the example D it is the animal again (fox) but with the feature "gibier".

One can also uses the dictionary the other way around and from an hyponym found its specific value. For instance one can ask *couper un arbre* (to cut a tree) and get back the answer *abattre*, or one can ask *personne qui vend des fleurs* (person who sells flowers) and get back *fleuriste* (florist). Another way of querying the system is to ask questions such as *je veux louer un véhicule volant* (I want to rent a flying machine), and get *location d'avion* (plane renting) as an answer.

3.8 EDR

3.8.1 Introduction

The EDR is an electronic dictionary for English and Japanese. It is composed of five types of dictionaries:

- the English and Japanese (monolingual) *Word Dictionaries*
- the English/Japanese and Japanese/English *Bilingual Dictionaries*
- the *Concept Dictionary*
- the English and Japanese *Co-occurrence Dictionaries*, and
- the *Technical Terminology Dictionaries.*

The major repository of semantic information is the *Concept Dictionary.*

Size The Japanese Word Dictionary contains 250,000 words, and the English Word Dictionary contains 190,000 words. The Japanese-English Bilingual Dictionary contains 230,000 words, and the English-Japanese Bilingual Dictionary contains 190,000 words. The Concept Dictionary contains information on the 400,000 concepts. The Japanese Co-occurrence Dictionary contains 900,000 phrases, and the English Co-occurrence Dictionary contains 460,000 phrases. The Japanese Corpus contains 220,000 sentences, and the English Corpus contains 160,000 sentences.

3.8.2 Description

The purpose of the *Concept Dictionary* is to describe, classify and interrelate the concepts that are referred to in the *Word Dictionaries*, the *Bilingual Dictionaries* and the *Co-occurrence Dictionaries.* It is composed of three separate dictionaries:

- the *Headconcept Dictionary* which gives a description of each concept in words, e.g.

```
<Record Number>                       CPH0314159
<Concept Identifier>                  3d0ecb
<Headconcept>
        <English Headconcept>         borrow
        <Japanese Headconcept>        ...
<Concept Explication>
        <English Concept Explication>  to use a person's property after
                                       promising to ...
        <Japanese Concept Explication> ...
<Management Information>
        <Management History Record>   Date of record update "93/04/26"
```

- the *Concept Classification Dictionary* with a classification of concepts in terms of hypo/hyperonymic relations

```
<Record Number>                       CPC0271828
<Super-concept Identifier>            4445bc [Concept identifier that
                                      indicates 'something written']
<Sub-concept Identifier>              4445a0 [Concept identifier that
                                      indicates 'piece of correspondence']
<Management Information>
        <Management History Record>   Date of record update "92/03/05"
```

- the *Concept Description Dictionary* with info regarding the relation between concepts, e.g.

```
<Record Number>                       CPT0577216
<Description Information>
        <Description Category>        E
        <Description Tupple>
                <Concept Identifier of Concept 1>
                                      3d0ecb [Concept identifier
                                      that indicates 'borrow']
                <Relation Label>      object
                <Concept Identifier of Concept 2>
                                      0e5097 [Concept identifier
                                      that indicates 'book']
                <Truth Value>         1
<Management Information>
        <Management History Record>   Date of record update "92/05/10"
```

The Concept Classification Dictionary In Concept Classification, multiple inheritance is allowed. Some intermediate concepts in the EDR IS-A hierarchy may not not directly represent a word sense. There are approximately 6,000 intermediate concepts provided in the Concept Classification Dictionary. The following shows the first and second level headings for the concept classification of basic words. The five categories on the first level are indicated below as 1, 2, 3, 4, and 5. An example of third level classification is provided for category 1-3 (independent body/group).

```
1       3aa911          subject whose behavior resembles that of a human
        1-1     30f6b0  human/person
        1-2     30f6bf  animal
        1-3     3aa912  independent body/group
                1-3-1   30f746  organization
                1-3-2   3cfacc  a collection/group of people
                1-3-3   3f960d  human race
                1-3-4   444614  meeting/conference
```

```
                1-3-5   3aa930  object/thing that moves independently

        1-4     4444b6  supernatural being

2       3d017c          matter
        2-1     444d86  thing (concrete object)
        2-2     444ab5  a matter
        2-3     444daa  name that discriminates one thing from another
        2-4     0e7faa  an objective subject

3       30f7e4          event/occurrence
        3-1     30f7e5  phenomenon
        3-2     30f83e  action/deed
        3-3     30f801  movement
        3-4     3f9856  change in state/change in condition
        3-5     3aa963  condition/state

4       30f751          location/locale/place
        4-1     3aa938  physical location/actual location/actual space
        4-2     30f753  position that is defined through the relation of several
                        things or objects
        4-3     30f767  area/territory/domain
        4-4     3f9651  part of something
        4-5     3f9658  direction
        4-6     444a9d  abstract location

5       30f776          time
        5-1     3f9882  point in time
        5-2     444dd2  a period of time that occurs at intervals
        5-3     444dd3  a period of time with a beginning and ending point
        5-4     30f77b  time/duration of time
        5-5     444dd4  a time measurement/time that is indicated in units
        5-6     4449e2  a period
        5-7     30f7d6  elapse of time (considered from a historical point of view)
```

The Concept Description Dictionary Semantic relations between a verbal and a nominal concept are described by means of the following *relations*:

```
AGENT   That which acts on its own volition and is the subject that brings about
        an action
        Ex. The father eats.
                AGENT  MAIN

OBJECT  That which is affected by an action or change
        Ex. (He/She) eats an apple.
                     MAIN    OBJECT
A-OBJECT
        That which has a particular attribute
        Ex. A tomato   is red.
              A-OBJECT    MAIN

IMPLEMENT
        That which is used in a voluntary action such as tools or other
        implements
        Ex. (I) cut  with a knife.
                MAIN        IMPLEMENT

MATERIAL
```

```
        That which is used to make up something
        Ex. (He/she) makes butter from milk.
                     MAIN  OBJECT      MATERIAL

SOURCE  Location from which an event or occurrence begins
        Ex. (I) come from Kyoto.
                MAIN      SOURCE

GOAL    Location from which an event or occurrence ends
        Ex. (I) go   to Tokyo.
                MAIN    GOAL

PLACE   Place (physical location) at which something occurs
        Ex. (I) play in the room.
                MAIN        PLACE

SCENE   Place (abstract location) at which something occurs
        Ex. (I) act  in a drama.
                MAIN      SCENE

BASIS   That which is used as the standard of comparison
        Ex. Roses are more beautiful than tulips.
        [ [MAIN more]
          [OBJECT [ [MAIN beautiful]
                    [A-OBJECT rose]]]
                    [BASIS [ [MAIN beautiful]
                             [A-OBJECT tulip]]]

MANNER  Way in which an action or change occurs
        Ex. (I) speak slowly.
                MAIN  MANNER

        Ex. (I) watch for 3 hours.
        [ [MAIN watch]
          [MANNER [ [MAIN hour]
                    [NUMBER 3]]]]

TIME    Time at which something occurrs
        Ex. (I) wake up at 8 o'clock.
        [ [MAIN wake up]
          [TIME [ [MAIN o'clock]
                  [MODIFIER 8]]]]

TIME-FROM
        Time at which something begins
        Ex. (I) work from 9 o'clock.
        [ [MAIN work]
          [TIME-FROM [ [MAIN o'clock]
                       [MODIFIER 9]]]]

TIME-TO Time at which something ends
        Ex. (I) work until 9 o'clock.
        [ [MAIN work]
          [TIME-TO [ [MAIN o'clock]
                     [MODIFIER 9]]]]

QUANTITY
```

```
        Amount (quantity) of a thing, action, or change
        Ex. (There are) 3 kgs of apples.
        [ [MAIN apple]
          [QUANTITY [ [MAIN kg]
                      [NUMBER 3]]]]

        Ex. (I) lost 3 kgs.
        [ [MAIN lose]
          [QUANTITY [ [MAIN kg]
                      [NUMBER 3]]]]

MODIFIER
        Modification
        Ex. the book on the desk
        [ [MAIN book]
          [MODIFIER [ [MAIN on]
                      [MODIFIER desk]]]]

NUMBER  Number
        Ex. 3      kgs
            NUMBER MAIN

AND     Coordination between concepts
        Ex. (I) go to Rome and Naples.
        [ [MAIN go]
          [GOAL [ [MAIN Naples]
                  [AND Rome]
                  [ATTRIBUTE focus]]]]

        Ex. The mountains are beautiful and the water is clear.
        [ [MAIN [ [MAIN clear]
                  [A-OBJECT water]]]
          [AND    [ [MAIN beautiful]
                    [A-OBJECT mountain]]]]

OR      Selection between concepts
        Ex. (I will) go to Rome or Naples.
        [ [MAIN go]
          [GOAL [ [MAIN Naples]
                  [OR Rome]
                  [ATTRIBUTE focus]]]]

        Ex. (I will) go to school or go to the library.
        [ [MAIN [ [MAIN go]
                  [GOAL library]]]
          [OR   [ [MAIN go]
                  [GOAL school]]]]

CONDITION
        Condition of an occurrence or fact
        Ex. (I) It rained so (I) went home.
        [ [MAIN [ [MAIN went]
                  [GOAL home]]]
          [CONDITION rain]]

PURPOSE Purpose or reason for an action or occurrence
        Ex. (I) go to see a movie.
```

```
        [ [MAIN go]
          [PURPOSE [ [MAIN see]
                     [OBJECT movie]]]]

COOCCURRENCE
        Simultaneous occurrence of events or actions
        Ex. (I) cried while I was going home.
        [ [MAIN cry]
          [COOCCURRENCE [ [MAIN go]
                          [GOAL home]]]]

SEQUENCE
        Sequential occurrence of events or actions
        Ex. (I) went to the library and borrowed a book.
        [ [MAIN borrow]
          [OBJECT book]
          [SEQUENCE [ [MAIN go]
                      [GOAL library]]]]

POSSESSOR
        Possession or ownership
        Ex. (my) father's  book
                 POSSESSOR MAIN

BENEFICIARY
        Beneficiary (receiver) of an action or occurrence
        Ex. (I) buy  a book for my father.
                MAIN               BENEFICIARY

UNIT    Unit
        Ex. (This costs) 500 yen per dozen.
        [ [MAIN yen]
          [NUMBER 500]
          [UNIT [ [MAIN dozen]
                  [NUMBER 1]]]]

FROM-TO Range of items specified
        Ex. the cities from Osaka to Tokyo
        [ [MAIN cities]
          [MODIFIER [ [MAIN Tokyo]
                      [FROM-TO Osaka]]]]
```

Two types of concept relations are distinguished: "E" and "I". E-relations are based on concepts which are actual word senses, e.g. "3d0ecb" and "0e5097" as concept identifiers for specific uses of the words *borrow, book.* I-relations are based on intermediate concepts which are not actual word senses, but may be superordinate of actual word senses or other intermediate concepts, such as the concept identifier "30f6ae" which is a superordinate for any concept describing a stationary object.

Paradigmatic information involving conceptual aspects of word knowledge, e.g. semantic frame information, is provided in other dictionaries as shown in the Cooccurrence Dictionary entry below.

```
<Record Number>                       ECC157145
<Headword Information>
        <Headphrase>                  eaten   @d-object   lunch
<Co-occurrence Constituent Information>
<Constituent #> <Morpheme> <Stem>  <POS>  <Idiom Flag> <Concept Info>
```

```
 1              eaten      eat     VERB   0            3bc6f0
 2              lunch      lunch   NOUN   0            3bec74
<Syntactic Tree Information>
        <Syntactic Sub-tree>
                <Governing Constituent> 1/eaten
                <Relator Constituent>   2/@d-object
                <Dependent Constituent> 2/lunch
<Semantic Information>
        <Semantic Sub-frame>
                <Concept of Governing Element>  1/3bc6f0/eaten
                <Semantic Relation>             object
                <Concept of Dependent Element>  1/3bec74/lunch
<Co-occurrence Situation Information>
        <Frequency>                     3;2;173;65
        <Example Sentence>              {003000002264/ have you (eaten) <lunch>}
<Management Information>
        <Management History Record>     DATE="95/3/31"
```

3.8.3 Comparison with Other Lexical Databases

The EDR is one of the most complete large-scale lexical databases available. It combines rich bilingual and monolingual dictionary information (e.g. including details about syntactic and semantic subcategorization) with a WordNet-style organization of word senses into a hierarchy of concepts (e.g. synsets). Unlike EuroWordNet (§3.6.2) and even more than WordNet1.5 (§3.6.3), the EDR concept hierarchy includes nodes (intermediate concepts) which do not correspond to actual word sense. In EDR these concepts explicitly marked, which is not the case in the WordNet1.5. The EDR also provides a detailed characterization of semantic relations (e.g. *agent, source, goal*) which is usually only found in experimental lexical such as those developed in ACQUILEX (§3.12.3), DELIS (§3.12.5) and EUROTRA (§3.11.1), or partially encoded as in EuroWordNet.

3.8.4 Relation to Notions of Lexical Semantics

As mentioned with reference to LLOCE, the combination of syntactic and semantic information given in the EDR is particularly well suited for addressing questions concerning the syntax-semantics interface. Another strong relation to lexical semantic notions concerns the use of semantic relations, although the classification adopted goes well beyond the characterization of event participant roles.

3.8.5 LE Uses

The EDR was specifically developed for advanced processing of natural language by computers and has been used in a variety of applications including Machine Translation, lexical acquisition and Word Disambiguation.

3.9 Higher Level Ontologies

3.9.1 Introduction

In the last several years a number of higher or upper level ontologies have become generally available to the knowledge representation and natural language research communities.

Representations of the sorts of things that exist in the world and relations between them are necessary for a variety of natural language understanding and generation tasks, including syntactic disambiguation (e.g. prepositional phrase attachment), coreference resolution (only compatible types of things can corefer), inference based on world knowledge for interpretation in context, and to serve as language-independent meaning representations for text generation and machine translation. While NL applications in different domains appear to require domain-specific conceptualisations there is some hope that a common upper level of domain-independent concepts and relations can be agreed: such a shared resource would greatly reduce the load on individual NL application developers to reinvent a model of the most general and abstract concepts underlying language and reasoning. A collectively refined resource should also benefit from increased comprehensiveness and accuracy.

This section reviews four current candidates for upper level ontologies: Cyc, Mikrokosmos, the Generalised Upper Model, and Sensus. These are by no means the only candidates, but give an indication of the work going on in this area, especially work of relevance to NLP (since ontologies are also being explored both purely theoretically, and with a view to application in areas other than NLP, such as simulation and modelling (e.g. in molecular biology) and knowledge sharing and reuse, not all work on ontologies of relevance here). Recent general review articles on ontologies are [Vic97] and [Noy97].

Ontologies are not lexical resources *per se*. They are generally regarded as conceptualisations *underlying* language, so that mappings from lexicons into ontologies need to be provided. One of the advantages of this is that ontologies can serve an interlingual role, providing the semantics for words from multiple languages. But there are murky philosophical waters here. And, there are practical problems for any attempt to evolve standards for lexical semantics: should such semantics be anchored in an underlying ontological framework? If so, which ? And would this presuppose arriving first at a standard for ontologies?

3.9.2 Cycorp

Cycorp (the inheritor of Cyc from MCC which ran the Cyc project for 10 years) has made public an upper ontology of approximately 3,000 terms, a small part of the full Cyc knowledge base ('many tens of thousands' more concepts), but one which they believe contains 'the most general concepts of human consensus reality' [Cyc97]. They have not made available most of the ('hundreds of thousands of' axioms relating concepts nor any of the domain-specific microtheories implemented in the Cyc KB. They have not made available the Cyc lexicon which contains over 14,000 English word roots with word class and subcategorization information plus their mappings into the KB, nor the other components of their NL system – a parser and semantic interpreter.

Each concept in the KB is represented as a Cyc constant, also called a term or unit. Each term has *isa* links to superclasses of which it is an instance, plus *genls* links to superclasses of which it is a subclass. Two of the most important Cyc classes are collections and relations (predicates and functions). In addition to *isa* and *genls* links, collections frequently also have have links to subsets (usually just illustrative examples in the published version). Associated with predicate terms in the hierarchy is information about the predicate's arity and the types of its arguments. There may also be links to more general and more specific predicates. Functions also have information about their argument and result types.

Here is the textual representation of two sample Cyc constants – a collection and a relation. Each has a heading, an English gloss, then one or more relational attributes indicating links to other constants in the KB.

#$Head-AnimalBodyPart
The collection of all heads of *#$Animal*s.

isa: *#$AnimalBodyPartType #$UniqueAnatomicalPartType*

genls: *#$AnimalBodyPart #$BiologicalLivingObject*

some subsets: *#$Head-Vertebrate*

#$hairColor <*#$Animal*> <*#$ExistingObjectType*> <*#$Color*>
(*#$hairColor* ANIMAL BODYPARTTYPE COLOR) means that the hair which the *#$Animal* ANIMAL has on its BODYPARTTYPE has the *#$Color* COLOR. E.g., (*#$hairColor #$SantaClaus #$Chin #$WhiteColor*). This is normally *#$Mammal* hair, but certain *#$Invertebrate*s also have hair.

isa: *#$TernaryPredicate #$TangibleObjectPredicate*

arg2Genl: *#$AnimalBodyPart*

The Cyc KB organised as a collection of lattices where the nodes in all the lattices are Cyc constants and the edges are various sorts of relation (*isa*, *genls*, *genlpred*).

3.9.3 Mikrokosmos

The *Mikrokosmos ontology* [Mik97, Mah95a, Mah95b] is part of the Mikrokosmos knowledge-based machine translation system currently under development at the Computer Research Laboratory, New Mexico State University. It is meant to provide a language-neutral repository of concepts in the world to assist in the process of deriving an interlingual text meaning representation for texts in a variety of input languages. It is derived from earlier work on the ONTOS ontology [Car90].

The ontology divides at the top level into **object**, **event**, and **property**. Nodes occurring beneath these divisions in the hierarchy constitute the concepts in the ontology and are represented as frames consisting of slots with facets and fillers. Concepts have slots for an NL definition, time-stamp, links to superordinate and subordinate concepts, and an arbitrary number of other other properties (local or inherited). These slots have *facets* each of which in turn has a *filler*. Facets capture such things as the permissible semantic types or ranges of values for the slot (**sem**), the actual value (**value**) if known, and default values **default**.

The Mikrokosmos web site puts the current size of the ontology at about 4500 concepts. The ontology is being acquired manually in conjunction with a lexicon acquisition team, and a set of guidelines have evolved for acquiring and placing concepts into the ontology.

3.9.4 The PENNMAN Upper Model

The *PENMAN upper model* originated in work done in natural language generation at ISI in the 1980's [Bat90]. It emerged as a general and reusable resource, supporting semantic classification at an abstract level that was task- and domain-independent. One of its key features was the methodology underlying its construction, according to which ontologies should be created by careful analysis of semantic distinctions as revealed through grammatical alternations in and across languages. The PENMAN upper model was written in LOOM a knowledge representation language developed at ISI.

The original PENMAN upper model was then merged with the KOMET German upper model [Hen93] to create a single unified upper model. This in turn has been further generalised through consideration of Italian and is now referred to as the Generalized Upper Model [Gum97, Bat94].

3.9.5 The Sensus ontology

The *Sensus ontology* (formerly known as the Pangloss ontology) is a freely available 'merged' ontology produced by the Information Sciences Institute (ISI), California [Sen97, Kni94, HovFC]. It is the result of merging:

- the PENMAN Upper Model
- the ONTOS ontology
- the LDOCE semantic categories for nouns
- WordNet
- the Harper-Collins Spanish-English Bilingual Dictionary

The topmost levels of the ontology (called the Ontology Base (OB)) consist of about 400 terms representing generalised distinctions necessary for linguistic processing modules (parser, analyser, generator). The OB is the result of manually merging the PENMAN upper model with ONTOS. The middle region of the ontology consists of about 50,000 concepts from WordNet. An automated merging of WordNet and LDOCE with manual verification was carried out and the result of this merging, given the earlier merging of OB and WordNet, is an ontology linked to a rich English lexicon. A final merge with the Harper-Collins Spanish-English Bilingual Dictionary links Spanish words into the ontology (one of the aims of the work is to support Spanish-English machine translation).

Little detail of the structure of the ontology or of individual entries is available in published form. The electronic source for the ontology consists of some various word and concept definition files. The OB files contain entries of the form:

```
(DEFCONCEPT ARTIFACT
  :DEFINITION "physical objects intentionally made by humans"
  :DIRECT-SUPERCLASS (INANIMATE-OBJECT))
```

and entries in WordNet derived files are of the form:

```
(DEFCONCEPT |tolerate|
  :DEFINITION " put up with something or somebody unpleasant  "
  :DIRECT-SUPERCLASS (|countenance,let|)
  :FRAMES ((8 0) (9 0))
  :WN-TYPE VERB.COGNITION
  )
```

It is not clear how the WordNet derived entries link into the OB.

Comments in the interface/KB access code suggest that much richer information is available including part-whole relations, instance and member relations, constraints on verbal arguments, etc. But none of this data appears to be in the public release data files.

3.9.6 Comparison with Other Lexical Databases

The ontologies described here are different from usual lexical resources in that they focus on knowledge from a non-lexical perspective. An exception is MikroKosmos, which has rich lexical resource linked to the ontology which come more close to the lexical resources discussed here. However, the distinction is not that clear-cut. We have seen that EDR and WordNet1.5 contain both lexicalized concepts and non-lexicalized concepts, and can thus partly be seen as language-neutral structures as well.

3.9.7 Relation to Notions of Lexical Semantics

The kind of semantics described in the higher-level ontologies comes closest to the taxonomic models described in §2.7. It is also closely related to the work of [Sch73, Sch75], which formed the basis for a non-linguistic approach to conceptual semantics. As a non-lexical approach (with the exception of MikroKosmos) the resources clearly do not relate to §2.5 and 2.4.

3.9.8 LE Users

One of the prime uses to which the *Cyc ontology* is to be put is natural language understanding: in particular the Cycorp Web pages refer to several enhanced information retrieval applications (see §4.2), including "knowledge-enhanced" searching of captioned information for image retrieval and information retrieval from the WWW, parts of which could be converted into Cyc's internal format and used to supplement Cyc itself. Another application is thesaurus management, whereby the extensive Cyc ontology is used to support "conceptual merging" of multiple (perhaps industry-specific) thesauri.

Mikrokosmos is primarily designed to support knowledge-based machine translation (KBMT – see §4.1) and is being used in Spanish-English-Japanese translation applications.

The *Penman upper model* and *the Generalised Upper Model* were originally designed to assist in natural language generation applications (see §4.5), though their authors believe these models have broader potential utility for NL systems.

Pangloss and Penman from which Sensus was derived were applications in machine translation and text generation respectively, and *Sensus* is intended to support applications in these areas (see §4.1 and 4.5 for discussions of machine translation and text generation).

3.10 Unified Medical Language System

3.10.1 Introduction

Unified Medical Language System (UMLS) is a set of knowledge sources developed by the US National Library of Medicine as experimental products. It consists of four sections: a metathesaurus, a semantic network, a specialist lexicon and an information sources map, and contains information about medical terms and their interrelationships.

3.10.2 Description

The Metathesaurus The Metathesaurus contains syntactic and semantic information about medical terms that appear in 38 controlled vocabularies and classifications, such as SNOMED, MeSh, and ICD. It is organised by concept, and contains over 330,000 concepts and 739,439 terms. It also contains syntactic variations of terms, represented as *strings*. The representation takes the form of three levels: a set of general concepts (represented by a code), a set of concept names (represented by another, related, code) and a set of strings (represented by another code and the lexical string itself). An illustrative example is given in Figure 3.3. Meanings and relationships are preserved from the source vocabularies, but some additional information is provided and new relationships between concepts and terms from different sources are established.

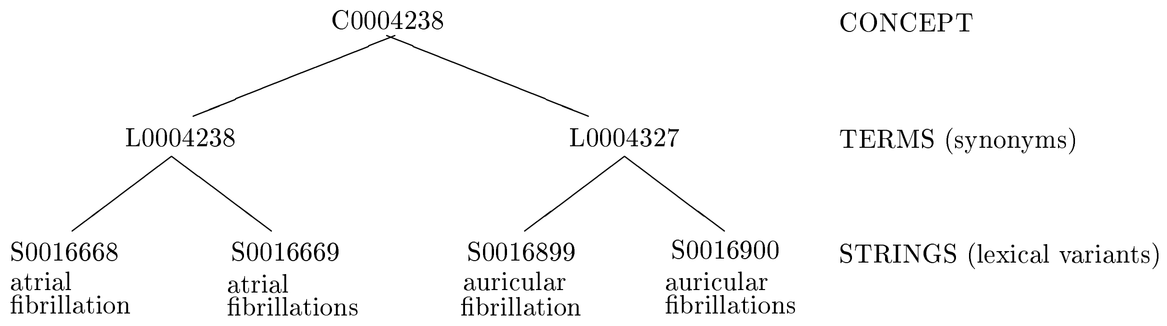


Figure 3.3: Fragment of Concept Hierarchy.

The relationships described between concepts in the metathesaurus are the following:

- X is broader than Y
- X is narrower than Y
- X and Y are "alike"
- X is a parent of Y
- X is a child of Y
- X is a sibling of Y
- X and Y have some other relation

An example record from the relational file is given below.

$$C0001430|CHD|C0022134|isa|MSH97|MTH$$

This indicates that there is an is-a relationship between the concept *nesidioblastoma* (C0001430) and the term *adenoma* (C0022134), that the former is a child of the latter, the source of the relationship comes from the MeSh subject headings (MSH97), and that this relationship was created specifically for the Metathesaurus (MTH).

The Semantic Network The semantic network contains information about the semantic types that are assigned to the concepts in the Metathesaurus. The types are defined explicitly by textual information and implicitly by means of the hierarchies represented. The semantic types are represented as nodes and the relationships between them as links. Relationships are established with the highest level possible. As a result, the classifications are very general rather than explicit ones between individual concepts.

In the semantic network, relations are stated between semantic types. The primary relation is that of hyponymy, but there are also five major categories of non-hierarchical relations:

- physical, e.g. *contains*
- spatial, e.g. *location of*
- functional, e.g. *prevents*
- temporal, e.g. *co-occurs with*
- conceptual, e.g. *diagnoses*

The following example shows how terms can be decomposed into their various concepts and positioned in the hierarchy.

D-33— Open Wounds of the Limbs
DD-33620 Open wound of knee without complication 891.0
(T-D9200)(M-14010)(G-C009)(F-01450)
DD-33621 Open wound of knee with complication 891.1
(T-D9200)(M-14010)(G-C008)(F-01450)

The Specialist Lexicon The specialist lexicon provides detailed syntactic information about biomedical terms and common English words. An individual lexical entry is created for each spelling variant and syntactic category for a word. These are then grouped together to form a unit record for each word, defined in a frame structure consisting of slots and fillers. Full morphological and syntactic information is provided, e.g. syntactic category, number, gender, tense, adjectival type, noun type, etc.

The Information Sources Map The information sources map contains details of the original sources of the terms. It consists of a database of records describing the information resources, with details such as scope, probability utility and access conditions. The sources themselves are varied and include bibliographic databases, factual databases and expert systems.

3.10.3 Comparison with Other Lexical Databases

SNOMED thus represents an implicit hierarchy of medical terms and their relationships by means of a coding system, enabling the identification of synonyms, hyponyms and hyperonyms. This makes it related to WordNet, although its coverage is very different. All the information in SNOMED is contained in UMLS, but represented in a more explicit tree-like structure. However, UMLS also includes information from a wide variety of other sources, and establishes relationships between these. UMLS further provides explicit morphological, syntactic and semantic information.

	All PoS	Nouns	Verbs	Adjectives	Adverbs	Other
Number of Entries	2193	941	444	465	269	yes
Number of Senses	2881	1322	740	396	270	yes
Morpho-Syntax	yes					
Synsets	no					
Sense Indicators	yes					
- Indicator Types	1					
Semantic Network	no					
Semantic Features	yes					
- Features	41					
Multilingual Relations	yes					
Argument Structure	yes					
Selection Restrictions	yes					
Domain Labels	no					
Register Labels	no					

Table 3.10: EUROTRA - Spanish

3.10.4 Relations to Notions of Lexical Semantics

As with the Higher-Level Ontologies discussed in the previous section, the structuring as synonyms, hyponyms and hyperonyms relate it to cognitive taxonomic models referred to in §2.7.

3.10.5 LE Uses

Both NLM and many other research groups and institutions are using UMLS in a variety of applications, including natural language processing, information extraction and retrieval, document classification, creation of medical data interfaces, etc. NLM itself uses it in several applications [UMLS97], including Internet Grateful Med, an assisted interactive retrieval system, SPECIALIST, an NLP system for processing biomedical information, and the NLM/AHCPR Large-Scale Vocabulary Test.

3.11 Lexicons for Machine-Translation

In this section we discuss several computerized lexicons that have been developed for Machine Translation applications: Eurotra, Cat-2, Metal, Logos and Systran (see §4.1). They have a high degree of formalization as compared to traditional dictionaries but the information is specifically structured to solve translation problems.

3.11.1 Eurotra Lexical Resources

Eurotra is a transfer based and syntax driven MT system which deals with 9 languages (Danish, Dutch, German, Greek, English, French, Italian, Spanish and Portuguese). Monolingual and Bilingual Lexical resources were developed for all the languages involved, size and coverage of those were similar for all.

We will only supply figures for Spanish as merely orientative in Table 3.10.

Eurotra dictionaries are organized according to a number of levels of representation: Eurotra Morphological Structure (EMS), Eurotra Constituent Structure (ECS), Eurotra Relational Structure (ERS) and Interface Structure (IS). The IS is the basis for transfer and although it reflects deep syntactic information it is also the level were semantic information is present.

The Eurotra IS level is an elaboration of dependency systems in that every phrase is made up of a governor optionally followed by dependants of two types: arguments and modifiers. Arguments of a given governor are encoded in the lexicon. The relations between governors and their arguments are not explicitly stated. The set of arguments, are:

```
arg1:       subject (experiencer/causer/agent)
arg2:       object (patient/theme/experiencer)
arg_2P:     2nd participant (goal/receiver (non-theme))
arg_2E:     2nd entity (goal/origin/place (non-theme))
arg_AS:     secondary stative predication on subject
arg_AO:     secondary stative predication on object
arg_Pe:     dative perceiver with raising predicates
arg_ORIGIN: oblique
arg_GOAL:   oblique
arg_MANNER: oblique
```

Not all labels have the same theoretical status nor correspond to the same level of depth in analysis. Thus,

- Sometimes ERS and IS functions express identical grammatical relations. This is the case of subject or object attribute (arg_AS and arg_AO respectively).
- Sometimes IS functions neutralize surface variation, so that different ERS syntactic functions go to the same IS argument. Thus arg2 includes NPs, VPs, SCOMPs and PPs (as in 'John wants bananas (NP)', 'John wants to go (VP)' 'John said that ... (SCOMP)' 'John believes in God (PP)' also for 'dative shift alternation' there is only one IS representation:

  John told Mary (arg2P) a story (arg2)

  John told a story (arg2) to Mary (arg2P)
- Sometimes IS roles establish more fine-grained distinctions than ERS functions. Thus, unergative and unaccusative verbs share the same ERS-frame but differ at IS where the subject is projected into arg1 for the former and into arg2 for the latter (eg, John (arg2) arrived). Also adjuncts are further semantically specified (origin, goal, manner, etc.)

Essentially the semantic information encoded in E-dictionaries is used for disambiguation purposes. These include (i) **structural ambiguity**, (ie. argument modifier distinction, specially in the case of PP-attachment) and (ii) lexical ambiguity in lexical transfer, that is **collocations** (restricted to verb support constructions), **homonymy** and **polysemy** (this is further explained in §5.3.4).

All information is encoded as Feature-Value pairs, in ASCII files. Here are some examples:

```
absoluto_1 =
{cat=adj,e_lu=absoluto,e_isrno='1',e_isframe=arg1,e_pformarg2=nil,term='0'}.
```

Information encoded depends on the category. For all categories, the category (cat=), lema (e_lu=) and reading number (e_isrno=) is encoded. Other information is, for nouns and verbs: deep syntactic argument structure (e_isframe) as explained above, argumental strongly bound prepositions required by the lexical item (e_pformargX), selectional restrictions for all the arguments (semargX=) and the semantic type of the lexical item (sem=).

Reading number refers to a meaning distinction usually also reflected in a difference in the encoding of the other atributes. In the case of *centro* ("center") the meaning distinction is referred to in the "sem" attribute" is: coordinate vs. place (*lug*). Besides, the reading "place" has no argumental structure while the reading "coord" can have one argument ("e_isframe=arg1"), and this has to be "concrete" in oposition to "abstract entity".

```
centro_1 =
{cat=n,e_lu=centro,e_isrno='1',e_gender=masc,person=third,nform=norm,
nclass=common,class=no,e_isframe=arg1,e_pformarg1=de,e_pformarg2=nil,
e_pformarg3=nil,sem=coord,semarg1=conc,semarg2=nil,semarg3=nil,
exig_mood=nil,e_predic=no,wh=no,whmor=none,e_morphsrce=simple,
term='2000000538'}.

centro_2 =
{cat=n,e_lu=centro,e_isrno='2',e_gender=masc,person=third,nform=norm,
nclass=common,class=no,e_isframe=arg0,e_pformarg1=nil,e_pformarg2=nil,
e_pformarg3=nil,sem=lug,semarg1=nil,semarg2=nil,semarg3=nil,
exig_mood=nil,e_predic=no,wh=no,whmor=none,e_morphsrce=simple,term='0'}.
```

Other strictly monolingual information encoded for nouns is: gender, person, type of noun ("nform" and "nclass"), if it requires a specific verbal mood (exig_mood) when creating a subordinate clause, if the noun is predicative ("e_predic"), information about relatives (wh and whmor), morphological derivative information (e_morphsrce, refers to morphological source, i.e, derivate...), and terminological identification: "term".

```
basar_1 =
{cat=v,e_lu=basar,e_isrno='1',e_isframe=arg1_2_PLACE,e_pformarg1=nil,
e_pformarg2=nil,e_pformarg3=en,e_pformarg4=nil,p1type=nil,p2type=nil,
semarg1=anim,semarg2=ent,semarg3=ent,semarg4=nil,e_vtype=main,
vfeat=nstat,term='0',erg=yes}.
```

As said before, information about strongly bound prepositions is encoded for all the arguments, and in case the verb preposition is weakly bound, 2 features corresponding to 2 possible complements might refer to a class of prepositions such as "origin", "goal", etc. As for nouns, selectional restrictions are encoded but no semantic typing of the verb itself. Specific monolingual information is encoded in the following attributes: "e_vtype", refers to the traditional *main* vs. *auxiliar* distinction, and "erg" refers to ergative verbs. Aspectual characterization of the verb is encoded in "vfeat", with possible values *stative*, *non stative*.

3.11.2 CAT-2 Lexical Resources

The CAT2 system, developed at IAI (Saarbruecken), is a direct descendant of Eurotra and was designed specifically for MT [Ste88], [Zel88], [Mes91]. The CAT2 system exploits linguistic

	All PoS
Number of Entries	German 20000
Number of Entries	English 30000
Number of Entries	French 30000
Number of Senses	German 40000
Number of Senses	English 50000
Number of Senses	French 50000
Senses/Entry	German 2
Senses/Entry	English 1.6
Senses/Entry	French 1.25
Morpho-Syntax	yes
Synsets	no
Sense Indicators	no
Semantic Network	no
Semantic Features	yes
- Feature Types	60
Multilingual Relations	yes
Argument Structure	yes
- Semantic Roles	7
Semantic Frames	no
Selection Restrictions	yes
Domain Labels	yes
- Domain Tokens	5000
Register Labels	yes
- Register Tokens	4

Table 3.11: CAT-2

information of different kinds: phrase structure, syntactic functional information and semantic information. The figures supplied in Table 3.11 provide an indication of size and coverage.

Semantic information is essentially used for reducing syntactic ambiguity, disambiguation of lexical entries, semantic interpretation of prepositional phrases, support verb constructions, lexical transfer and calculation of tense and aspect.

A verbal entry for the IS level example is:

```
apply1  =
%% He applied the formula to the problem.
{lex=apply,part=nil,VOW}\&
({slex=apply,head={VERB}}
;{slex=applying,head={VN_ING}}
;{slex=application,head={TION_R}}
;{slex=applicant,head={ANT_N}}
;{slex=applier,head={VN_AGENT}}
;{slex=application,head={TION_A}}
;{slex=unapplicable,head={UNABLE}}
;{slex=applicable,head={ABLE}}
;{slex=applicable,head={ELL_ABLE}}
;{slex=appliability,head={ABILITY}})\&
{sc={a={AGENT},b={THEME},c={GOAL,head={ehead={pf=to}}}},
 trans={de=({lex=applizieren};{lex=wenden,head={prf=an}}),fr={lex=appliquer}}}.
```

3.11.3 METAL Lexical Resources

Metal is a commercial MT system which is offered in English-German, English-Spanish, German-English, German-Spanish, Dutch-French, French-Dutch, French-English, German-French. It delivers monolingual and transfer system lexicons of up to 200,000 entries for language pair, as indicated in Table 3.12. Terms are coded for morphological, syntactic, and semantic patterns, including specification of selectional restrictions. Metal offers a sophisticated subject-area code hierarchy.

Argument Structure A verbal frame consists of a list of roles. A role consists of a role identifier and a description of its possible syntactic fillers. These are somewhat surface-oriented and therefore, role mapping in transfer is performed in an explicit way.

Possible role values are:

```
$SUBJ   deep subject
$DOBJ   deep object
$IOBJ   the affected
$POBJ   prepositional object
$SOBJ   sentential object
$SCOMP  attribute of subject
$OCOMP  attribute of object
$LOC    locative
$TMP    temporal
$MEA    measure
$MAN    manner
```

	All PoS
Number of Entries	200,000
Number of Senses	
Senses/Entry	
Morpho-Syntax	yes
Synsets	no
Sense Indicators	no
Semantic Network	no
Semantic Features	yes
- Feature Tokens	15/14
Multilingual Relations	yes
Argument Structure	yes
Selection Restrictions	yes
Domain Labels	yes
Register Labels	yes

Table 3.12: METAL

Lexical semantic features METAL has a restricted set of lexical semantic features which essentially deal with (un)definiteness of NPs, Tense and Aspect. Semantic relations are treated under the syntactic assignment of syntactic roles.

Adjectives, nouns and adverbs are semantically classified. Semantic features (attribute/value pairs) include:

- TA (Type of adjective): Age, Colour, Counting, degree, Directional, Indefinite, Locative, Manner, Measurement, origin, Equantial, Shape, Size and Temporal.
- TYN (Semantic type for nouns): Abstract, Animal, Concrete, Body part, Human, Location, Material, Measure, Plant, Potent, Process, Semiotic system, Social institution, Temporal, Unit of measure.

3.11.4 Logos Lexical Resources

Logos is a commercial high-end MT system which is offered in English-German, English-French, English-Spanish, English-Italian, English-Portuguese, German-English, German-French and German-Italian. Lexical Resources contain app. 50,000 entries for English source, 100,000 for German source, plus an additional semantic rule database with app. 15,000 rules for English source and 18,000 for German source — as indicated in Table 3.13.

Logos is based on semantic analysis techniques using structural networks. Logos encodes Logos semantic types which allow to define selectional restrictions based on syntactic patterns. Dictionaries are extendible (Logos standard dictionary comprises 250 thematic dictionaries), and the system supplies with an automatic lexicographic tool (Alex), and a semantic database (Semantha).

3.11.5 Systran Lexical Resources

Systran is a highly structured MT system whose translation process is based on repeated scanning of the terms in each sentence in order to establish acceptable relationships between

	All PoS
Number of Entries	English 50,000
Number of Entries	German 100,000
Morpho-Syntax	yes
Synsets	no
Sense Indicators	yes
Semantic Network	yes
Semantic Features	yes
Multilingual Relations	yes
Argument Structure	yes
- Semantic Roles	yes
Semantic Frames	yes
Selection Restrictions	yes
Domain Labels	yes
Register Labels	no

Table 3.13: LOGOS

forms. Using basic dictionaries, the system is able to define terms by analyzing morphemes (combining their grammatical, syntactic, semantic and prepositional composition).

It is a commercially available system offered with the following pairs:

- English into French, German, Italian, Spanish, Polish and Dutch.
- French into English, German, Italian and Dutch.

The figures supplied in Table 3.11.5 provide an indication of size and coverage.

Semantic Encoding A first semantic encoding is used for syntactic parsing. Thus, verbs are marked as motion or direction verbs, and in turn encode the "semanto-syntactic" nature of its complements, for instance:

- ABSUB = Verb normally takes an abstract subject
- ANSUB = Verb normally takes an animate subject
- COSUB = Verb normally takes a concrete subject
- HUSUB = Verb normally takes a human subject

Nouns are marked too. The inventory of labels is:

- CON = Conrete
- ABS = Abstract
- CT = Countable
- MS = Mass
- HU = Human

	All PoS
Number of Entries	English 95,000
Number of Entries	French 76,000
Number of Entries	German 135,000
Morpho-Syntax	yes
Synsets	no
Sense Indicators	no
Semantic Network	no
Semantic Features	yes
Multilingual Relations	yes
Argument Structure	yes
Domain Labels	yes
Register Labels	yes

Table 3.14: Systran

- QUAN = Quantity
- TP = Time Period
- AN = Animate
- AMB = Animate/inanimate ambiguity
- GRP = Collective Noun

Adverbs are also characterized semantically

- TI = Time
- PL = Place
- MA = Manner
- DEG = Degree
- FREQ = Frequency
- MODA = Modality
- DIR = Direction
- FUT = Future time

Besides, more semantic information is also encoded as part of a complex expression. It comprises two types: semantic primitives and terminology codes. The common attribute for both types is SEM.

Semantic primitives These semantic categories are defined to give information on the concept behind the word. They are source-language-bound, and may be applied to any part of speech. Their main function is the selection of a proper target meaning. Semantic primitives are taxonomized, that is: each code is included in a tree structure called taxon. There are 5 taxons:

`THINGS, PROCES, LOCATN, QUALITY, BEINGS`

Each of the taxons is the root of a tree which branches off to a number of subordinate nodes. For instance:

- THINGS: MATERL,PLANTS; INFORM; FDPROD;DEVICE
- MATERL: CMBUST, CHELEM, CHCOMP
- DEVICE: CONTNR, TRANSP

Terminology codes Terminology codes are defined to give information on the field a word is used in. They provide information of the source and correspond to the topical glossaries at target level. These codes are:

- ADMIN = administrative
- AGRIC = agriculture
- BACTL = bacteriology
- CHEMY = chemistry
- CONST = construction
- DPSCI = data processing
- ECLNG = CEC language
- ECONY = economy
- JURIS = juridical
- MEDIC = medical
- MIBIO = microbiology
- MILIT = military
- TECHY = technical
- TRADE = trade

Besides these terminological codes, subject field information is also supplied and is also related to the topical glossaries.

3.11.6 Comparison with Other Lexical Databases

From the figures mentioned above, it is obvious that Eurotra dictionaries cannot be considered a complete semantic database to be used in real-world applications. However they should be valued because of the information contained which was agreed for all the languages involved in the project. Furthermore, the rich syntactic specification may serve as a good basis for integrating syntactic and semantic information. This is illustrated by the CAT-2 dictionaries, which have become a large multilingual database which combine translational information encoding correspondences that goes beyond the word unit and morphosyntactic information.

3.11.7 Relation to Notions of Lexical Semantics

Eurotra IS representation is a deep syntactic representation and not a semantic one. However, in a number of areas, attempts were made to provide an interlingual semantic solution to the translation problem. The areas which have been singled out for semantic analysis were those in which a morpho-syntactic approach proved to be insufficient to cope with the translation problems. These areas were mainly:

- Tense and Aspect (§2.2)
- Mood and Modality
- Determination and Quantification (§2.7.4)
- Negation
- Aktionsart (§2.2)
- Lexical Semantic Features (§2.7, 2.5.2)
- Modification

Following Eurotra-D, CAT2 uses semantic relations as a basis for monolingual and bilingual disambiguation. In addition, the system suggests an extensive semantic encoding of nouns using hierarchical feature structures.

The semantic coding of nouns follows Cognitive Grammar principle [Zel88]. The semantic coding of argument roles follows Systemic Grammar [Ste88]. Support verb constructions follow the analysis of [Mes91].

3.12 Experimental NLP lexicons

3.12.1 Introduction

In this section several experimental computer lexicons are described, which have been developed for NLP applications in general. They try to encode rather sophisticated and complex lexical data, using complex representation formalisms, such as Types Feature Structures (TFS), and derivational mechanisms, such as lexical rules. Because of the complexity of the data, the coverage in entries and senses is mostly low, but the potential use of the rich data is very high.

3.12.2 the Core Lexical Engine

CoreLex is an ontology that implements the basic assumptions laid out in Generative Lexicon theory [Pus95a], primarily the view that *systematic polysemy* should be the basic principle in ontology design for lexical semantic processing. The idea for CoreLex itself originates out of an NSF-ARPA funded research project on the Core Lexical Engine, a joint research project of Brandeis University and Apple Computers Inc. [Pus95b]. Results of this research have been published in various theoretical and applied papers [Pus94b] [Joh95] [Pus96]. The research described in [Bui98], however, is the first comprehensive attempt to actually construct an ontology according to some of the ideas that arose out of this accumulated research and investigate its use in both classification and semantic tagging.

In CoreLex lexical items (currently only nouns) are assigned to systematic polysemous classes instead of being assigned a number of distinct senses. This assumption is fundamentally different from the design philosophies behind existing lexical semantic resources like WordNet that do not account for any regularities between senses[2]. A systematic polysemous class corresponds to an underspecified semantic type that entails a number of related senses, or rather *interpretations* that are to be generated within context. The underspecified semantic types are represented as qualia structures along the lines of Generative Lexicon theory.

Acknowledging the systematic nature of polysemy allows one to:

- structure semantic lexicons more efficiently
- give better explanations for non-literal use of language as in metaphor and metonymy
- build more robust natural language processing systems that can generate more appropriate interpretations within context

In order to achieve these goals one needs a thorough analysis of systematic polysemy on a large and useful scale. The CoreLex approach represents such an attempt, using readily available resources as WordNet and various corpora, to establish an ontology of 126 underspecified semantic types corresponding to 324 systematic polysemous classes that were derived from WordNet. The strategies for deriving such an ontology of systematic polysemous classes from WordNet can be summarized by the following three stages: 1. Reducing WordNet senses to a set of 'basic types'; 2. Organizing the basic types into systematic polysemous classes, that is, grouping together lexical items that share the same distribution of basic types; 3. Representing systematic polysemous classes through *underspecified semantic type* definitions, that extend into *qualia structure* representations.

Size and Coverage

Table 3.15 provides an overview of the size and coverage of CoreLex. Currently, only nouns are covered, although initial work on verbs and adjectives has started. The number of senses per entry is not an applicable feature for CoreLex, because its theoretical underpinning is exactly to give underspecified representations for polysemous words, instead of discrete senses. Homonyms in CoreLex would still have such discrete senses, but they are currently not considered. Underspecified semantic types are represented by use of qualia structure [Pus95a].

[2] WordNet 1.5 in fact includes an experiment in representing related senses, called *cousins*. The scope of

	Nouns
Number of Entries	39,937
Number of Senses	126
Senses/Entry	n/a
Semantic Network	
- Number of Tops	39
- Semantic Features	Yes
- Feature Types	Qualia

Table 3.15: Numbers and figures for nouns in CoreLex

Representation Formalism

CoreLex is implemented as a flat ASCII database of three tables that can easily be turned into a relational database, for instance using the PERL programming language. The three tables relate nouns with underspecified semantic types (table 1), underspecified semantic types with systematic polysemous classes (table 2) and systematic polysemous classes with corresponding basic types (table 3).

Top Hierarchy

The hierarchy displayed in Figure 3.4 is used in deriving CoreLex. It extends the 11 WordNet top types with 28 further ones on several sublevels. Together they constitute a set of 39 *basic types*[3]. Figure 3.5 lists each of them together with their frequency in the nouns database, that is, how many noun instances there are for each type. Basic types marked by '*' are *residual*, meaning their frequencies include only those nouns that are strictly only of that (super)type and do not belong to any of its subtypes. For instance, ABS* has 8 instances that are defined by WordNet as belonging only to the supertype **abstraction** and which are not further specified into a subtype like **definite_quantity** or **linear_measure**.

An Example

As an illustration of the interaction between basic types, systematic polysemous classes and CoreLex underspecified semantic types, consider the type `'acr'` in Figure 3.12.2, which corresponds to the following four different polysemous classes. The underspecified semantic type `'acr'` binds together four basic types: **act**, **event**, **relation**, **state**. For example, in the following sentences, that were taken from the BROWN corpus, 'delicate transformation' and 'proper coordination' address simultaneously the **event** and the **act** of transforming/coordinating an object R_1, the transforming/coordinating **relation** between two objects R_2 and R_3 and the **state** of this transformation/coordination itself.

> Soon they all were removed to Central Laboratory School where their delicate transformation began.

this experiment is however small and its results are very preliminary

[3]The basic type `'tme'` (**time**) occurs twice, grouping together two subdomains of WordNet as one in CoreLex.

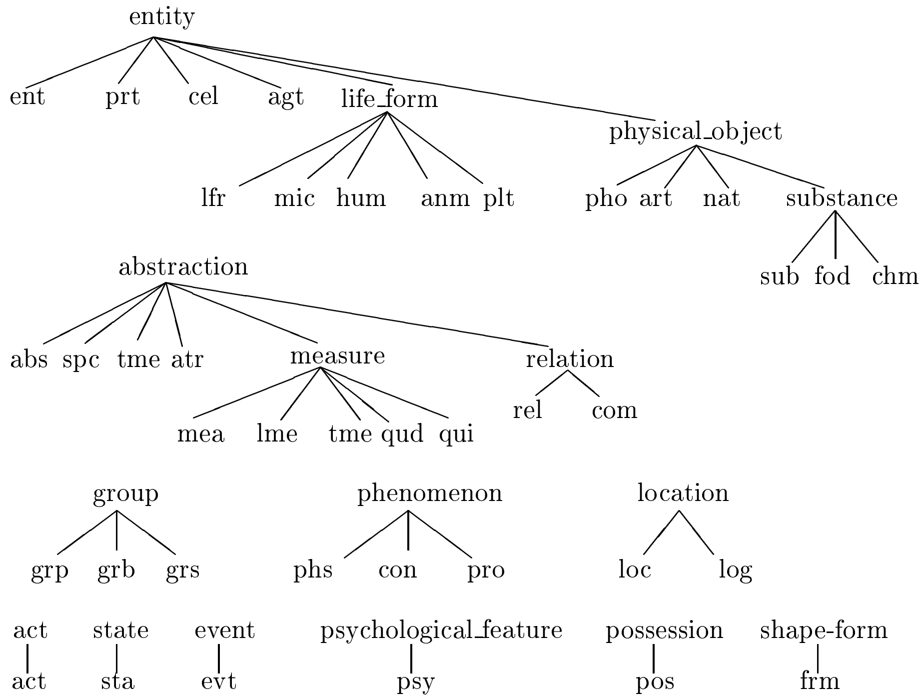


Figure 3.4: The hierarchies of basic types in WORDNET

> Revise and complete wildlife habitat management and improvement plans for all administrative units, assuring proper coordination between wildlife habitat management and other resources.

Since WORDNET was not developed with an underlying methodology for distinguishing between different forms of ambiguity, often CORELEX classes may include lexical items that do not directly belong there. This requires further structuring using a set of theoretically informed heuristics involving corpus studies and lexical semantic analysis [Bui97].

3.12.3 Acquilex

Research undertaken within the Acquilex projects (Acquilex-I, Esprit BRA 3030, and Acquilex-II, Esprit Project 7315) mainly aimed at the development of methodologies and tools for the extraction and acquisition of lexical knowledge from both mono- and bi-lingual machine readable dictionaries (MRDs) of various European languages (Dutch, English, Italian and Spanish). Within Acquilex-II, a further source of information was taken into account to supplement the information acquired from dictionaries: substantial textual corpora were explored to acquire information on actual usage of words. The final goal of the research was the construction of a prototype integrated multilingual Lexical Knowledge Base for NLP applications, where information extracted from different kinds of sources and for different languages was merged.

Acquilex did not aim at developing broad coverage lexical resources. The focus was on establishing a common and theoretically sound background to a number of related areas of research. Hence, in the specific case of this project, it makes more sense to consider those

Type	Corresponding Synsets	Freq
ART	artifact, artefact	8283
ACT	act, human_action, human_activity	6606
HUM	person, individual, someone, mortal, human, soul	6303
GRB	biological_group	4933
ATR	attribute	4137
PSY	psychological_feature	3456
COM	communication	3336
ANM	animal, animate_being, beast, brute, creature, fauna	2703
PLT	plant, flora, plant_life	2311
STA	state	2266
FOD	food, nutrient	1541
LOG	region 1 (geographical location)	1282
NAT	natural_object - water, - land,	1277
SUB*	substance, matter	1189
EVT	event	1082
PRT	part, piece	992
GRS	social_group - people	940
QUD	definite_quantity	777
PRO	process	773
CHM	compound, chemical_compound - chemical_element, element	699

Type	Corresponding Synsets	Freq
TME	time_period, period, period_of_time, amount_of_time - time_unit, unit_of_time - time	628
AGT	causal_agent, cause, causal_agency	624
POS	possession	571
LOC*	location, (any other location)	567
REL*	relation	506
FRM	shape, form	420
GRP*	group, grouping (any other group)	345
PHM*	phenomenon	342
QUI	indefinite_quantity	295
PHO*	object, inanimate_object, physical_object	186
MIC	microorganism	178
LME	linear_measure, long_measure	100
LFR*	life_form, organism, being, living_thing	61
CEL	cell	57
MEA*	measure, quantity, amount, quantum	38
ENT*	entity	28
CON	consequence, effect, outcome, result, upshot	21
SPC	space	21
ABS*	abstraction	8

Figure 3.5: Basic types in WordNet and their frequencies

information types which were extracted and/or formalised from different sources (see the following section) rather than giving detailed figures of encoded data.

Work carried out Work on dictionaries within the Acquilex project was divided into two consecutive steps:

1. development of methodologies and techniques and subsequent construction of software tools to extract information from MRDs and organise it into lexical databases (LDBs);

2. construction of theoretically-motivated LKB fragments from LDBs using software tools designed to integrate, enrich and formalise the database information.

Tools were constructed for recognising various kinds of semantic relations within the definition text. Starting from the *genus* part of the definition, hyponymy, synonymy and meronymy relations were automatically extracted and encoded within the mono-lingual LDBs. Also the *differentia* part of the definition was exploited, though to a lesser extent (i.e. only some verb and noun subclasses were considered), to extract a wide range of semantic attributes and relations characterising a word being defined with respect to the general semantic class it belongs to. Typical examples of relations of this class are: Made_of, Colour, Size, Shape and so forth; information on nouns *Qualia Structure*, e.g. Telic ([Bog90]; information on verb causativity/inchoativity; information on verb meaning components and lexical aspect; information on verb typical subjects/objects (cf. various Acquilex papers). Lexical information (such as collocations, selectional preferences, subcategorization patterns as well as near-synonyms and hyperonyms) was also extracted - although partially - from example sententes and semantic indicators (the latter used in bilingual dictionaries as constraints on the translation).

`act atr rel`	acceleration comparison
`act evt rel`	blend competition contraction transformation
`act rel`	abbreviation dealings designation discourse gait glide likening negation neologism neology prevention qualifying reciprocity sharing synchronisation synchronization synchronizing
`act rel sta`	coordination gradation inclusion involvement

Figure 3.6: Polysemous classes and instances for the type: '`acr`'

When considered in isolation, MRDs are useful but insufficient information sources for the construction of lexical resources, since it is often the case that knowledge derived from them can be unreliable and is in general asystematic. Hence, the use of corpora as additional sources of lexical information represented an important extension of the last phase of the project. Corpus analysis tools were developed with a view to (semi-)automatic acquisition of linguistic information: for instance, tools for part of speech tagging, or derivation of collocations or phrasal parsing. However, the real lexical acquisition task was not an accomplishment of this project which mainly focussed on the preparatory phase (i.e. development of tools).

The mono-lingual LDBs Most of the acquired data were loaded into the LDBs developed starting from MRDs at each site: some information (in particular taxonomical information) is now available extensively for all sources; other semantic relations were extracted and encoded only in relation to some lexico-semantic subclasses of words (e.g., Food and Drinks, motion and psychological verbs, etc.). The LDBs were given a structure which tries to preserve all the information extractable from the dictionary source, while expressing explicitly also structural relationships and leaving open the possibility to add new data, instead of having all the relationships and information explicitly stated from the start.

The multi-lingual LKB Part of the information acquired from different sources, in particular taxonomical data together with information extracted from the differentia part of definitions, was converted into a typed feature structure (TFS) representation formalism (augmented with a default inheritance mechanism and lexical rules) and loaded into the prototype multilingual Lexical Knowledge Base developed within the project. An example of formalised lexical entry is shown below, corresponding to the Italian entry for *acqua* 'water' (sense 1) which is defined in the Garzanti dictionary as *liquido trasparente, incoloro, inodoro e insaporo, costituito di ossigeno e idrogeno, indispensabile alla vita animale e vegetale* 'transparent, colourless, odourless, and tasteless liquid, composed of oxygen and hydrogen, indispensable for animal and plant life':

```
acqua G_0_1
< sense-id : dictionary > = ("GARZANTI")
< sense-id : homonym-no > = ("0")
< sense-id : sense-no > = ("1")
```

```
< lex-noun-sign rqs > < liquido_G_0_1< lex-noun-sign rqs >
< rqs : appearance > = transparent
< rqs : qual : colour > = colourless
< rqs : qual : smell > = odourless
< rqs : qual : taste > = tasteless
< rqs : constituency : spec > = "madeof"
< rqs : constituency : constituents : first_pred > = "ossigeno"
< rqs : constituency : constituents : rqs_first_pred > <
        ossigeno_G_0_0< lex-noun-sign rqs >
< rqs : constituency : constituents : rest_pred : first_pred > = "idrogeno"
< rqs : constituency : constituents : rest_pred : rqs_first_pred > <
        idrogeno_G_0_0b< lex-noun-sign rqs >
< rqs : constituency : constituents : rest_pred : rest_pred > =
        empty_list_of_preds_and_degrees.
```

Within the Acquilex LKB, lexical entries are defined as inheriting default information from other feature structures; those feature structures in their turn inherit from other feature structures. In lexical representation, default feature structures correspond to "genus" information; these feature structures are unified (through the default inheritance mechanism which is non-monotonic) with the non-default feature structure describing the information specific to the lexical entry being defined, which is contained in the "differentia" part of the definition. Hence, *acqua* inherits the properties defined for the lexical entry of *liquido* 'liquid'. The general properties are then complemented with information which is specific to the entry being defined; in the case at hand, these features specify colour, smell and taste as well as constituency for "acqua". A fully expanded definition of the same lexical entry is obtained by combining its TFS definition (i.e. the one shown above) with the TFS definition of each of its supertypes.

Different TFS lexicon fragments, circumscribed to semantic classes of verbs and nouns (e.g. motion verbs or nouns denoting food and drinks), are available for different languages. The table below illustrates, for each language, the coverage of the final TFS lexicons which have been developed within the project:

		Dutch	English	Italian	Spanish
Noun Entries	Number of LKB Entries Food subset	1190	594	702	143
	Number of LKB Entries Drink subset	261	202	147	254
Verb Entries	Number of LKB Entries Motion verbs subset		app. 360		303
	Number of LKB Entries Phsychological verbs subset		app. 200		

The fact that only part of the information extracted was converted into TFS form is also a consequence of the lack of flexibility of the class of TFS representation languages which causes difficulties in the mapping of natural language words - in particular word meanings which are ambiguous and fuzzy by their own nature - onto formal structures. In fact, the Acquilex experience showed the difficulty of constraining word meanings, with all their subtleties and complexities, within a rigorously defined organisation. Many meaning distinctions, which can be easily generalised over lexicographic definitions and automatically captured, must be blurred into unique features and values (see [Cal93]).

3.12.4 ET10/51

The ET10/51 "Semantic Analysis Using a Natural Language Dictionary" project (see [Sin94]) was aimed at the development of a methodology and tools for the automatic acquisition of lexical information from the Cobuild Student's Dictionary in view of the semi-automatic construction of lexical components for Natural Language Processing applications. Particular attention was on the extractability of information on the one hand, and on its exploitability within NLP applications on the other hand. As in the case of the other projects (see §3.12.3 and 3.12.5), the project did not aim at constructing a broad coverage lexical resource but rather at developing an appropriate lexical acquisition strategy for a corpus-based dictionary such as Cobuild. Again, it thus makes more sense to mention the information types which were extracted from dictionary entries and subsequently encoded according to the Typed Feature Structure Representation formalism (see section below), rather than to give detailed figures and numbers. Here suffice it to mention that the lexicon subset built within the project amounts to 382 entries representative of different parts of speech.

Unlike other work on the automatic analysis of machine readable dictionaries (see, for instance, the Acquilex projects) which focussed on semantic information which can be derived from the genus and differentia parts of the definition, in this project the acquisition work has mainly concentrated on syntagmatic links, namely the typical syntactic environment of words and the lexico-semantic preferences on their neighbours (whether arguments, modifiers or governors). In fact, due to the fact that Cobuild is a corpus-based dictionary and to particular structure of Cobuild definitions, this information type is systematically specified for all entries in the dictionary. However, also taxonomical information (i.e. hyperonymy, synonymy and meronymy relations) which could be extracted from the genus part of the definition was taken into account. Verb, noun and adjective entries were analysed and the extracted information was converted into a Typed Feature Structure Representation formalism following the HPSG theory of natural language syntax and semantics.

An example follows meant to illustrate the TFS representation format adopted within this project. The entry describes a transitive verb, **accent** (sense 4), which is defined in Cobuild as follows: "If you **accent** a word or a musical note, you emphasize it".

$$
\begin{bmatrix}
\text{PHON} & \langle \textit{accent} \rangle \\
\text{SYNSEM}|\text{LOCAL} & \begin{bmatrix}
\text{CAT} & \begin{bmatrix}
\text{HEAD} & \begin{bmatrix} \text{MAJOR} & \textit{verb} \\ \text{VFORM} & \textit{bse} \\ \text{PREF-VFORM} & \textit{active} \end{bmatrix} \\
\text{SUBJ} & \left\langle \text{NP} : \boxed{1} \begin{bmatrix} \textit{human} \end{bmatrix} \right\rangle \\
\text{COMPS} & \left\langle \text{NP} : \boxed{2} \begin{bmatrix} \begin{pmatrix} \textit{note} \\ \vee \\ \textit{word} \end{pmatrix} \end{bmatrix} \right\rangle \\
\text{LEX} & +
\end{bmatrix} \\
\text{CONTENT} & \begin{bmatrix}
\text{RESTR} & \left\{ \begin{bmatrix} \text{RELN} & \textit{accent} \\ \text{ARG.1} & \boxed{1} \\ \text{ARG.2} & \boxed{2} \end{bmatrix} \right\} \\
\text{LEXSEM} & \begin{bmatrix} \text{SYN} & \textit{emphasize} \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

As can be noted, this structure complies, to a large extent, with the general HPSG framework: it corresponds to the TFS associated with all linguistic signs, where orthographic ("PHON"), syntactic and semantic ("SYNSEM") information is simultaneously represented. The main differences lie in the insertion of Cobuild-specific features such as "DICTCOORD" (encoding the coordinates locating a given entry within a dictionary), "LEXRULES" (containing information about the lexical rules relevant to the entry being defined), "LEXSEM" (carrying information extracted from the genus part of the definition) and "U-INDICES" (i.e. usage indices, which characterize the word being defined with respect to its contexts of use, specified through the "REGISTER", the "STYLE" and the English variant ("DIAL-VAR") attributes). Other verb-specific attributes which have been inserted to represent Cobuild information are "PREF-VFORM" and "ACTION-TYPE", the former intended to encode the preferential usage of the verb being defined and the latter referring to the kind of action expressed by the verb, e.g. possible, likely, inherent, negative/unlikely, collective, subjective.

As in the case of Acquilex, also in this case the inadequacy of the formal machinery of a TFS representation language emerged, in particular with respect to the distinction between "constraining" and "preferential information". The distinction between constraints and preferences is not inherent in the nature of the data but rather relates to their use within NLP systems; e.g. the same grammatical specification (e.g. number or voice) can be seen and used either as a constraint or as a preference in different situations. Unfortunately, despite some proposals to deal with this typology of information, constraint-based formalisms as they are today do not appear suitable to capture this distinction (preferences are either ignored or treated as absolute constraints).

A sample of the TFS entries constructed on the basis of Cobuild information was then implemented in the Alep-0 formalism [Als91]. It emerged that, in its prototype version, this formalism presented several problems and limitations when used to encode lexical entries. The main objection was concerned with the expressivity of the formalism when dealing with lexical representations related to the inheritance between lexical entries. In fact, within the Alep framework inheritance between lexical entries is not supported, this mechanism being

	All PoS	Nouns	Verbs
Number of Entries	app.* 100	app. 15	app. 85
Number of Senses	app. 300	app. 50	app. 250
Morpho-Syntax	Yes		
Semantic Features	Yes		
Argument Structure	Yes		
Semantic Roles	Yes		
- Role Types	19		
Semantic Frames	Yes		
- Frame Types	3	1	3
Selection Restrictions	Yes		

Table 3.16: Numbers and figures for Delis (*app.= approximately).

restricted to the system of types. But, when dealing with the representation of semantic information within the lexicon, the choice of exploiting the taxonomic chain to direct the inheritance of properties between lexical entries appears quite natural; this was possible for instance within the Acquilex Lexical Knowledge Base (see [Cop91b]). In this way, many of the advantages of encoding the lexicon as a TFS system are lost since, potentially, each lexical entry could be used as a superordinate from which information could be inherited.

3.12.5 Delis

The LRE-Delis project aimed at developing both a method for building lexical descriptions from corpus material and tools supporting the lexicon building method. The main goal of the project was developing a method for making lexical description more verifiable and reproducible, also through linking of syntactic and semantic layers.
The project aimed at the development and assessment of the working method itself rather than the production of substantial amounts of data. Only information related to a relatively small number of verbs (plus a few nouns) was encoded. Lexical semantic descriptions of lexical items falling within some semantic classes (perception verbs and nouns, speech-act verbs, and motion verbs) were developed for various languages (Danish, Dutch, English, French, and Italian) by adopting the *frame semantics* approach (cf. [Fil92]; [Fil94]). Although a small set of lexical items was taken into consideration, several hundreds of sentences containing them were analysed and annotated in detail for each language (20+ types of semantic, syntactic and morphosyntactic annotations). A Typed Feature Structure dictionary was produced with entries for perception verbs of EN, FR, IT, DK, NL, related to the corpora sentences. Reports on the methodology followed, containing detailed discussion of the syntax/semantics of the other verb classes treated, are also available (e.g. [Hei95]).

Table 3.16 provides some numbers related to the kind of dataencoded:

The data encoded within Delis were acquired by manual work carried out on textual corpora. The methodology for corpus annotation, agreed on by all partners, is outlined in the CEES - Corpus Evidence Encoding Schema - ([Hei94]). This schema allows to:

- gather in a unique frame - following an HPSG-like approach - the properties belonging to different linguistic levels (namely, morphological, syntactic and semantic) of the item under analysis enabling to computationally retrieve them singularly or in correlation;

- encode these properties according to a standardized procedure, facilitating the comparison of the information in a multi-lingual environment.

Within DELIS, a list of aspects to be encoded for each verb and its surrounding context was agreed on for all the different linguistic layers:

- morphosyntactic data-base form and agreement features (based on EAGLES);
- syntactic information, described in terms of functional syntax (subject, object...) and of surface phrasal structures (NP, PP...);
- semantic information, encoded from two points of view:
  1. selection restrictions on verb arguments expressed in terms of features such as "human" and "concrete" (note that the set of features to be used in these descriptions was not obligatorily the same across all involved languages);
  2. thematic roles encoded according to Fillmore's theoretical approach of *frame semantics*.

One of the basic tasks of *frame semantics* is the schematic description of the situation types associated with the use of particular predicating words, by discovering and labelling elements of such situations in so far as these can be reflected in the linguistic structures that are built around the word being analysed. The DELIS approach made it possible to enucleate the common core of the linguistic behaviour associated with broad semantic classes - e.g. the perception class is mainly characterized by the Experiencer and the Percept roles, while the speech act class displays the three roles of Sender, Message and Receiver - and, at the same time, to discover the specific properties related to individual verb types.

As said above, many corpus sentences containing the words chosen were annotated. For perception verbs also TFS entries were produced. In the following an example of TFS is provided:

```
descry-att-tgt
[LEMMA: "descry"
 FEG: < fe
          [FE: exper-i
               [INTENTION: +
                SORT: human]
           GF: subj
           PT: np]
       fe
         [FE: p-target
              [EXPECTED: +
               SPECIFICITY: +
               SALIENCE: -
               DISTANCE: +
               INTEREST: +]
          GF: comp
          PT: np] >
```

```
 EVENT: vis-mod
         [MODALITY: vis
          DURATION: duration]].
```

The sense of the verb *descry* described in this FS involves an *intentional experiencer* (indicated by characterising the verb as 'att' = *attention*) and a *percept-target* ('tgt'). The attribute FEG ('Frame Element Group') has a list containing two frame elements as its values: 'exper-i' and 'p-target'. Some semantic features are encoded for each frame element, but also their 'grammatical function' and 'phrase-type'. Finally, also the 'event properties' of the verb are indicated: in this case we have a 'visual' MODALITY and a DURATION which is not further specified (although it could also be given a 'short' or 'long' value).

The most interesting observation, derived from the data emerging from an analysis of the corpus, however, is that meaning distinction cannot always rely on the information taken from phrasal types, grammatical functions and their thematic roles. As is demonstrated by the data discussed in various reports (e.g. [Mon94]), idiosyncratic meanings can be enucleated by taking other information into account, usually missing in traditional paper dictionaries, at the level of morphosyntax, semantics, collocations, statistics and the interactions between different levels of information (cf. also [Cal96]).

3.12.6 Comparison with Other Lexical Databases

In general we can state that the coverage in the experimental lexicons is much smaller than the other resources discussed here, but the richness and explictness of the data is much higher.

Corelex should be seen as the implementation of a particular theoretical approach to lexical semantics, capturing the dynamic properties of semantics. In this respect is it radically different from any of the other resources discussed here. Only in the case of EuroWordNet some of these ideas are being implemented in the form of the complex ILI-records (see §3.6.3). The underspecified types extracted in Corelex will be used as input in EuroWordNet (§3.6.3). The Acquilex multilingual LKB is a "prototype" database containing highly-structured and formalized data covering a well-defined set of syntactic/semantic classes of words. The semantic specific includes a QUALIA approach similar to Corelex, where meanings can also be derived by means of lexical rules. The language-specific lexical databases developed within the same project are on the one hand much richer in coverage, as traditional monolingual dictionaries, but are less formalized. As such they are closer to wordnets (§3.6). Furthermore, the English lexicons in Acquilex have been derived from the Longman dictionaries, showing that it is possible to derive complex lexicons from such resources.

The ET-10/Cobuild lexical database is constituted by a typology of entries of different parts of speech which were selected as representative of the different defining strategies adopted within the dictionary; this entails that the acquisition tools developed within the project should in principle be able to deal with the whole set of Cobuild entries. Hence, unlike other similar projects (such as Acquilex and Delis), here the set of formalised entries does not represent a semantically homogeneous dictionary subset but rather a typology of structurally different entries.

As said above, the data encoded within DELIS are only related to a small group of lexical items, chosen among words found in coherent semantic classes, and encoded just as an illustration of the working method followed. Thus the database itself, although rich of information on single verbs/nouns, does no longer appear as 'rich' when its coverage is considered, for instance when we compare it with resources such as EDR (§3.8) which contains quite similar semantic information for substantial portions of Japanese and English. Furthermore, within DELIS the focus of attention was mainly on the syntactic/semantic features of the different *frame* elements, whereas semantic relations such as those encoded in WordNet (§3.6.2) or EuroWordNet (§3.6.3) were not explicitly considered.

3.12.7 Relation to Notions of Lexical Semantics

Corelex is direct implementation of the Generative Approach described in §2.7. Something similar can be said for Acquilex, where rich Qualia structures have been built up, from which sense extensions can be derived via lexical rules. Furthermore, the research in Acquilex focused on properties that are prominent for describing the most salient semantic characteristics of words and/or as strongly connected with syntactic properties of words. Thus, much information is encoded both in the LKB and in the LDBs with respect to semantic relations such as synonymy, hyponymy and meronymy, which are central notions in lexical semantic research. Moreover, semantic information given both in the multilingual LKB for fragments of the various lexica and in the monolingual LDBs on noun quantification (§2.7.4), verb causativity/inchoativity (§2.5.2, 2.6.2), verb meaning components (§2.5.2) and lexical aspect (§2.2), etc. addresses questions concerning the syntax-semantics interface, which have been deeply investigated in these years.

In the field of lexical semantics it is commonly assumed that important semantic properties of a lexical item are reflected in the relations it contracts in actual and potential linguistic contexts, namely on the syntagmatic and paradigmatic axes [Cru86]. Cobuild defining strategy takes into account both descriptive dimensions and accounts for both of them within the same definition structure. Hence, the ET-10/Cobuild lexicon contains information about synonymy, hyponymy and meronymy as well as about the typical syntactic-semantic environment of a given word.

The combination of syntactic and semantic information encoded in the DELIS database can be useful to address questions concerning the syntax-semantics interface. Furthermore, there is a strong relation between the characterization of predicate arguments in terms of frame elements and the traditional notion of thematic relations.

3.12.8 LE Uses

As small-scale experimental lexicons they have not be used in realistic applications.

The main goal of the Acquilex project was the development and evaluation of different directions of research in related areas, ranging from automatic acquisition of lexical information from different sources and subsequent formalisation and multilingual linking in a LKB. Hence, its outcome mainly consists in the theoretical and methodological background for the creation of resources to be used within NLP applications.

The main goal of the ET-10 project was the development of a lexical acquisition strategy for the Cobuild dictionary and related tools. Hence, its outcome should be mainly considered from the methodological point of view. Yet, the acquisition tools developed within the project could in principle be usefully exploited to semi-automatically construct lexical resources for NLP applications.

Since the beginning, DELIS was conceived as a 'methodological' project whose purpose was to establish a theoretically motivated methodology for corpus-based computational lexicography and thus to prepare the ground for future development projects.

3.13 Bilingual Dictionaries

In the final section of this chapter, two traditional bilingual dictionaries will be described that have been used in various projects to derive information for NLP or that are directly used in computerized tools. The Bilingual Oxford Hachette French and the Van Dale Dutch-English dictionaries are just illustrative for many other bilinguals that can be used in this way.

3.13.1 The bilingual Oxford Hachette French dictionary

The bilingual Oxford-Hachette French Dictionary (French-English) (OHFD) is intended for general use and is not specific to any domain. It includes most abbreviations, acronyms, many prefixes and suffixes, and some proper names of people and places in cases of unpredictable translation problems. It also includes semi- technical terms which might be found in standard texts, but omits highly domain-specific technical terms. It is designed to be used for production, translation, or comprehension, by native speakers of either English or French. The electronic version of the OHFD is an sgml tagged dictionary. Therefore each element is tagged by function. For instance there are tags to indicate, part-of-speech, different meaning within a part of speech, pronunciation, usage, idiomatic expressions, domains, subject and object collocates, prepositions, etc. Unfortunatly, translations are not tagged at all, which makes the dictionary sometimes difficult to parse.

Size The English-French side has about 47 539 entries (most of the compounds are entries by themselves) which are divided into: 31061 nouns, 11089 adjectives, 5632 verbs, 2761 adverbs, others 165. There are 407 top level labels, a good part of which include a distinction between British English and American English. They can be just one concept (Bio), or a combination of a concept and language level (GB Jur).

The French-English side has 38944 entries (about 10.000 compounds which are part of entries themselves) which are divided into: 25415 nouns, 8399 adjectives, 4805 verbs, 1164 adverbs and 890 others. The top level labels are about 200 labels.

Homographs Homographs are represented in two ways:

- As separate entries. (We shall henceforth use "homograph" to denote these separated entries.). In this case, a number will be included with the headword lemma to distinguish it from its homographs, as in:
  row<hm>1</hm>...<ph>r@U</ph>...(<ic>line</ic>)...
  row<hm>2</hm>...<ph>raU</ph>... (<ic>dispute</ic>)...;

- As major subdivisions within a single entry. (We shall henceforth use "grammatical category" to denote such subdivisions.) In the OHFD, these are labelled with roman numerals, as in:
  row<hm>1</hm>...<ph>r@U</ph>...
  <s1 num=I><ps>n</ps> ...(<ic>line</ic>)...

<s1 num=III><ps>vi</ps> <ann><la>Naut</la>,
<la>Sport</la>...
The basis for deciding when to assign separate homograph entries to identically spelt words is a difference of pronunciation, not semantic or etymological independence. In addition, function words are given a separate entry from homographs even of the same pronunciation, as in:
mine<hm>1</hm>...<ph>maIn</ph>...<fi>le mien</fi>, <fi>la mienne</fi>...
mine<hm>2</hm>...<ph>maIn</ph>...<la>Mining</la> mine <gr>f</gr>...

Sense Counter Monosemous words have no overt identifier of their single sense, as in mineralogy:...<hg><ps>n</ps></hg> minéralogie <gr>f</gr>.</se>
This component is found only in the entries for polysemous words, as in:
migrant:...<s2 num=1><la>Sociol</la>... <s2 num=2><la>Zool</la>...
Senses are distinguished in considerable detail, although it should be remembered that in a bilingual dictionary the sense differentiation of the headword is often affected by target language (TL) equivalence. The original source language (SL) analysis of, for instance, the English word 'column' would yield eight or nine senses, covering architectural, military and newspaper columns, as well as columns of figures and columns of smoke; with French as the TL, there is only one 'sense' in the 'column' entry, since every sense of the English word has the French equivalent 'colonne'.

Word Usage Labels They include:

- Stylistic Label: An example of a style label is "littér" (literary) in
milling:...<la>littér</la> <co>crowd</co> grouillant...
Other items may be marked as belonging to administrative, or technical, or poetic language.
The OHFD also marks formal terms as "fml" (French "sout" for "soutenu"). However, in order to avoid terms such as "colloquial" or "familiar", which are open to individual interpretation, the OHFD has devised a system of symbols showing points on a scale from (roughly) "informal" to "informal, may be taboo" .
The presence of these labels, which may be attached to SL and to TL items, allows both the decoder and the encoder to align words of similar style in the two languages.

- Diatechnical Label: Semantic domain labels are exceedingly frequent in the OHFD, as may be seen in:
matrix:...<la>Anat</la>, <la>Comput</la>, <la>Ling</la>, <la>Math</la>, <la>Print</la>, <la>Tech</la> matrice <gr>f</gr>; <la>Miner</la> gangue <gr>f</gr>...
Moreover, the dictionary tapes actually contain a more comprehensive coverage of semantic domain labels than appears in the printed text. Whenever the lexicographers believed that the lemma was frequently (or exclusively) used when a particular subject was under discussion, they marked the lexical unit with an appropriate semantic domain marker.

- Diatopic Label: This type of labelling is used in the print dictionary to mark such regional varieties as Belgian French, Swiss German and American English, as "US" in
math:... <la>US</la> =<xr><x>maths</x></xr>...

or “GB” in
bedroom:...<le>a two &hw. flat <la>GB</la> <u>ou</u> apartment</le> un trois pièces...

- Diachronic Label: This type of labelling allows words or usages to be marked as “old-fashioned”, ”obsolete” or “archaic”, etc. In the following, “mae west” is marked as old-fashioned:
  mae west:...&dated....
  This marker’s presence (it may be attached to SL and to TL items) allows both the decoder and the encoder to align words of similar currency in the two languages.

- Evaluative Label: This warning label is used to indicate a favourable (”appreciative”) or unfavourable (”pejorative”) attitude in the speaker or writer, as”(pej)” in:
  macho:...<la>pej</la>macho;...
  where it shows that to describe someone as “macho” is not a compliment (in English at least). Its presence (it may be attached to SL and to TL items) allows both the decoder and the encoder to align words indicating similar attitudes in the two languages.

- Frequency Label: These indicate the uncommon or rare word forms, as in:
  stride:...<s1 num=III><ps>vtr</ps> (<gr>prét</gr> <fs>strode</fs>,
  <gr>pp</gr> <la>rare</la> <fs>stridden</fs>)...

- Figuration Label: These indicate figurative, literal, or proverbial uses as in:
  mist:...<la>fig</la> (<ic>of tears</ic>) voile...

Cultural Equivalent For certain culture-specific concepts, the source lemma does not have a direct semantic equivalent, but there is an analogous concept in the target language culture which serves as a translation, as in:
high school:...<la>US Sch</la> &appr. lycée <gr>m</gr>;...

Sense Indicators The indicator may be a synonym or paraphrase in the form of a single word, as ”information” or”data” in:
material:... (<ic>information, data</ic>) documentation...
or a phrase, as “become proficient in” in:
master: ...(<ic>learn, become proficient in or with</ic>) matriser <co>subject, language, controls, computers, theory, basics, complexities</co>;...
The OHFD also includes sense clue labels as additional information. The sense clue is usually a brief phrase, as “of specified nature” or “requiring solution” or “on agenda” in:
matter... <s2 num=1><la>gen</la> chose <gr>f</gr>; (<ic>of specified nature</ic>) affaire <gr>f</gr>; (<ic>requiring solution</ic>) problème <gr>m</gr>; (<ic>on agenda</ic>) point <gr>m</gr>;...
These may also be used to more finely distinguish subsenses within the same substituting indicator, like “in chess” and “in draughts” here:

man :...<s2 num=7><la>Games</la> (<ic>piece</ic>) (<ic>in chess</ic>) pièce <gr>f</gr>; (<ic>in draughts</ic>) pion <gr>m</gr></s2>; ...

Subcategorisation It includes indication of prepositional usage, subject and object collocate for semantic complementation. The OHFD attempts to show all the structures necessary if the full semantic potential of the translation equivalent(s) is to be expressed grammatically. Examples of these structures ("through, au moyen de") are to be seen in:
mediate:...diffuser <co>idea, cult</co> (<pp><sp>through</sp> au moyen
mediate:...diffuser <co>idea, cult</co> (<pp><sp>through</sp> au moyen de, par</pp>)...
Similar to this type of information is information relating to the obligatory grammatical environment of the translation word, such as "(+subj)" in marvel:... <ls>to &hw. that</ls> s'étonner de ce que (<gr>+ subj</gr>)...

Collocators The type of collocator which may be offered depends on the word class of the headword; in the print dictionary the following types of collocators are used (the relationship is of course with one lexical unit, i.e. a single combination of lexical component and semantic component, a monosemous word or a polysemous word in one of its senses):

- Verb headwords have as collocators nouns that are typical subjects of the verb as in :
  merge:...<co>roads, rivers</co> se rejoindre; ...,
  or nouns that are typical objects of the verb as in
  merge: ...<lo>to &hw. sth into <u>ou</u> with sth</lo> incorporer qch en qch <co>company, group</co>...
- Adjective headwords have as collocators nouns that typically are modified by the adjective as in
  messy:...(<ic>untidy</ic>) <co>house, room</co> en désordre;
- Noun headwords have as collocators one of the following:
  - (for deverbal nouns) nouns that are the object of the cognate verb as in management: ... (<ic>of business, company, hotel</ic>) gestion ... or that are the subject of the cognate verb as in
    maturation: ...(<ic>of whisky, wine</ic>) vieillissement ...;
  - (for nouns that are nominalisations of adjectives) nouns modified by the cognate adjective as in
    mildness:...(<ic>of protest</ic>) modération;
  - (for concrete nouns with real-word referents) nouns to which the headword is related by meronymy or nouns naming objects that stand in some other real-world relationship to the object that is the referent of the headword noun as in
    mug:...(<ic>for beer</ic>) chope...
  - (for nouns used to modify other nouns) the collocators given are typical of the semantic set(s) thus modified as in
    mango:...(<ic>tree</ic>) manguier ... [<lc>grove</lc>] de manguiers;
  - for adverbs: typical verbs and/or adjectives modified by the adverb as in
    marvellously:...<co>sing, get on</co> à merveille; <co>clever, painted</co> merveilleusement;

Collocations Often, even within a single semantic sense, the lemma will translate differently depending on words which appear with it. For example, in:

> accident: ...[<lc>figures</lc>, <lc>statistics</lc>] se rapportant aux accidents; [<lc>protection</lc>] contre les accidents;...

the lemma should be translated as "se rapportant aux accidents" when it appears with "statistics", but as "contre les accidents" when it appears with "protection.
Collocation (tagged by <lc> in the OHFD) should not be confused with either compounds (multi-word lexemes) (tagged <cw>), which include and translate the co-occurring words as part of the lemma, nor with collocators (tagged <co>), which help to identify distinct senses of the lemma.

Multi-Word Lexeme Multi-word lexical units occurring as sublemma forms may generate almost all the lexicographic items that normally constitute a full dictionary entry, with the exception of a phonetic transcription. The three principal types of multi-word sublemmas are:
(a) compounds, as in:
mud:...<cw>mudbank</cw>... banc <gr>m</gr> de vase; ...<cw>mud bath</cw>...(<ic>for person, animal</ic>) bain <gr>m</gr>de boue;...
(b) phrasal verbs, as in:
miss:...<pvp><lp>&hw. out</lp> être lésé; ... (c) idioms, as in:
miss:...<id>
to &hw. the boat <u>ou</u> bus
&coll. rater le coche</id>;...
Multi-word lexemes may range from fixed phrases (e.g. "by and large") through idiomatic verb phrases ("raining cats and dogs") to whole sentences, such as proverbs or phatic phrases ("Have a nice day").

Gloss This is given when there is no direct TL equivalent of the lemma, as in:
mid-terrace:...[<lc>house</lc>, <lc>property</lc>] <gl>situé au milieu d'un alignement de maisons identiques et contiguës</gl>...

3.13.2 Van Dale Bilingual Dutch-English

The Van Dale Bilingual Dictionaries are developed for native Dutch speakers. This means that the resources contain only very limited information on the Dutch words and much more information on the Foreign-Language target words. The Dutch-English dictionary is described here [Mar86].

The Van Dale Dutch-English is a traditional bilingual dictionary. It is structured using a tagged field structure, which makes it relatively easy to extract a computer tractable database from it. However, the scope of the fields is not always explicit and the values within the fields are often undifferentiated and consist of free-text.

The entry-structure is homograph-based but homographs are distinguished only when the part-of-speech differs and/or the pronunciation. Sub-homographs are used when senses differ in major grammatical properties such as valency, countability, predicate/attributive usage. The figures supplied in Table 3.17 provide an indication of size and coverage.

In addition to some grammatical information on the Dutch words and the English translations, the dictionary contains a large amount of semantic information restricting the senses and the translations:

Entries	90925
Homographs	2967
Sub-homographs	6769
Senses	127024
Main Translations	145511
Secondary Translations	104181
Examples	111226

Table 3.17: Number of Entries, Senses and Translations in the Van Dale Dutch-English Dictionary

Sense-indicators (53368 tokens) to specify the Dutch senses or polysemous entries. These contain bits and pieces from original definitions (often a genus word).

Biological gender marker for English translations. This is necessary to differentiate translations when the source and target language have different words for male or female species: 286 translations are labeled as male, 407 translations as female.

Usage labels for domain, style and register Applies to both Dutch senses and their English translations.

Dialect labels for Dutch senses and their English translations

Context markers (23723 tokens, 16482 types). These are semantic constraints differentiating the context of multiple translations, and to limit the scope of translations having a narrower context than the Dutch source sense.

The usage labels and the domain labels are mostly stored in the same field. Differentiation has to be done by some parsing. The usage labels form a limited closed set of abbreviations and codes, the domain labels are free text. For the main-translations about 400 different types of values occur.

The context markers and sense-indicators are in textual form (Dutch). Their interpretation varies from domain-labels, selection restrictions, semantic classifications and semantic properties. The difference in interpretation is however not coded but can partially be inferred, e.g.:

- a noun used as a constraint or sense-indicator for a verb or adjective is mostly a selection restriction;
- a noun used as a constraint for a noun is a classification;
- an adjective constraining a noun is a feature;
- an adverb constraining a verb indicates a more specific manner.

Finally, a lot of collocational information is stored in the examples and their translations. The typical combination words are marked and distinguished per part-of-speech. If the combination is compositional then the correlating meaning of the entry is given, in the case of idiosyncratic collocation there is a mark. The examples and their translations can be seen as partially structured context specification for the Dutch and English word pairs.

3.13.3 Relations to Notions of Lexical Semantics

Bilingual resources often contain information which disambiguate the usage of words in the target languages, but only in so far it is necessary to select the correct alternative. The information takes the form of semantic classes, selection restrictions, register and domain labels or morpho-syntactic information, but it requires considerable processing to differentiate between them. Somewhat more sophisticated information is available in the form of the examples and the translations of the examples. The combinatoric constraints provide very useful information comparable to Mel'cuk's lexical semantic functions [Mel89].

3.13.4 LE Uses

Obviously, bilingual dictionaries are useful input for constructing Machine-Translation systems, although a lot of additional work has to be carried out to formalize and complete the information. In Acquilex, it has neverthless been used to automatically extract equivalence-relations between English and Dutch word-senses (see [Cop95b]).

Because elements are tagged by functions the OUP is a very convenient dictionary to retrieve information from. The Oxford Hachette has been successfully used to design an intelligent dictionary lookup, Locolex, first developed in the framework of the COMPASS European project.

Furthermore, [Mak95] show that it is possible to achieve high degree of sense-disambiguation using the rich annotations in the examples and their translations in bilingual dictionaries.

Part III

Semantic Requirements for NL and IS Applications

Semantic Requirements for NL and IS Applications

The following two chapters review the requirements, both actual and potential, that language engineering (LE) applications place on lexical semantic resources. Given the scope and rapdily changing nature of LE applications any such review is bound to be partial. We have chosen to address the topic in two ways. First, in Chapter 4, we discuss areas of LE application within which lexical semantics has or can have some role. The application areas addressed are machine translation, information retrieval, information extraction, summarisation and natural language generation. For each we give a brief overview of the application, survey the approaches to the application that have been undertaken to date, point to related application areas and techniques, and discuss the role that lexical semantics has or could play in the application.

In Chapter 5 we review various LE component technologies that utilise lexical semantic information and that are being, or may be, used in a variety of different LE application areas. These techniques address such tasks as semantic word clustering, multiword recognition, word sense disambiguation, proper name recognition and parsing and coreference. Unlike the application areas identified in Chapter 4, these component technologies are of no intrinsic interest to an end user; rather they are building blocks from which LE applications are constructed. Their value lies in being resuable across applications. So, for instance, robust word sense disambiguation could benefit both machine translation systems and information retrieval systems. For each component technology reviewed we give a brief overview, survey existing approaches, discuss the role of lexical semantics in the component technology, and point to applications that are or could utilise the technique.

Chapter 4

Areas of Application

4.1 Machine Translation

4.1.1 Introduction

Our survey focusses on four types of systems:

- Interlingual MT;
- Knowledge-based MT (KBMT) ;
- Language-based MT (LBMT), and
- Example-based MT (EBMT).

4.1.2 Survey

Interlingual MT As an illustrative example of this approach, we report on work done by Bonnie Dorr and the system PRINCITRAN [Dor93, Dor95a, Dor95b]. They use an automatic classification of verbs using a rich semantic typology based on Levin's work [Lev93]. They defined thematic grids based on a more fine-grained verbal classification. Lexical entries for SL and TL include thematic roles. Thematic roles are necessary for parsing, for building an interlingual semantic representation. For instance argument positions are associated with a wider variety of semantic roles, i.e. intransitives are not uniformly marked 'ag' (agent), but may be marked 'th' (theme), depending on the real logical role of the argument. Hypothesis used: semantically related verbs have similar, or identical thematic grids, i.e. verbs of the same semantic class have the same thematic specification. Dorr experimented with methods combining LDOCE codes (see §3.2) with Levin's verb classes.

KBMT To be successful it requires a large amount of hand-coded lexical knowledge. The cost-effectiveness of this task can be alleviated by partial automation of the knowledge acquisition process for the build up of concept dictionaries, see [Lon95] with reference to the KANT system. Terminological coverage crucial. KANT is interlingua based. They identify a set of domain semantic roles related with prepositions, and then potentially ambiguous role assignment. So, the distinguished semantic roles are associated with specific syntactic patterns (these two steps are done manually). KBMT seeks to translate from the meaning of the source text derived with the help of a model of background knowledge (see [Car]).

The semantic analysis derived from the source string is augmented by information from the ontology and domain models. The success of KBMT depends on the following: having a good model of **conceptual universals** of the application domain across cultures and a **linguistic model** which can describe the mapping between linguistic form and meaning. The use of a model of background knowledge and the lack of a principled methodology for conceptual modelling present the major difficulty in this approach. Another way of using the background knowledge in MT is proposed by ATR's transfer-driven MT, where translation is performed at the linguistic level rather than the conceptual level as in KBMT.

LBMT Language-based MT has two subtypes of systems: lexeme-oriented and grammar-oriented. The modern descendants of lexeme-oriented approaches assume that translation can be captured by lexical equivalence and units of translation are words and specific phrases (see [Tsu93]). The grammar-oriented approach considers the structural properties of the source text in disambiguation and translation but confines itself to intra-sentential structures. Its advantage over the lexeme approach can be seen in two respects: the concept of translation equivalence exists not only in terms of source strings but also in the structure of the source strings. The unit of translation is extended to the sentential scope. The problem with both types of LBMT is that they do not give much attention to the context of language use. The context can be textual, such as discourse organisation, social.

EBMT Example-based machine translation proposes translation by examples collected from past translations. The examples are annotated with surface descriptions, e.g. specific strings, word class patterns, word dependencies (see [Sat]), predicate frames (see [Jon]) for the purpose of combining translated pieces. Transfer is effected on *similarity* in terms of these descriptions. The knowledge used for transfer is specific rather than general. The immediate criticism about specific knowledge representation is that its reliance on concrete cases limits the coverage of the knowledge model. EBMT builds its transfer on approximate reasoning i.e. there is no single correct translation, but degrees of acceptability among renditions. EBMT finds the translation of an s-string by its probability of translations calculated on the basis of a corpus of aligned bilingual texts. The best rendering of a source string is found by comparing it with examples of translation. The key words are compared by calculating their distance according to a thesaurus (see [Fur]).

4.2 Information Retrieval

4.2.1 Introduction

Information retrieval (IR) systems aim to provide mechanisms for users to find information in large electronic document collections (we ignore voice, audio, and image retrieval systems here). Typically this involves retrieving that subset of documents (or portions thereof) in the collection which is deemed relevant by the system in relation to a *query* issued by the user. The query may be anything from a single word to a paragraph or more of text expressing the user's area of interest. Generally IR systems work by associating a set of terms (*index terms*) with each document, associating a set of terms with the query (*query terms*) and then performing some similarity matching operation on these sets of terms to select the documents which are to be returned to the user.

One of the main problems with which the end-users are faced is to parse efficiently their intentional questions in the query language that the computer systems allow. Some people suggest specific logic or hypertext-based technologies to assist end-users in their task. (see [Rij]). Many authors propose methods based on semantic networks used to express declarative knowledge relevant to concepts and their relationships, (see [Sow92]).

4.2.2 Survey

Conceptual IR & Indexing

BADGER [BAD] is a text analysis system which uses linguistic context to indentify concepts in a text. The key point of the system is that single words taken out of context may not relate to the same concept as the phrase to which that word belongs. It therefore, aims to find linguistic features that reliably identify the relevant concepts, representing the *conceptual content* of a phrase as a case frame, or concept node (CN).

CRYSTAL [Sod95] automatically induces a dictionary of CNs from a training corpus. The CN definitions describe the local context in which relevant information may be found, specifying a set of syntactic and semantic constraints. When the constraints are satisfied for a portion of text, a CN is instantiated.

Woods, W. at Sunlabs [Woo] uses semantic relationships among concepts to improve IR. Use of NLP and knowledge representation techniques to deal with differences in terminology between query and target. Development of a prototype system for indexing and organising information in structured conceptual taxonomies.

DEDAL [Bau, DED] is a knowledge-based retrieval system which uses a conceptual indexing and query language to describe the content and form of design information. DE-KART is a KA tool which refines the knowledge of DEDAL by increasing the level of generality and automation.

Use of Thesauri

Thesauri are used to expand query or index terms to permit broader matching.

Use of SNOMED/UMLS [Ama95] use SNOMED (see §3.10) and a formal grammar to create templates with a combination of syntactic and semantic labels.

[Nel95] use the UMLS metathesaurus (see §3.10) to identify concepts in medical knowledge.

Use of WordNet [Sta96] present a system to identify relevant concepts in a document and to index documents by concept using WordNet (§3.6.2). Lexical chains are generated which describe the context surrounding the concept. Documents are indexed using WordNet synsets as tokens, as opposed to terms.

Use of Semantic Verb Frames

Semantic verb frames to permit more constrained retrieval on entities playing certain roles in verbal case frames, see [Jin] and other semantic classification programs [Hat].

4.2.3 Role of Lexical Semantics

Lexical semantic relations, as discussed in §2.3 and 3.6 (see [Kro92, Kro97]).

4.2.4 Related Areas and Techniques

Word clustering and word sense disambiguation techniques, as discussed in §5.1 and 5.3 (see also [Kro92, Kro97, Res97]).

4.3 Information Extraction

4.3.1 Introduction

Information extraction (IE) systems extract certain predefined information about entities and relationships between entities from a natural language text and place this information into a structured database record or *template*. In IE there is the strong necessity of striking a balance between extensive generic lexical resources (e.g. WordNet, discussed in §3.6.2 and application dependent resources. Until now limited application-related resources have been shown to be more effective in applied systems, as most of the systems have been applied to very restricted domains (e.g. in the MUC competitions). But scaling up the IE technology to wider domains (e.g. search in the Net) would necessarily need the use (or re-use) of extensive resources.

4.3.2 Survey

The main information needed by an IE system are mainly related to subcategorization frames (for verbs, nouns and adjectives), information related to nominalization (e.g. produce/production) or adjectivization (pagare/pagabile). Most of these information are not provided by WordNet-like resources but are needed. In any case tools for adapting a generic lexicon to the actual need of specific domains are necessary, as applicative domains often show deviations with respect to the normal use of language in terms of:

- the kind of subcategorization frame: the frame may change according to specific uses; for example the verb "indicare" in standard Italian is a normal intransitive verb, while in the financial domain has an additional argument related to the value introduced by the preposition "a" (e.g. "i titoli sono stati indicati al 2%"); also role restrictions can change;
- meaning: very often words assume additional meanings in specific domains; for example the verb "indicare" in italian means "to point"; but in the finacial domain it is used to introduce prices for listed shares;
- familiarity for example the verb "to index" is considered rare in standard English (is not even listed in WordNet), but it is very familiar in finance and computer science.

The use of generic resources in analysing texts in restricted domains also introduce the problem of the relation between the domain description available or needed by the system (in order to reason on the extracted information; e.g. the knowledge base) and the generic lexical semantic definition given by the generic resources. Partial overlaps can be found, but the domain specific description is likely to be more precisely defined and reliable.

4.3.3 Role of Lexical Semantics

- many IE systems have domain-specific lexicons that store e.g. subcat patterns with highly domain-relevant verbs in such a fashion as to permit them to use the patterns directly in a pattern matching engine and then map the matched patterns directly into a template;

- coreference which is necessary for IE (and task is now formally part of the MUC evaluations) requires syno-, hyper-, and hyponymic lexical information;

- IE relies on proper name and multiterm identification (see below);

4.4 Text Summarization

4.4.1 Introduction

With the proliferation of online textual resources, an increasingly pressing need has arisen to improve online access to textual information. This requirement has been partly addressed through the development of tools aiming at the automatic selection of document fragments which are best suited to provide a summary of the document with possible reference to the user's interests. Text summarization has thus rapidly become a very topical research area.

4.4.2 Survey

Most of the work on summarization carried out to date is geared towards the extraction of significant text fragments from a document and can be classified into two broad categories:

- **domain dependent approaches** where a priori knowledge of the discourse domain and text structure (e.g. weather, financial, medical) is exploited to achieve high quality summaries, and

- **domain independent approaches** where a statistical (e.g. vector space indexing models) as well as linguistic techniques (e.g. lexical cohesion) are employed to identify key passages and sentences of the document.

Considerably less effort has been devoted to "text condensation" treatments where NLP approaches to text analysis and generation are used to deliver summary information of the basis of interpreted text [McK95].

Domain Dependent Approaches

Several domain dependent approaches to summarization use Information Extraction techniques ([Leh81, Ril93]) in order to identify the most important information within a document. Work in this area includes also techinques for Report Generation ([Kit86]) and Event Summarization ([May93]) from specialized databases.

Domain Independent Approaches

Most domain-independent approaches use statistical techniques often in combination with robust/shallow language technologies to extract salient document fragments. The statistical techniques used are similar to those employed in Information Retrieval and include: vector space models, term frequency and inverted document frequency ([Pai90, Rau94, Sal97]). The language technologies employed vary from lexical cohesion techniques ([Hoe91, Bar97]) to robust anaphora resolution ([Bog97]).

4.4.3 Role of Lexical Semantics

In many text extraction approaches, the essential step in abridging a text is to select a portion of the text which is most representative in that it contains as many of the key concepts defining the text as possible (*textual relevance*). This selection must also take into consideration the degree of *textual connectivity* among sentences so as to minimize the danger of producing summaries which contain poorly linked sentences. Good lexical semantic information can help achieve better results in the assessment of textual relevance and connectivity.

For example, computing *lexical cohesion* for all pair-wise sentence combinations in a text provides an effective way of assessing textual relevance and connectivity in parallel [Hoe91]. A simple way of computing lexical cohesion for a pair of sentences is to count *non-stop* (e.g. closed class) words which occur in both the sentences. Sentences which contain a greater number of shared non-stop words are more likely to provide a better abridgement of the original text for two reasons:

- the more often a word with high informational content occurs in a text, the more topical and germane to the text the word is likely to be, and

- the greater the times two sentences share a word, the more connected they are likely to be.

The assessment of lexical cohesion between sentences in a text can be significantly improved by using semantic relations such as synonymy, hyp(er)onymy and other thesaural relations in addition to simple orthographic identity [Hoe91, Mor91, Hir97, Bar97] as well as semantic annotations such as subject domains [San98] in addition to simple orthographic identity.

4.4.4 Related Areas and Techniques

Related area of research are: Information Retrieval, Information Extraction and Text Classification.

4.4.5 Glossary

Summarization, Abridgement, Information Retrieval, Information Extraction, Text Classification, Lexical Cohesion.

4.5 Natural Language Generation

4.5.1 Introduction

The problem of automatic production of natural language texts becomes more and more salient with the constantly increasing demand for production of technical documents in multiple languages; intelligent help and tutoring systems which are sensitive to the user's knowledge; and hypertext which adapts according to the user's goals, interests and prior knowledge, as well as to the presentation context. This section will outline the problems, stages and knowledge resources in natural language generation.

4.5.2 Survey

Natural Language Generation (NLG) systems produce language output (ranging from a single sentence to an entire document) from computer-accessible data usually encoded in a knowledge or data base. Often the input to a generator is a high-level communicative goal to be achieved by the system (which acts as a speaker or writer). During the generation process, this high-level goal is refined into more concrete goals which give rise to the generated utterance. Consequently, language generation can be regarded as a goal-driven process which aims at adequate communication with the reader/hearer, rather than as a process aimed entirely at the production of linguistically well-formed output.

Generation Sub-Tasks

In order to structure the generation task, most existing systems divide it into the following stages, which are often organised in a pipeline architecture:

Content Determination and Text Planning: This stage involves decisions regarding the information which should be conveyed to the user (content determination) and the way this information should be rhetorically structured (text planning). Many systems perform these tasks simultaneously because often rhetorical goals determine what is relevant. Most text planners have hierarchically-organised plans and apply decomposition in a top-down fashion following AI planning techniques. However, some planning approaches (e.g., schemas [McK85], Hovy's structurer [Hov90]) rely on previously selected content – an assumption which has proved to be inadequate for some tasks (e.g., a flexible explanation facility [Par91, Moo90])

Surface realisation : Involves generation of the individual sentences in a grammatically correct manner, e.g., agreement, reflexives, morphology.

However, it is worth mentioning that there is no agreement in the NLG community on the exact problems addressed in each one of these steps and they vary between the different approaches and systems.

Knowledge Sources

In order to make these complex choices, language generators need various knowledge resources:

- **discourse history** – information about what has been presented so far. For instance, if a system maintains a list of previous explanations, then it can use this information to avoid repetitions, refer to already presented facts or draw parallels.
- **domain knowledge** – taxonomy and knowledge of the domain to which the content of the generated utterance pertains.
- **user model** – specification of the user's domain knowledge, plans, goals, beliefs, and interests.
- **grammar** – a grammar of the target language which is used to generate linguistically correct utterances. Some of the grammars which have been used successfully in various NLG systems are: (*i*) unification grammars—Functional Unification Grammar [McK85], Functional Unification Formalism [McK90]; (*ii*) Phrase Structure Grammars—Referent Grammar (GPSG with built-in referents) [Sig91], Augmented Phrase Structure Grammar [Sow84]; (*iii*) systemic grammar [Man83]; (*iv*) Tree-Adjoining Grammar [Jos87, Nik95]; (*v*) Generalised Augmented Transition Network Grammar [Sha82].
- **lexicon** – a lexicon entry for each word, containing typical information like part of speech, inflection class, etc.

The formalism used to represent the input semantics also affects the generator's algorithms and its output. For instance, some surface realisation components expect a hierarchically structured input, while others use non-hierarchical representations. The latter solve the more general task where the message is almost free from any language commitments and the selection of all syntactically prominent elements is made both from conceptual and linguistic perspectives. Examples of different input formalisms are: hierarchy of logical forms [McK90], functional representation [Sig91], predicate calculus [McD83], SB-ONE (similar to KL-ONE) [Rei91], conceptual graphs [Nik95].

4.5.3 Related Areas and Techniques

Machine Translation (§4.1) , Text summarisation (§4.4).

Chapter 5

Component Technologies

5.1 Word Clustering

5.1.1 Introduction

Word clustering is a technique for partitioning sets of words into subsets of semantically similar words and is increasingly becoming a major technique used in a number of NLP tasks ranging from word sense or structural disambiguation to information retrieval and filtering.

In the literature, two main different types of similarity have been used which can be characterised as follows:

1. **paradigmatic**, or substitutional, similarity: two words that are paradigmatically similar may be substituted for one another in a particular context. For example, in the context *I read the book*, the word *book* can be replaced by *magazine* with no violation of the semantic well-formedness of the sentence, and therefore the two words can be said to be paradigmatically similar;

2. **syntagmatic** similarity: two words that are syntagmatically similar significantly occur together in 2text. For instance, *cut* and *knife* are syntagmatically similar since they typically co-occur within the same context.

Hereafter we will refer to type 1. similarity as "semantic similarity" tout court while referring to type 2. similarity more loosely as "semantic relatedness". As we will see, the two aspects are obviously interconnected (and in fact they are often not kept apart in the literature) although we find it useful for classificatory purposes to maintain the distinction.

Both types of similarity, computed through different methods, are used in the framework of a wide range of NLP applications.

5.1.2 Survey of Approaches

The methods and techniques for clustering words that we will consider here will be parted according to the kind of similary they take into account:

- semantic similarity;

- semantic relatedness.

Semantic similarity

Typically, semantic similarity is computed either on the basis of taxonomical relationships such as hyperonymy and synonymy or through distributional evidence. The former approach presupposes prior availability of independent hierarchically structured sources of lexico-semantic information such as WordNet [Mil90a]. With the latter approach, the semantic similarity between two words W1 and W2 is computed on the basis of the extent to which their typical contexts of use overlap. The notion of taxonomy-based semantic similarity crucially hinges on words' membership of more general semantic classes. By contrast, distributionally-based semantic similarity rests on the assumption that words entering into the same syntagmatic relation with other words can be seen as semantically similar, although in this case the similarity may be grounded on some covert properties orthogonal to their general semantic superclass.

Taxonomy-based semantic similarity In this section, methods which assess semantic similarity with respect to hierarchically structured lexical resources will be discussed.

In Rada et al. [Rad89] and Lee et al. [Lee93], the assessment of semantic similarity is done with respect to hyperonymy (IS-A) links. Under this approach, semantic similarity is evaluated in terms of the distance between the nodes corresponding to the items being compared in a taxonomy: the shorter the path from one node to another, the more similar they are. Given multiple paths, the shortest path is taken as involving a stronger similarity.

Yet, several problems are widely acknowledged with this approach. First, other link types than hyperonymy can usefully be exploited to establish semantic similarity. For instance, Nagao [Nag92] uses both hyperonymy and synonymy links to compute semantic similarity where higher similarity score is associated with synonymy; with hyperonymy, similarity is calculated along the same lines as Rada et al. [Rad89], i.e. on the basis of the length of the path connecting the compared words. Yet other scholars have attempted to widen the range of relationships on the basis of which semantic similarity can be computed; see, among others, [Nir93b] who also use morphological information and antonyms.

Another problem usually addressed with the path-length similarity method is concerned with the underlying assumption that links in a taxonomy represent uniform distances, which is not always the case. In real taxonomies, it has been noted that the "distance" covered by individual taxonomic links is variable, due to the fact that, for instance, certain sub-taxonomies are much denser than others. To overcome the problem of varying link distances, alternative ways to evaluate semantic similarity in a taxonomy have been put forward.

Agirre and Rigau [Agi96] propose a semantic similarity measure which, besides the length of the path, is sensitive to

- the depth of the nodes in the hierarchy (deeper nodes are ranked closer), and
- the density of nodes in the sub-hierarchies involved (concepts in a denser subhierarchy are ranked closer than those in a more sparse region).

This similarity measure is referred to as "conceptual density" measure.

Resnik [Res95] defines a taxonomic similarity measure which dispenses with the path length approach and is based on the notion of information content. Under his view, semantic similarity between two words is represented by the entropy value of the most informative concept subsuming the two in a semantic taxonomy, the WordNet lexical database (see §3.6.2) in the case at hand. For example, (all senses of) the nouns *clerk* and *salesperson* in WordNet are connected to the first sense of the nouns *employee, worker, person* so as to indicate that *clerk* and *salesperson* are a kind of *employee* which is a kind of *worker* which in turn is a kind of *person*. In this case, the semantic similarity between the words *clerk* and *salesperson* would correspond to the entropy value of *employee* which is the most informative (i.e. most specific) concept shared by the two words.

The information content (or entropy) of a concept c — which in WordNet corresponds to a set of such as *fire*_v_4, *dismiss*_v_4, *terminate*_v_4, *sack*_v_2 — is formally defined as $-log\,p(c)$. The probability of a concept c is obtained for each choice of text corpus or corpora collection K by dividing the frequency of c in K by the total number of words W observed in K which have the same part of speech p as the word senses in c :

$$p(c_p) = \frac{freq(c_p)}{W_p}$$

The frequency of a concept is calculated by counting the occurrences of all words which are potential instances of (i.e. subsumed by) the concept. These include words which have the same orthography and part of speech as the synonyms defining the concept as well as the concept's superordinates. Each time a word W_p is encountered in K, the count of each concepts c_p subsuming W_p (in any of its senses) is increased by one:

$$freq(c_p) = \sum_{c_p \in \{x | sub(xW)\}} count(W_p)$$

The semantic similarity between two words $W1_p, W2_p$ is expressed as the entropy value of the most informative concept c_p which subsumes both $W1_p$ and $W2_p$, as shown below.

$$sim(W1_p, W2_p) = \max_{c_p \in \{x \mid sub(x, W1_p) \wedge sub(x, W2_p)\}} [-\log\, p(c_p)]$$

The specific senses of $W1_p, W2_p$ under which semantic similarity holds is determined with respect to the subsumption relation linking c_p with $W1_p, W2_p$. Suppose for example that in calculating the semantic similarity of the two verbs *fire, dismiss* using the WordNet lexical database we find that the most informative subsuming concept is represented by the synonym set containing the word sense *remove*_v_2. We will then know that the senses for *fire, dismiss* under which the similarity holds are *fire*_v_4 and *dismiss*_v_4 as these are the only instances of the verbs *fire* and *dismiss* subsumed by *remove*_v_2 in the WordNet hierarchy.

Such an approach, combining a taxonomic structure with empirical probability estimates, provides a way of tailoring a static knowledge structure such as a semantic taxonomy onto specific texts.

Distributionally-based semantic similarity The thrust of distributionally-based semantic similarity rests on the idea that the semantic content of a target word T can to a large extent be characterised in terms of (a description of) how T accompanies with other words in a corpus. Two target words are then considered to be semantically similar if they consort with a critical set of common words. Approaches differ in terms of i) how the notion of distribution is formally characterised and ii) what distance metric is adopted to assess the proximity of two distributions.

In the work reported by [Bro91], each target word T_i is characterised by the words that immediately follow it in a text. More formally, the context of T_i is denoted by a vector C(T_i) defined as follows:

$$C(T_i) = <\mid w_1 \mid, \mid w_2 \mid, ..., \mid w_n \mid>$$

where $\mid w_j \mid$ counts how often w_j follows T_i in a reference corpus. For any two such vectors $\mathrm{C}(T_i)$ and $\mathrm{C}(T_k)$, [Bro91] define a distance metric measuring their pairwise distance in terms of (minimal loss of) Average Mutual Information (AMI), where AMI is the value averaging the mutual information of individual word pairs. The idea at the basis of their clustering method is to find groups in which the loss of average mutual information is small. In general, the loss is smaller when the members of the group have similar vectors.

In related work, words are not clustered in terms of frequency counts for the word immediately following T, but rather by looking at a text window of about one thousand words surrounding T. The problem with this and similar approaches is that they are considerably greedy without being completely reliable. Since it is hard to know a priori which word to pull out of the context of a target word to best characterise its sense contextually, blind algorithms are deployed that take into account the largest possible environment around T. Still, this approach can be proved to fail to take into account all relevant contextual evidence, since relevant words are shown by [Gal92a] to be possibly found as far away as 1000 words from the target word.

An attempt to characterise word usage on a sounder linguistic basis is reported by [Per92] who work on examples of verb-object relation found in a corpus. The vector associated with a target noun n_k to characterise its sense contains the (conditional) distribution of verbs for which it serves as direct object in a corpus. The metric used in this study is relative entropy and the representation of the resulting classes is a tree structure of groups.

A further development of this approach is reported in Grefenstette, who characterises the target word in terms of more than one syntactic relation, including subject, object, indirect-object and modifier. For each target noun T_i the resulting representation is a vector $\mathrm{C}(T_i)$ of counts for every other word w_j and each of the four possible syntactic relations between T_i and w_j. [Sek92] apply a simplified version of the same technique to triples with the following structure:

[Headword, MODIFICATION, Modifier]

The triples are automatically extracted from a shallow-parsed corpus. Accordingly two target words are taken to be semantically close if they occur as modifiers of the same headword. The authors, unfortunately, do not use this technique to construct word classes.

[Fed96, Fed97] make the same assumption as [Sek92] but apply it to examples of verb-object and verb-subject relations. Distributionally-based semantic similarity is computed on both terms of these relations. Thus, verbs are taken to be semantically similar if they share some words serving as direct object or subject. By the same token, nouns are semantically similar if they serve as subject and/or object for a shared subset of verbs. A notion of semantic proportional analogy is developed resting on both types of word clustering: this notion gives, for a given target word in context, its closest semantic analogue(s). Unlike all approaches considered so far, whereby word clusters are defined once and for all by averaging out counts for words in context, semantic proportional analogy varies depending on the context. Thus, the same word T selects a different closest analogue depending whether T is considered in the company of the object O_i or the object O_j. This technique is reported to be robust enough to be able to overcome both noisy and sparse data problems.

Semantic relatedness

As we saw above, two words can be considered to be semantically related if they typically co-occur within the same context. In the section on distributionally-based semantic similarity, we showed that word co-occurrence patterns are instrumental for identifying clusterings of semantically similar words. In this section, word co-occurrence patterns will be considered as such together with their use for the resolution of a number of NLP problems.

In the literature, various scholars have suggested that ambiguity resolution (e.g. prepositional phrase attachment) is strongly influenced by the lexical content of specific words ([For82, Mar80] to mention only a few). Yet, for this assumption to be assessed in practice, the necessary information about lexical associations was to be acquired. Recently, an upsurge of interest in corpus-based studies led a number of scholars to collect lexical co-occurrence data from large textual corpora; among them - to mention only a few - [Chu89, Cal90, Hin93, Sma93, Rib94, Tsu92]. Collected patterns were successfully used for different disambiguation purposes. For instance, [Hin93] show how structural ambiguities such as prepositional phrase attachment can be resolved on the basis of the relative strength of association of words, estimated on the basis of distribution in an automatically parsed corpus. Similar results are reported in [Tsu92].

5.1.3 Relevant notions of lexical semantics

All types of word clustering reviewed so far have been widely studied from the theoretical point of view in the field of lexical semantics where it is commonly assumed that the semantic properties of a lexical item are fully reflected in the relations it contracts in actual and potential linguistic contexts, namely on the syntagmatic and paradigmatic axes.

Sense relations are of two fundamental types: paradigmatic and syntagmatic. Paradigmatic relations such as synonymy, hyperonymy/hyponymy, antonymy or meronymy occupy focal positions in discussions of lexical semantics (see, for instance, [Cru86]). They reflect the way reality is categorised, subcategorised and graded along specific dimensions of lexical variation. By contrast, syntagmatic aspects of lexical meaning form a less prominent topic of lexical semantics which in the literature is generally referred to as co-occurrence restrictions. Two different kinds of co-occurrence restrictions are commonly distinguished: selection restrictions, which are defined in very general terms as “any of various semantic constraints reflected in the ability of lexical items to combine in syntactic structures” [Tra94]; collocational restrictions, which are “unusually idiosyncratic or language- specific” (ibidem) restrictions.

Nowadays, a new trend in the lexical semantics literature is emerging which questions the watertight classification of meanings presupposed by traditional approaches to lexical semantics. Under this view, word senses are defined in terms of clear, typical, central cases (called “prototypes”) rather than in terms of their boundaries which very often appear as unclear and undefinable (see [Cru94]).

5.1.4 NLP applications using word clustering techniques

The various word clustering techniques described so far have a large number of potentially important applications, including:

- helping lexicographers in identifying normal and conventional usage;
- helping computational linguists in compiling lexicons with lexico-semantic knowledge;

- providing disambiguation cues for:
  - parsing highly ambiguous syntactic structures (such as noun compounds, complex coordinated structures, complements attachment, subject/object assignment for languages like Italian);
  - sense identification;
- retrieving texts and/or information from large databases;
- constraining the language model for speech recognition and optical character recognition (to help disambiguating among phonetically or optically confusable words).

5.2 Multiword Recognition and Extraction

5.2.1 Introduction

This section examines approaches to multiword recognition and extraction and automatic term recognition (ATR). We will examine linguistic and statistical approaches to ATR. There are no purely statistical approaches to ATR. Statistical approaches come rather from the areas of collocation extraction and IR. The second and third sections examine collocation extraction and the sub-area of IR that relates to ATR, indexing, and how they influence multiword ATR.

5.2.2 Survey of Approaches

Linguistic Approaches

Researchers on multiword ATR seem to agree that multiword terms are mainly noun phrases, but their opinions differ on the type of noun phrases they actually extract. In the overview that follows, most systems rely on syntactic criteria and do not use any morphological processes. An exception is Damerau's work [Dam93].

Justeson and Katz [Jus95] work on noun phrases, mostly noun compounds, including compound adjectives and verbs albeit in very small proportions. They use the following regular expression for the extraction of noun phrases

$$((Adj|Noun)^{+}|((Adj|Noun)^{*}(Noun-Prep)^{?})(Adj|Noun)^{*})Noun \tag{5.1}$$

They incorporate the preposition *of*, showing however, that when *of* is included in the regular expression, there is a significant drop on precision (this drop is too high to justify the possible gains on recall). Their system does not allow any term modification.

Daille et al. [Dai94] also concentrate on noun phrases. Term formation patterns for *base Multi-Word Unit* (base-MWU), consist mainly of 2 elements (nouns, adjectives, verbs or adverbs). The patterns for English are:

1. Adj Noun
2. Noun Noun

while for French

1. Noun Adj

2. Noun Noun
3. Noun de (det) Noun
4. Noun prep (det) Noun

They suggest that MWU of length 3 or more are mostly built from base-MWU using one of the following operations:

1. overcomposition, ([side lobe] regrowth)
2. modification, (interfering [earth (-) station])
3. coordination, (packet assembly/disassembly)

However, their current work deals with base-MWU only.

Bourigault [Bou92] also deals with noun phrases mainly consisting of adjectives and nouns that can contain prepositions, usually *de* and *à*, and hardly any conjugated verbs. He argues that terminological units obey specific rules of syntactic formation. His system does not extract only terms.

In [Dag94a], noun phrases that are extracted consist of one or more nouns that do not belong to a stoplist. A stop list is also used by [Dam93]. Damerau uses morphological analysis for inflectional normalisation.

Statistical Approaches

The most common statistics used, is frequency of occurrence of the potential multiword term ([Jus95, Dag94a, Dai94, Eng94, Per91]).

[Dai94] investigate more statistical scores, since the frequency of occurrence would not retrieve infrequent terms. Several scores have been tested, among which are the one itemized below where:

(w_1, w_2) is the pair of words, a is the frequency of occurrence of both w_1 and w_2
b is the frequency of occurrence of w_1 only
c is the frequency of occurrence of w_2 only, and
d is the frequency of occurrence of pairs not containing neither w_1, nor w_2.

- *Mutual Information*, first defined by [Fan61], and introduced again by [Chu90] as *association ratio*, for measuring the degree of cohesiveness of two words

$$IM(w_1, w_2) = \log \frac{a}{(a+b)(a+c)} \tag{5.2}$$

  Mutual Information tends to extract frozen compounds from the general language.

- Φ^2 *coefficient*, introduced by [Gal91] for the concordance of parallel texts

$$\Phi^2(w_1, w_2) = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)} \tag{5.3}$$

- *Loglike coefficient*, introduced by [Dun93]

$$\begin{aligned} Loglike = {} & a \log a + b \log b + c \log c + d \log d \\ & -(a+b)\log(a+b) - (a+c)\log(a+c) \\ & -(b+d)\log(b+d) - (c+d)\log(c+d) \\ & +(a+b+c+d)\log(a+b+c+d) \end{aligned}$$

- *Diversity*, first introduced by [Sha48], and later proposed for the extraction of frozen compounds or collocations.

As no combination of scores improved the results and they used the Loglike criterion.

[VanE94] uses statistics for finding the pairs of terms from the source and target language. The translations of the terms of the source language are ranked according to the following measure

$$\frac{freq_{local}(t_l|s_l)}{freq_{global}(t_l)} \tag{5.4}$$

where s_l is the source language term, t_l a target language term, and $freq_{local}$, $freq_{global}$, local and global frequencies correspondingly.

[Dam93] uses the difference of two association ratios, one for a corpus consisting of various subdomains and one for a subcorpus of a specific domain.

$$\log \frac{P_s(x,y)}{P(x)P(y)} - \log \frac{P_t(x,y)}{P(x)P(y)} = \log \frac{P_s(x,y)}{P_t(x,y)} \tag{5.5}$$

The probabilities P are estimated by the frequencies normalised by the size of the corpus, t stands for the total corpus and s for the subject sub-corpus.

5.2.3 NLP applications using Multiword Recognition/Extraction

LEXTER

LEXTER [Bou92] takes as input a corpus tagged with a part-of-speech tagger, that consists of 1700 texts from the Research Development Division of Electricité de France, with a total of 1,200,000 words. It works in two stages. During *analysis*, the maximum length noun phrases are extracted, taking into consideration potential "terminological frontiers" (pronouns, conjuction, conjugated verbs, etc.). He suggests that more complicated rules can be added to find the boundaries (e.g. between noun phrases related to the same verb or noun). Some of these rules were intuitively suggested and after being tested for validity on the corpus, were added to LEXTER.

The second stage, *parsing*, extracts substrings from the noun phrases extracted from the previous stage, as additional likely terminological units. These are extracted according their position within the maximum length noun phrases.

[Bou92] argues for the non-necessity of complete syntactical analysis, but the use of a surface grammatical one.

TERMIGHT

TERMIGHT [Dag94b] has been designed as a tool for the extraction of terms for human and machine translation. It consists of a monolingual and a bilingual part.

As a tool, it seems to be more concerned with issues like speed and how easy it is for users. The text is tagged and terms are extracted according to a regular expression and a stop-list.

Termight has a high recall, partly expected since there is no threshold on the frequency of occurrence of the candidate terms, but partly not, since a lot of terms are expected to contain adjectives, which are not treated at the current version of Termight.

As for the bilingual part, TERMIGHT identified the candidate translations of a term, based on word alignment. The candidate translations for each source term are displayed, sorted according to their frequency as translations of the source term.

Daille et al., 1994

Daille et al. work on English and French corpora [Dai94], each consisting of 200,000 words of the field of telecommunications. Only 2-word terms are considered. They are extracted according to morpho-sysntactic criteria, allowing variations on terms. All the variations add up as a list to each term. To the candidate terms extracted, a statistical score (likelihood ratio) is to be applied as an additional filter.

Justeson and Katz, 1995

No parsing or tagging is used by [Jus95], due to their error rate. Instead a lexical database of about 100,000 entries is used, assigning to each word, a part-of-speech, after (basic) morphological analysis is applied if/as needed. At the assignment of the part-of-speech, preference is given to nouns then adjectives then prepositions.

The approach gives preference on recall over precision, unless a high improvement on precision can be gained with a low loss on recall. This is actually the case where the preposition is excluded from the regular expression.

Van der Eijk, 1994

His work is on bilingual terminology (between Dutch and English) [VanE94]. The texts are aligned at sentence level (segments of one or more sentences), and after being tagged, the noun phrases of the form $np \rightarrow w_a^* w_n^+$ are extracted.

The following tables are created: a table that holds the global frequencies of the target language terms and source language term, a table that holds the local frequencies of the target language terms. The candidate translation terms are ranked according to $\frac{freq_{local}(t_l|s_l)}{freq_{global}(t_l)}$, where t_l stands for translation terms, and s_l for source terms. The score should be greater to 1 for the target term to be extracted as a translation to the source term. The assumption is that the translated term is more likely to be more frequent in the target text segments ligned to the source text segments that contain the source term, than in the entire target text.

TERMINO

TERMINO adopts a morphosyntatic approach. The morphological analyser finds the stemming and does the part-of-speech tagging. The syntactic part consists of the parser and the synapsy detector. The parser resolves the remaining lexical ambiguity and gives the syntactic structure. A *synapsy* is a "polylexical unit of syntactic origin forming the nucleous of a noun phrase" ([Dav90]:145). It comprises a noun head that may be preceded by an adjectival phrase or/and may be followed by an adjectival phrase or prepositional phrase complement.

The synapsy detector consists of two parts. The first part, the synapsy builder, is activated each time a noun phrase is encounted by the parser. At this stage the head of the noun phrase is assigned a syntactic structure. the second part, the ysnapsy comparator, applies empirical criteria to filter out some of the noise. This criteria include frequency and category, as well as stop lists for the adjectival modifiers and the position of head.

Term Identification using Contextual Cues–Frantzi & Ananiadou

Till now the information (linguistic and statistical) used for the extraction of terms, was 'internal', ie. coming from the candidate term itself. We see how the incorporation of 'external' information derived from the context of the candidate term. It is embedded to the *C-value* method for automatic term recognition, in the form of weights constructed by statistical characteristics of the context of the candidate term. The environment of words has been previously used for the construction of thesaurus [Gre94]. In that case, words that share the same context are viewed as synonymous. In our case, "extended word units can be freely modified while multiword terms cannot" [Sag90]. We therefore say that terms of the same domain share some common context: the form "shows" of the verb "to show" in medical domains, is almost always followed by a term. So, if we know that "to show" is such a verb (that appears with terms), we can increase the possibility of a string being a term, if it appears with this verb.

The verb of the previous example is carrying information within the medical domain. There are cases where a particular environment that carries such information can be found in more than one domains, like the form "is called" of the verb "to call", that is often involved in definitions of terms in various domains. Our claim is that context, since it carries such information should be involved in the procedure for the extraction of terms. We incorporate context information to the approach of Frantzi & Ananiadou [Fra96a] for the extraction of multiword terms in a fully automatic way[1]. The corpus used is tagged. From the tagged corpus, the n-grams using the following regular expression are extracted $(Noun|Adjective)^{+}Noun$ The choice of the regular expression affects the precision and recall of the Our choice is a compromise between the two. For these n-grams, *C-value*, a statistical measure for the extraction of terms, based on the frequency of occurrence, and "sensitive" to *nested terms*[2] is evaluated [Fra96a].

According to [Fra96b], the *C-value* integrates the parameter of the length of the n-gram. The length was used as a parameter when *C-value* was applied for the extraction of collocations [Fra96b]. Its weight ie weakened as shown in 5.6, where:

a is the examined n-gram,
$|a|$ the length, in terms of number of words, of a,
$f(a)$ the frequency of a in the corpus,
b_i the candidate extracted terms that contain a,
$c(a)$ the number of those candidate terms.

[1] We have to point out here that in a fully automatic approach (compared to a manual), issues of compromise between the precision and recall come at play.

[2] Nested terms are sub-strings of other terms.

$$C\text{-}value'(a) = \begin{cases} \log_2 |a| \cdot f(a) & |a| = max, \\ \log_2 |a| \cdot (f(a) - \frac{1}{c(a)} \sum_{i=1}^{c(a)} f(b_i)) & \\ & otherwise. \end{cases} \tag{5.6}$$

The output of the application of *C-value* on these n-grams is a list of potential terms. The higher the *C-value* of an n-gram, the more possible for it to be a term.

From that list, the higher ranked terms are considered for the context evaluation. By context, we mean the verbs, adjectives, nouns, that appear with the candidate term. We attach a weight to those verbs, adjectives, nouns. Three parameters are considered for the evaluation of these weights: the number of candidate terms the word (verb, adjective, noun) appeared with, its frequency as a context word, and its total frequency in the corpus. The above are combined as shown in 5.7, where:

w is the noun/verb/adjective to be assigned a weight,
n the total number of candidate terms considered,
$t(w)$ the number of candidate terms the word w appears with,
$ft(w)$ w's total frequency appearing with candidate terms,
$f(w)$ w's total frequency in the corpus.

$$Weight(w) = 0.5 \cdot (\frac{t(w)}{n} + \frac{ft(w)}{f(w)}) \tag{5.7}$$

For each of the n-grams of the previously created list, its context words, (verbs, adjectives or nouns) are extracted. These context words have from the previous stage a weight assigned to them (that can be 0 if the word was not met when the weights were assigned). The sum of these weights will give the context weight $wei(a)$ for each n-gram, as shown in 5.8 where C_a is the context of the n-gram a.

$$wei(a) = \sum_{b \epsilon C_a} weight(b) + 1 \tag{5.8}$$

The n-grams will be re-ranked as shown in 5.9 where:
a is the examined n-gram,
$C\text{-}value'(a)$, the previously calculated $C\text{-}value'(.)$,
$wei(a)$, the context weight for a,
N, the size of the corpus in terms of number of words.

$$NC\text{-}value(a) = \frac{1}{log(N)} \cdot C\text{-}value'(a) \cdot wei(a) \tag{5.9}$$

Rank Xerox Multiword lexeme recognition and extraction

To recognize and extract Multi-word lexeme (MWL) we use the finite state technology ([Kar93, Kar92] which provides an efficient and fast implementation environment.

MWL recognition Multi-word expressions cannot be properly understood – or even translated – if they are not recognized as complex lexical units. We call such expressions *multi-word lexemes (MWL)*. These include *idioms* (to rack one's brains over sth), *proverbial sayings* (birds of a feather flock together), *phrasal verbs* (to come up with), *lexical and grammatical collocations* (to make love, with regard to), *compounds* (on-line dictionary).

Some MWLs always occur in exactly the same form and can therefore be easily recognised by their lexical pattern. This is the case for expressions like *footloose and fancy free* or *out of the blue*. However, most MWLs allow different types of variation and modification[3]. To be able to recognize such MWLs in a text, occurrences deviating from the standard or base form of the MWL have to be identified, e.g. different inflections, word orderings and modified uses. For example, in *casser sa pipe* (to kick the bucket), no plural is possible for the noun, the verb cannot be replaced by its near-synonym *briser*, nor can the phrase be passivised without losing its idiomatic meaning. Yet, the verb itself can be inflected.

Simple string matching methods are too weak to identify MWLs because most of them are not completely fixed. Besides, the variations they can undergo are, in most cases, lexicographically not well defined. A dictionary entry usually provides the reader with one form of the expression – not necessarily the base or canonical form –, giving no details about allowed variations, except sometimes lexical variants. This type of missing information can be stated with local grammar rules which have more general expressiveness than traditional descriptions.

Local grammar rules describe restrictions of MWLs compared to general rules by implicitly stating allowed variations of the MWL compared to the default case of a completely fixed MWL. In the default case, all restrictions apply, i.e. no variation at all is allowed, and the MWL is represented by the surface form of all lexical components in a fixed order. Violations to standard grammatical rules, e.g. missing constituents or agreement violations, need not be stated explicitly, though if necessary they can be expressed to distinguish the idiomatic from a literal use of the lexical pattern. To write the local grammar rules we use the two-level formalism IDAREX (**ID**ioms **A**s **R**egular **EX**pressions) developed as part of the FSC finite state compiler at Rank Xerox Research Centre [4]. The local grammar rules we write are restricted to cover at most sentence length patterns. They are formulated as generally as possible, allowing for overgeneration. Although more specific and restrictive rules could be written, this is unnecessary because we assume that there is no ill-formed input. Indeed, it does not matter if the rules allow more variations than the ones that will actually appear in texts as long as idiomatic and literal uses can be distinguished. For instance, as long as we are not concerned with the semantic representation of MWLs, the local grammar rule for the French expression *peser dans la balance* accepts semantically correct phrases such as *peser lourd dans la balance* or *peser énormément dans la balance*, but also the semantically ill-formed *peser *ardemment dans la balance*. More generally, local grammar rules are also useful for syntactic parsing, e.g. by describing complex adverbials such as dates *Le lundi 21 aout au matin*[5] or any other expressions that do not follow the general syntax. In many cases the syntactic parser would just fail because it would not be able to analyse properly the multi-word expression embedded in a larger phrase. For instance in German, the general syntax states that a determiner should precede any count noun. This rule is infringed in the MWL *von Haus aus* (originally).

Regarding the techniques we use, the two-level morphological approach based on finite state technology together with the IDAREX formalism, have the advantage of providing us with a compact representation. As we saw, we can define general variables, such as "any adverb" (ADV) or more specific morphological variables, such as "only verbs in the third

[3]By *variation* we understand an exchange or syntactic re-structuring of components; by *modification* we understand the addition of modifying words to the MWL.

[4]For a more detailed description of the formalism see [Kar93, Tap94, Seg95].

[5]For a related treatment of French dates see [Mau93].

person singular" (Vsg3). This relieves the lexicographer from the burden of explicitly listing all the possible forms. Functional variables provide a means to formulate generalizations about patterns that can occur for a whole class of MWLs. Besides, the two levels enable us to express facts either with the surface form or with the lexical form. Therefore, when we want to say that a given form is fixed, we just have to use the surface form without bothering with all the features on the lexical side.

In this technology, operations like addition, intersection, substraction and composition are allowed on the networks generated from regular expressions. Although we have not used this possibility in our work on local grammars yet, it is very powerful. For instance, if we are concerned about the semantics of a MWL and want to be more restrictive with the rules, we can build new regular expressions and substract the resulting networks from the one we already built. Such additional regular expressions would, for example, express facts about the compatibility of semantic classes of adjectives and nouns.

MWL extraction Much of the terminology found in a corpus is composed of noun phrases. One extension of our NLP suite is a noun phrase extraction step which can follow part-of-speech tagging [Sch96]. In order to perform this step, transducers have been compiled from finite-state expressions which are basically grammar rules describing the contour and patterns of noun phrases for each language for which a lexicon and tagger are made. The patterns can include surface forms as well as part-of-speech tags. When these transducers are applied to tagged text, noun phrase boundaries are inserted. For example, consider the Dutch text:

> *De reparatie- en afstelprocedures zijn bedoeld ter ondersteuning voor zowel de volledig gediplomeerde monteur als de monteur met minder ervaring. (The repair and adjustment procedures are meant to aid the fitter who has completed his degree work as well as the less experienced fitter.)*

After part-of-speech tagging, the noun phrase transducers will recognize and isolate the following noun phrases: *reparatie-en afstelprocedures, ondersteuning, volledig gediplomeerde monteur, monteur and ervaring.* The current noun phrase mark-up was designed basically for terminology extraction from technical manuals. It covers relatively simple noun phrase detection, i.e. some constructions such as relative clauses are not included.

Because one can easily add a new regular expression to handle more constructions, more elaborate patterns including verbs can be extracted. The same automatic means have been used to extract collocations from corpora, in particular, support verbs for nominalizations. In English, an example of proper support verb choice is *one makes a declaration* and not *one does a declaration. Make* is said to support the nominalization declaration which carries the semantic weight of the phrase. We used NLP suites followed by syntactic pattern matching slightly more complicated than the noun phrase extractors of the previous section, in order to extract verbal categorization patterns for around 100 nominalizations of communication verbs in English and French [Gre96].

Similar approaches are used to identify more sepcific items such as: dates, proper names. They use a combination of regular expressions as described above and specific lexical ressources including, for instance, semantic information.

5.3 Word Sense Disambiguation

5.3.1 Introduction

One of the first problems that is encountered by any natural language processing system is that of lexical ambiguity, be it syntactic or semantic. The resolution of a word's syntactic ambiguity has largely been solved in language processing by part-of-speech taggers which predict the syntactic category of words in text with high levels of accuracy (for example [Bri95]). The problem of resolving semantic ambiguity is generally known as word sense disambiguation and has proved to be more difficult than syntactic disambiguation.

The problem is that words often have more than one meaning, sometimes fairly similar and sometimes completely different. The meaning of a word in a particular usage can only be determined by examining its context. This is, in general, a trivial task for the human language processing system, for example consider the following two sentences, each with a different sense of the word *bank*:

- The boy leapt from the bank into the cold water.
- The van pulled up outside the bank and three masked men got out.

We immediately recognise that in the first sentence *bank* refers to the edge of a river and in the second to a building. However, the task has proved to be difficult for computer and some have believed that it would never be solved. An early sceptic was Bar-Hillel [Bar64] who famously proclaimed that "sense ambiguity could not be resolved by electronic computer either current or imaginable". He used the following example, containing the polysemous word *pen*, as evidence:

Little John was looking for his toy box.
Finally he found it.
The box was in the *pen*.
John was very happy.

He argued that even if *pen* were given only two senses, 'writing implement' and 'enclosure', the computer would have no way to decide between them. Analysis of the example shows that this is a case where selectional restrictions fail to disambiguate "pen", both potential senses indicate physical objects in which things may be placed (although this is unlikely in the case of the first sense), the preposition *in* may apply to both. Disambiguation, in this case, must make use of world-knowledge; the relative sizes and uses of pen as 'writing implement' and pen as 'enclosure'. This shows that word sense disambiguation is an AI-complete problem[6].

However, the situation is not as bad as Bar-Hillel feared, there have been several advances in word sense disambiguation and we are now at a stage where lexical ambiguity in text can be resolved with a reasonable degree of accuracy.

The Usefulness of Word Sense Disambiguation We can distinguish "final" and "intermediate" tasks in language processing: final tasks are those which are carried out for their own usefulness examples of final tasks are machine translation, automatic summarisation and

[6] A problem is AI-complete if its solution requires a solution to the general AI problems of reasoning about arbitrary world knowledge.

information extraction; intermediate tasks are carried out to help final tasks, examples are part-of-speech tagging, parsing, identification of morphological root and word sense disambiguation, these are tasks in which we have little interest in their results *per se*.

The usefulness of intermediate tasks can be explored by looking at some of the final tasks with which they are likely to help. We shall now examine three tasks which it has been traditionally assumed word sense disambiguation could help with: information retrieval, machine translation and parsing.

Information Retrieval It has often been thought that word sense disambiguation would help information retrieval. The assumption is that if a retrieval system indexed documents by senses of the words they contain and the appropriate senses in the document query could be identified then irrelevant documents containing query words of a different sense would not be retrieved. Strzalkowski [Str95] has recently found evidence that NLP may help in information retrieval. However, other researchers have found that word sense disambiguation leads to little, if any, improvement in retrieval performance. Krovetz and Croft [Kro92, Kro97] manually disambiguated a standard IR test corpus and found that a perfect word sense disambiguation engine would improve performance by only 2%. Sanderson [San94b] performed similar experiments where he artificially introduced ambiguity into a test collection, he found that performance was only increased for very short queries (less than 5 words). The reason for this is that the statistical algorithms often used in information retrieval are similar to some approaches to word sense disambiguation[7] and the query words in long queries actually help to disambiguate each other with respect to documents. Sanderson also dicovered that "the performance of [information retrieval] systems is insensitive to ambiguity but very sensitive to erroneous disambiguation" (p 149).

Machine Translation In contrast, researchers in machine translation have consistently argued that effective word sense disambiguation procedures would revolutionise their field. Hutchins and Sommers [Hut92] have pointed out that there are actually two types of lexical semantic ambiguity with which a machine translation system must contend: there is ambiguity in the source language where the meaning of a word is not immediately apparent but also ambiguity in the target language when a word is not ambiguous in the source language but it has two possible translations.

Brown et. al. [Bro91] constructed a word sense disambiguation algorithm for an English-French machine translation system. This approach performed only as much disambiguation as was needed to find the correct word in the target language (ie. it resolves only the first type of ambiguity in machine translation). Brown found that 45% of the translations were acceptable when the disambiguation engine was used while only 37% when it was not. This is empirical proof that word sense disambiguation is a useful intermediate task for machine translation.

Parsing Parsing is an intermediate task used in many language processing applications and accurate paring has long been a goal in NLP. It seems that if the semantics of each lexical item were known then this could aid a parser in constructing a phrase structure for that sentence. Consider the sentence "The boy saw the dog with the telescope.", which is often used as an example of the prepositional phrase attachment problem. A

[7] In particular the *bag of words* approaches mentioned below

parser could only correctly attach the prepositional phrase to the verb "saw" by using semantic world knowledge, and it seems that semantic tags would provide part of that knowledge. Unfortunately there seems to have been little empirical research carried out on the usefulness of word sense disambiguation for parsing.

We can see then that word sense disambiguation is likely to be of benefit to several important NLP tasks, although it may not be as widely useful as many researchers have thought. However, the true test of word sense disambiguation technology shall be when accurate disambiguation algorithms exist, we shall then be in a position to experiment whether or not they add to their effectiveness.

5.3.2 Survey of Approaches to Word Sense Disambiguation

It is useful to distinguish some different approaches to the word sense disambiguation problem. In general we can categorise all approaches to the problem into one of three general strategies: *knowledge based, corpus based and hybrid.* We shall now go on to look at each of these three strategies in turn.

Knowlegde based

Under this approach disambiguation is carried out using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary, thesaurus or it may be hand-crafted. This is one of most popular approaches to word sense disambiguation and amongst others, work has been done using existing lexical knowledge sources such as WordNet [Agi96, Res95, Ric95, Sus93, Voo93], LDOCE [Cow, Gut91], and Roget's International Thesaurus [Yar92].

The information in these resources has been used in several ways, for example Wilks and Stevenson [Wil97], Harley and Glennon [Har97] and McRoy [McR92] all use large lexicons (generally machine readable dictionaries) and the information associated with the senses (such as part-of-speech tags, topical guides and selectional preferences) to indicate the correct sense. Another approach is to treat the text as an unordered *bag of words* where similarity measures are calculated by looking at the semantic similarity (as measured from the knowledge source) between all the words in the window regardless of their positions, as was used by Yarowsky [Yar92].

Corpus based

This approach attempts to disambiguate words using information which is gained by training on some corpus, rather that taking it directly from an explicit knowledge source. This training can be carried out on either a *disambiguated or raw corpus*, where a disambiguated corpus is one where the semantics of each polysemous lexical item is marked and a raw corpus one without such marking.

Disambiguated corpora This set of techniques requires a training corpus which has already been disambiguated. In general a machine learning algorithm of some kind is applied to certain features extracted from the corpus and used to form a representation of each of the senses. This representation can then be applied to new instances in order to disambiguate them. Different researchers have made use of different sets of features, for example [Bro91] used local collocates such as first noun to the left and right, second word to the left/right and so on. However, a more common feature set used by [Gal92a] is to take all the words in a window of $\pm n$ words around the ambiguous words, treating the context as an unordered *bag of words*.

Another approach is to use Hidden Markov Models which have proved very successful in part-of-speech tagging. Realizing of course that semantic tagging is a much more difficult problem than part-of-speech tagging, [Seg97] decided nonetheless to perform an experiment to see how well words can be semantically disambiguated using techniques that have proven to be effective in part-of-speech tagging. This experiment involved the following steps:

1. deriving a *lexicon* from the WordNet data files which contains all possible semantic tags for each noun, adjective, adverb and verb. Words having no semantic tags (determiners, prepositions, auxiliary verbs, etc.) are ignored.

2. constructing a *training corpus* and a *test corpus* from the semantically tagged Brown corpus (manually tagged by the WordNet team) by extracting tokens for the HMM bigrams.

3. computing a HMM model based on the training corpus, runnig the tagger on the test corpus and comparing the results with the original tags in the test corpus.

The general problem with these methods is their reliance on disambiguated corpora which are expensive and difficult to obtain. This has meant that many of these algorithms have been tested on very small numbers of different words, often as few as 10.

Artificial Corpora A consequence of the difficulty in obtaining sense tagged corpora has meant that several researchers have found innovative ways of creating *artificial corpora* which contain some form of semantic tagging.

The first type of artificial corpus which has been used extensively is the parallel corpus. A bilingual corpus consists of two corpora which containing the same text in different languages (for example one may be the translation of the other, or they may have been produced by an organisation such as the United Nations or the European Union who routinely transcript meetings in several languages). Sentence alignment is the process of taking such a corpus and matching the sentences which are translations of each other and several algorithms exist to carry this out with a high degree of success (eg. [Cat89], [Gal92b]). A bilingual corpus which has been sentence aligned becomes an aligned parallel corpus. This is an interesting resource since it consists of many examples of sentences and their translations. These corpora have been made use of in word sense disambiguation (see [Bro91] and [Gal92b]) by taking words with senses which translate differently across languages. They used the Canadian Hansard, the proceedings of the Canadian Parliament which is published in both French and English, and words such as “duty” which translates as “devoir” in the sense of 'moral duty' and “droit” when it means 'tax'. They took all the sentence pairs with “duty” in the English sentence and split then into two groups, roughly corresponding to senses, depending upon which word was in the French sentence of the pair. In this way a level of disambiguation suitable for a Machine Translation application could be tested and trained without hand-tagging.

There are two ways of creating artificial sense tagged corpora. The first way is to disambiguate the words by some means, as happens in the case of parallel corpora; the other approach is to add ambiguity to the corpus and have the algorithm attempt to resolve this ambiguity to return to the original corpus. Yarowsky [Yar93] used this method by creating a corpus which contained "pseudo-words". These are created by choosing two words ("crocodile" and "shoes" for the sake of argument) and replacing each occurance of either with their concatenation ("crocodile/shoes").

Raw Corpora It is often difficult to obtain appropriate lexical resources (especially for texts in a specialised sublanguage), and we have already noted the difficulty in obtaining disambiguated text for supervised disambiguation. This lack of resources has led several researchers to explore the use of unannotated, raw, corpora to perform *unsupervised disambiguation*. It should be noted that unsupervised disambiguation cannot actually label specific terms as a referring to a specific concept: that would require more information than is available. What unsupervised disambiguation can achieve is word sense *discrimination*, it clusters the instances of a word into distinct categories without giving those categories labels from a lexicon (such as LDOCE sense numbers or WordNet synsets).

An example of this is the dynamic matching technique [Rad96] which examines all instances of a given term in a corpus and compares the contexts in which they occur for common words and syntactic patterns. A similarity matrix is thus formed which is subject to cluster analysis to determine groups of semantically related instances of terms.

Another example is the work of Pedersen [Ped97] who compared three different unsupervised learning algorithms on 13 different words. Each algorithm was trained on text with was tagged with either the WordNet or LDOCE sense for the word but the algorithm had no access to the truce senses. What it did have access to was the number of senses for each word and each algorithm split the instances of each word into the appropriate number of clusters. These clusters were then mapped onto the closest sense from the appropriate lexicon. Unfortunately the results are not very encouraging, Pedersen reports 65-66% correct disambiguation depending on the learning algorithm used. This result should be compared against that fact that, in the corpus he used, 73% of the instances could be correctly classified by simply choosing the most frequent sense.

Hybrid approaches

These approaches can be neither properly classified as knowledge or corpus based but use part of both approaches. A good example of this is Luk's system [Luk95] this uses the textual definitions of senses from a machine readable dictionary (LDOCE) to identify relations between senses. He then uses a corpus to calculate mutual information scores between these related senses in order to discover the most useful. This allowed Luk to produce a system which used the information in lexical resources as a way of reducing the amount of text needed in the training corpus.

Another example of this approach is the unsupervised algorithm of Yarowsky [Yar95]. This takes a small number of seed definitions of the senses of some word (the seeds could be WordNet synsets or definitions from some lexicon) and uses these to classify the "obvious" cases in a corpus. Decision lists [Riv87] are then used to make generalisations based on the corpus instances classified so far and these lists are then re-applied to the corpus to classify more instances. The learning proceeds in this way until all corpus instances are classified. Yarowsky reports that the system correctly classifies senses 96% of the time.

5.3.3 Relevant notions of lexical semantics

5.3.4 NLP applications using WSD Techniques

Semantic disambiguation in Locolex

One application of semantic tagging is in the framework of an intelligent on line dictionary lookup such as LocoLex [Bau95]. LocoLex is a tool that has been developed at RXRC and which looks up a word in a bilingual dictionary taking the syntactic context into account. For instance, in a sentence such as *They like to swim* the part of speech tagger in LocoLex determines that *like* is a verb and not a preposition. Accordingly, the dictionary lookup component provides the user with the translation for the verb only. LocoLex also detects multi-word expressions [8]. For instance, when *stuck* appears in the sentence *my own parents stuck together* the translation displayed after the user clicks on *stuck* is the one for the whole phrase *stick together* and not only for the word *stick*.

Currently LocoLex is purely syntactic and cannot distinguish between the different meanings of a noun like *bark*. If, in addition to the current syntactic tags, we had access to the semantic tags provided by WordNet for this word (natural event or plants) and if we were able to include this label in the online dictionary, this would improve the bilingual dictionary access of Locolex even further.

Current bilingual dictionaries often include some semantic marking. For instance looking at the OUP-Hachette English French dictionary, under *bark* we find the label Bot(anical) attached to one meaning and the collocator (of dog) associated with the other one. It is possible that some type of automated matching between these indications and the WordNet semantic tags[9] would allow the integration of a semantic tagger into LocoLex.

Using only existing dictionary labels might still not be completely satisfying for machine translation purpose. Indeed looking back at the example *my own parents stuck together*, even if we retrieved the multi-word expression meaning it will be difficult to decide which translation to choose with existing dictionary indications[10]. For instance for *stick together* the Oxford-Hachette English French dictionary gives:

```
stick together
1. (become fixed to each other)
(pages) se coller
2. (Coll) (remain loyal)
se serrer les coudes (Fam) =EAtre solidaire
3. (Coll) (not separate)
rester ensemble
```

One could go one step further by using the sense indicators in the Oxford-Hachette dictionary: (become fixed to each other) (remain loyal) (not separate). These sense indicators are remains of definitions and often turn to be synonyms of the entry. They are about 27.000 of them and building a HMM tagger for them is not possible. We can still reduce their number by grouping them into classes of higher level. For instance we could group together : *old man, old person, young man, old woman*, etc. under *person*. Then we can use a statistical method

[8] Multi-words expressions including: idiomatic expression such as *to sweep something under the rug*, phrasal verbs *to spit up*, compounds *warning light*.

[9] Or some other derived tag set.

[10] Especially considering that WordNet provides only two senses of *stick together*: S35 and S41.

such as the one described in [Yar95] to choose the most appropriate meaning in context. How to evaluate the result on large corpora is still pending.

Another step can be achieved by using the verb subcategorization frame together with selectional restriction for its arguments and shallow parsing.
At RXRC we have developped a shallow parser for French (Ait-Moktar and Chanod, 1997). The advantages of using shallow parsing are many. Among them:

Shallow parsing is robust and based on linguistic knowledge. Therefore it handles any kind of text. The system is fast (about 150 words per second)

Our shallow parser is incremental. Results can be checked at each step insuring that the input of the next step is the desired one.

Our shallow parser provides in a reliable way functional information such as subject and object [11].

If we consider the example:

J'ai assisté à la réunion de ce matin.

The verb *assister* has the following subcategorisation frames:

- Two arguments: a subject (noun phrase) and a prepositional phrase. The prepositional phrase is introduced by the preposition *à*. The arguments are mandatory.
- Two arguments: a subject (noun phrase) and an object (noun phrase). The arguments are mandatory.

When we parse the above sentence with a shallow parser we get the following structure:

```
[VC  [NP j' NP]/SUBJ :v ai assist=E9 v: VC]  [PP =E0 la r=E9union PP]
[PP de ce matin PP] .
```

from the above parse we learn that there is a subject, it is a noun phrase, there are two prepositional phrases, one of them introduced by the preposition *à*. Therefore we can select the meaning associated with only the first sub categorization frame for *assister*, *to attend a meeting.*

Using just verb subcategorisation frame moved us one step further in the process of semantic selection.

Still, in some cases even subcategorisation frame is not enough and one needs to have access to ontologies in order to express selectional restriction. In other words one needs to have more information regarding the semantic type of the verb argument. Consider now the sentence:

[11] The accuracy of the system for Subject assignment is 93% success for newspapers and 97% success for technical texts.

Je bouche le trou avec du ciment (I fill the hole with concrete)

The shallow parser creates the following analysis:

```
[VC  [NP je NP]/SUBJ :v bouche v: VC]  [NP le trou NP]/OBJ  [PP avec du
ciment PP]
```

If we look in the Oxford Hachette biligual dictionary we have all the meanings below (in this case translations) associated with the transitive use of the verb *boucher*:

boucher vtr

mettre un bouchon à to cork, to put a cork on [bouteille];

obstruer [to block tuyau, passage, aération, fen=EAtre, vue];
en encrassant to clog (up) [gouttière, artère, pore];
en comblant to fill [trou, fente]; ...
lit to fill the holes;
fig (*dans un budget, une soirée*) to fill the gaps
....

This example illustrates the representation and the treatment of collocates. The underlined words are encoded in the Oxford-Hachette as the object collocate, in other words they indicate what is the semantic type of the object. Looking at the analysis produced by the shallow parser we know that the head of the object is the word '*trou*. In this case just a simple pattern matching with the list of possible object collocates will tell us which meaning to choose and therefore which translation (*to fill*).
But if we had the sentence *boucher les fissures* we would need to know that *trou* and *fissure* are member of the same semantic cluster and that the verb *boucher* accepts this cluster as typical object in one of its meaning

Lexical Disambiguation in Eurotra

Eurotra's Interface Structure (IS) representation is not a semantic representation. However, in a number of areas, attempts were made to provide an interlingual semantic solution to the translation problem. Essentially the semantic information encoded in E-dictionaries is used for disambiguation purposes. These include (i) **structural ambiguity**, (ie. argument modifier distinction, essentially PP-attachtment) and (ii) **lexical ambiguity** in lexical transfer, that is collocations (restricted to verb support constructions) and polysemy.

Lexical ambiguity: Collocations (Verb Support) Treatment of collocations within Eurotra is limited to the support verb constructions. Support verbs in Eurotra are associated to predicative nouns. A predicative noun is a noun which has an argument structure with, at least, a deep subject argument. A predicative noun can appear in a structure of the kind: Det-def N [SUBJ-GEN Na] (Prep Nb) (Prep Nc)

The SUBJ-GEN stands for the subjective genitive expressed in English by the preposition 'of' or as a preposed genitive:

the attack of the enemies

the enemies' attack

It is understood that for every predicative noun there is a non-empty set of of support verbs (Vsup) such that the following conditions hold:

- the SUBJ-GEN of the predicative N is the subject of the correponding Vsup.

  the enemies made an attack

- The predicative noun may take a relative clause contaning the support verb:

  the attack which the enemies made

Following [Gro81], in Eurotra the main function of the support verbs is to provide with tense, number and person information. Support verbs are understood to be semantically contentless. Also following Gross, in Eurotra support verbs are distinguished by means of their aktionsart. Here is the only case where a 'euroversal' semantic classification is followed:

neutral: John HAS influence over Mary

inchotive: John GAINS influence over Mary

durative: John RETAINS influence over Mary

iterative: John REPEATED his attacks against Mary

terminative: John LOST influence over Mary

Support verb constructions are given a 'raising verb' representation. The entry for a predicative nouns such as 'influence' bears information about the kinds of support verbs it demands:

```
eg: 'influence':
{cat=n, predic=yes, is_frame=arg_12, pform_arg1=over, svneut=have,
svincho=gain, svdur=keep, avterm=lose, sviter=none}
```

Lexical Ambiguity: polysemy Lexical Semantic Features (LSF) are present at IS because they are used to distinguish readings in analysis and in generation. There are two different approaches to LSFs, (i) Universal Semantic Feature theory, and (ii) Restricted Semantic Feature theory. The former provides with a set of universal hierarchically organized fine-grained features. The latter provides with a set of language specific features.

In the USF approach, attribute/value pairs must be assigned in identical way in all monolingual distionaries. Lexical transfer is performed from lexical unit to lexical unit with unification of all semantic features:

```
eg: {lu=computer, human=no}     <=>     {lu=ordinateur, human=no}
    {lu=computer, human=yes}    <=>     {lu=elaboratore, human=yes}
```

Eurotra legislation follows the RSF approach. So, lexical semantic features are language specific. They are not part of the information that is transferred but they are an important part of the dictionary. Lexical semantic features are essentially monolingual and each language group is free to choose them. In Spanish, these include:

```
semarg1/2/3/4 = parti,med,tiem,coord,prop,cual,abt,rel,co_em,res,act,n_ag,pers,
      org,lug,no_hum,instr,masa,fu,nat,parte,todo,ent,sem,no_sem,abs,esc,
      no_parti, sit,est,no_prop,din,conc,anim,ag,agde,hum,no_lug,ind,
      no_anim,cont,art,nil.
```

There is one entry for every reading. Readings are distinguished by means of a identificational feature (isrno). Lexical transfer is performed from reading to reading by means of reading numbers. LSFs are not transferred. There may be different readings in all levels:

```
at the ECS level:'walk'
{lu=walk, ecsrno=1, cat=v...}
{lu=walk, ecsrno=2, cat=n...}

at the IS level: 'adopt'
{lu=adopt, isrno=1, rsf_human_of_arg1=yes, rsf_human_of arg2=yes...}
{lu=adopt, isrno=2, rsf_human_of_arg1=yes, rsf_human_of arg2=no...}
```

For the 'computer' example above, lexical transfer goes from reading to reading, despite the semantic features involved in each entry:

```
eg: {lu=computer, isrno=1, human=no}    <=> {lu=ordinateur, isrno=1, human=ho}
    {lu=computer, isrrno=2, human=yes}  <=> {lu=elaboratore, isrrno=1, human=yes}
```

A note on Eurotra-D: Eurotra allows further subtyping IS with semantic relations (SR) [Ste88]. The 'Development of the EUROTRA-D System of Semantic relations' by Steiner, Eckret, Weck and Winter aims at automizing to a large extend the compilation of transfer dictionaries in order to reduce the work and to be more independent of bilingual knowledge. EUROTRA-D suggests a detailed monolingual classification of verbs according to a classification system based on syntactic and semantic criteria. The idea is that the resulting description of verbs makes automatic mapping from one lexeme in one language onto a lexeme in another language. The encoding of different readings of a verb is used to partly automize the compilation of transfer dictionaries.

Essentially, EUROTRA-D assumes four major semantic Predicate Types and a set of subtypes:

- Relational: locational, associative, classificatory, identifying, existential
- Mental:Precessor oriented, phenomenon oriented, two phenomena, phenomenon only
- Communication: Sender+message, sender+receiver
- Action: natural phenomenon, agent only, affected only,attribuant only, agent centered, affected centered

Each Predicate Type assigns a unique set of SRs to its subcategorized elements. Thus, for instance, the SRs assigned by Communication type predicates include: message, sender, receiver and promt.

SRs serve to distinguish readings, so two readings are different iff their associated sets of SRs are distinct in either, number, type or canonical order (the order of SRs is given by the order of their syntactic realization).

SRs are used for lexical disambiguation/transfer purposes. At IS representation, the ARG2 of move like verbs occur with different syntactic and semantic constituents:

The director moved the firm

The director moved to Europe

The firm moved to Europe

For some languages, also the ARG1 has different semantic 'constituens':

The director moved

The firm moved

At IS level, we would have three readins which hardly distinguish examples above:

move1= ...frame=ARG1,ARG2,ARG3

move2= ...frame=ARG1,ARG2

move3= ...frame=ARG1

Enriched IS representations allow to obtain a hihger number of readings:

```
move_1= frame=  ARG1(SR=3RDPARTYAGENT),ARG2(SR=ATTRIBUANT),
                ARG3=(SR=LOCATION)
                'the director moved the firm to Europe'

move_2= frame=  ARG1(SR=3RDPARTYAGENT),ARG2(SR=ATTRIBUANT),
                'the director moved the firm'

move_3= frame=  ARG1(SR=AGENT_ATTRIBUANT), ARG2(SR=LOCATION)
                'the director moved to Europe'

move_4= frame=  ARG1(SR=AFFECTED-ATTRIBUANT), ARG2(SR=LOCATION)
                'the firm moved to Europe'

move_5= frame=  ARG1(SR=AFFECTED-ATTRIBUANT)
                'the firm moved'
```

ET-10/75 Project: (collocations) The ET-10/75 project centers on the use of Mel'cuk's Lexical Functions (LFs) as interlingua representations in MT dictionaries for a subset of collocations (adj-noun and verb-noun). The project is implemented in ALEP.

Collocations constitute an important challange for MT. There is no simple way of mapping collocations of one language onto equivalent expressions of another.

To cope with these mismatches we can (i) add information in bilingual dictionaries so that, for example, English 'pay' translates to Spanish 'prestar (eg, 'pay attention' 'prestar atención). The problem here is that we have to specify in which contexts 'pay' and 'prestar' are equivalent. (ii) we can list collocations giving the corresponding expression in the target language. (iii) we can adopt an interlingua approach.

Following Mel'cuk, the ET-10/75 project suggests for a interlingua approach to collocations. Mel'cuk works in Meaning Text theory and provides with Extended Combinatory Dictionaries (Russian and French).

ET-10/75 has a wide notion of collocation. Collocations are compositional and semantically transparent (the meaning of the whole reflects the meaning of the parts), 'frequent', allow material in between (eg, 'pay little attention', cannot be syntactically or semantically predicted, may allow syntactic 'processes' (eg. passivization, extractions etc) ...

The kind of Collocations they deal about are (lexical collocations):

N of N: flock of SHEEPS

Adj N: narrow SCOPE

V N: commit MURDER

The syntactic relations in collocations involve head/modifier (N-of-N and Adj-N) or head/argument relations (V-N).

In both cases, the N is the base and selects the collocate. This means that for head/modifier collocations the head (N) selects the collocate/modifier and for head/argument collocations the argument (N) selects the head/collocate.

Lexical Functions (Mel'cuk) LF are used to systematically describe certain semantic and collocational relations existing between lexemes. They apply at the deep syntactic level of the Meaning Text model.

A LF is a relation between an argument and a value, where the value is a linguistic expression (the 'collocate' in the examples above) which expresses the meaning which corresponds to the LF. (Mel'cuk suggests for 60 LFs).

LF can be 'semantically' classified:

evaluative qualifiers: eg. Bon(cause)=worthy

distributional quant: eg. Mult(sheep)=flock, Sing(rice)=grain

involving preposition: eg. Loc(distance)=at

involving V operators: eg. Oper(attention)=pay

LFs are further refined by using subscripts and superscripts and extended forming compound LFs.

ET-10/75 investigates to what extend LFs (the 'semantic' relations between base/collocates) 'follow' certain syntactic rules. It seems that for each kind of 'syntactic' collocation there is a specific set of LF so that certain LFs apply on certain collocations. They study whether certain LF only apply to particular syntactic categories and only outputs particular categories (for instance Loc outputs prepositions).

Arguments and values of LF are uninflected lexemes. (they stand out the advantage of dealing with lemmatized texts when looking for collocations and dealing with 'lemmatized' LFs in order to 'simplify' the system).

Translation of Collocations When a collocation of one language corresponds to a collocation in the target language things are easy. Problems arise when there are mismatches:

- lexicalization: collocation = single unit

  ET-10/75 suggests for 'merged' LF: the application of a LF in one language corresponds to the application of its merged counterpart in the other (eg. Antibon(interpret)=misinterpret == Antibon(interpretieren)=falsch interpretieren).

- non-default translation: one = many as in the English/Spanish examples:

  bunch of grapes - racimo de uvas

  bunch of bananas - piña de plátanos

  bunch of key - manojo de llaves

  In these cases a 'default' correspondance is not possible and pair correspondances are explicitly specified.

- regular compound formation (eg, German resorts to compounds where English resorts to collocations). In ET-10/75 collocations and compounds are given the same semantic form. This allows approaching Collocation/Compounds mismatches are approached via 'merged' LF (eg, merged LF = non-merged LF.

- collocation = non-collocation. No LF analysis is possible.

- collocation = paraphrase mismatches are approached in terms of a 'Paraphrasing System'. The 'Paraphrasing System' which consists of: (i) a set of (60) lexical paraphrasing rules of the kind Causative = Decausative, contains = belongs ... so that when in one language there is no Causative constructions, decausative constructions is produced. (ii) syntactic paraphrasing rule: the ones which tale care of the 'changes' caused by lexical substitutions invoked by lexical PRs.

5.4 Proper Noun Recognition and Classification

Recognising and classifying proper nouns involves identifying which strings in a text name individuals and which classes these individuals fall into. Typical name classes include organisations, persons, locations, dates and monetary amounts. However, further classes can include book and movie titles, product names, restaurant and hotel names, ship names, etc. The task is made difficult by the unpredictable length of names (company names can be twelve or more words long), ambiguity between name classes (*Ford* can be a company, a person, or a location), embedding, where e.g. a location name occurs within an organisation name, variant forms, and unreliability of capitalisation as a cue, e.g. in headlines in English and everywhere in German.

Being able to recognise and classify proper names correctly clearly has relevance for a number of application areas:

- precision in IR systems should increase if multiword names are treated as unitary terms and if variant forms can be linked;

- IE systems rely heavily on PN recognition and classification as MUC-6 results have shown.

Most PN recognition systems use lexical resources such as gazetteers (lists of names) but these are necessarily incomplete, as new names are constantly coming into existence. Therefore further techniques must be used, including syntactic analysis and semantic classification based on verb and preposition complement roles.

Part IV

Guidelines for Lexical Semantic Standards

Chapter 6

Guidelines for Lexical Semantic Standards

The goal of this chapter is to establish guidelines for the deliberation of standards in the encoding of lexical semantic information. Our primary concern is to determine which aspects of lexical semantic information are needed most in Machine Translation and Information Systems, and are available for feasible development of large scale lexical resources. Readers who wish to consult the results now can proceed directly to §6.9 where a listing of the guidelines is given.

Word sense identity is perhaps the most pervasive notion underpinning the provision of lexical semantic standards. Word sense identification does in fact underly the classification of words into concepts and is elemental in capturing generalisations about regularities in word usage shifts — e.g. metonymies, metaphors and verbal diatheses. Unfortunately, a full standardisation of formal, operational and substantive criteria for word sense identification is not an objective which can be achieved within the scope of the present work. Such an enterprise would require the direct involvement of an international congregation of lexicographers representing the major language resource producers. Moreover, it is not clear that a common protocol for word sense identification can be reached without a specification of intended use and a shared top-level reference ontology.

In order to provide operative criteria and pave the way towards common protocols for word sense identification, our standardisation effort will focus on basic conceptual notions which are general enough to ensure compatibility with any specific word sense classification, and sufficiently grounded in language technology applications so as to ensure utility. Such basic conceptual notions have been established by assessing the relevance of lexical semantic information types used in the applications and enabling technologies reviewed in the previous chapters, as indicated in Table 6.1.

Basic conceptual notions have also been partitioned into three priority bands to provide an indication of relative weight, according to the following criteria (see Table 6.1):

- high priority: notions which are used in most of the applications and enabling technologies reviewed and are (likely to be) utilized in language technology products (soon);

- medium priority: notions which are used experimentally in some of the applications and enabling technologies reviewed, and

	BWT	HYP	ANT	MER	SeF	TNS	TME	ASP	DOM	COO	QUA
MT	*	*		*	*	*	*	*	*	*	*
IR	*	*		*	*						
IE	*	*		*	*	*		*			
TS	*	*							*		
LG	*	*	*	*	*	*	*	*	*	*	*
WC	*	*	*							*	
MR	*	*			*				*		
WD	*	*	*		*				*	*	
PN	*	*		*							
PA	*	*			*			*		*	
CR	*	*			*						*

High Priority Basic Notions
BWT = Base Word Types
HYP = Hyponymy/Synonymy
SeF = Semantic Frames

applications
MT = Machine Translation
IR = Information Retrieval
IE = Information Extraction
TS = Text Summarization

Medium Priority Basic Notions
MER = Meronymy
ANT = Antonymy
DOM = Subject Domains
COO = Co-occurrence relations
TNS = Tense
TME = Time (temporal relations)
ASP = Aspect
QUA = Quantification

enabling technologies
LG = Generation
WC = Word Clustering
MR = Multiword Recognition
WD = Word Sense Disambiguation
PN = Proper Nouns
PA = Parsing
CR = Co-reference

Table 6.1: Relevance of lexical semantic notions to applications and enabling technologies.

- low priority: notions which are currently not generally utilized in applications or enabling technologies, but are of potential value for future developments of language technologies.

The chapter is mostly concerned with high and medium priority lexical semantic notions.

6.1 Hyponymy, Synonymy and Base Types

Hyponymy and synonymy are fundamental semantic relations that make up the core of semantic lexicons. A standard lexical semantic definition of the notions is given in (§2.3), mainly based on the work of [Cru86]. We can paraphrase these definitions as follows:

identity (W1 = W2): If two words have the same meaning, they are synonyms, e.g. *fiddle* and *violin*.

inclusion (W1 is included into W2): If one word is more general than the other, we speak of hyponymy, where the more general term is the hyperonym or hypernym, e.g.

musical instrument and the more specific term is the hyponym, e.g. *piano*. More general implies that the meaning of the hyperonym is fully *included* by the meaning of the hyponym but the hyponym also implies more (a *piano* is more than a *musical instrument*).

disjunction (W1 and W2 have no element in common): Words that share the same hyperonym and have no overlap in meaning are called co-hyponyms, e.g. *piano* and *violin*.

Similarity or generality of meaning can then be further defined in different ways in the form of a more-specific subsumption relation:

extensional sets if two concepts refer to the same sets of entities they are synonymous, if one concept refers to a genuine subset of the other concept there is a hyponymy relation.

intensional definition if two concepts have the same set of features they are synonymous, if one concept is defined by the genuine feature subset of the other concept there is a hyponymy relation.

substitution definition if two concepts can be interchanged in a linguistic context without a change in meaning they are synonymous, if one concept can substitute the other but not the other way around they are hyponymous.

These definitions partly overlap in their operational result and their usefulness depends on the kind of usage in an NLP system. An Information System that keeps track of a database administration needs to know to what individuals words can refer and therefore must operationalize intensional definitions to extensional sets. It may very well be that different intensional definitions lead to the same extensional set: *my boss and my wife can be the same person.* An Information Retrieval System must be able to predict what words can be substituted for each other to capture variation in expression but does not need to capture the extensional effect.

An operational test for hyponymy which can be used in general is the following:

- X is a subtype of Y if the following expressions are correct:

  X is a kind of Y or X is a type of Y for nouns,

  X-ing is a way of Y-ing for verbs.

A pragmatic consequence of the hyponymy relation is that the more general word can be used to substitute or replace a more specific word without changing the truth of the sentence, but not the other way around:

- If X is a subtype of Y and the sentence "There was an X" is true then, as a logical consequence, also the sentence "There was an Y" is true.

- If X is a subtype of Y and the sentence "There was an Y" is true then it is not necessarily the case that also the sentence "There was an X" is true.

If two words are mutually substitutable, we speak of synonymy:

- If X and Y are synonyms then:

  if the sentence "There was an X" is true then also the sentence "There was an Y" is true, and,

- if the sentence "There was an Y" is true then also the sentence "There was an X" is true.

It is a common believe that there are very few absolute synonyms, if any, in a language, but words may be synonyms in given contexts. We then view the synonymy relation as a ternary relation: W1 and W2 are synonyms in the context C.

Another important characteristic of synonymy and hyponymy is the transitivity of the relation:

- if A is a synonym of B, and B is a synonym of C, then A is also a synonym of C.
- if A is a hyperonym of B, and B is a hyperonym of C, then A is also a hyperonym of C.

Transitivity of hyponymy relations makes it possible to organize the word meanings of a language as a hierarchical network or classification scheme, e.g.:

```
thing
.... animate object
.... .... human
.... .... plant
.... .... animal
....inanimate object
.... .... natural object
.... .... artifact
.... .... .... building
.... .... .... .... apartment building 1
.... .... .... .... bakery 2
.... .... .... .... bathhouse 2
.... .... .... .... boathouse 1
.... .... .... .... chancellery 1
.... .... .... .... city hall 1
.... .... .... .... courthouse 2
.... .... .... .... farm building 1
.... .... .... .... glasshouse 1
.... .... .... .... hotel 1
.... .... .... .... ruin 2
.... .... .... .... school 1
.... .... .... .... smoke house 1
.... .... .... .... temple 2
.... .... .... .... theater 1
.... .... .... .... etc...
```

Hierarchical networks are very powerful structures because classifications at the top can be inherited to large numbers of word meanings that are directly or indirectly related to these top levels. In the above example, we can for all hyponyms below *building* make the same semantic inferences:

as buildings they have roofs and walls.

as artifacts they are created by humans for some purpose.

as inanimate objects they have a fixed shape, do not live, act, or die, but can be destroyed.

as things have colour, smell, taste, weight, size.

By augmenting important hierarchy nodes with such basic semantic features, it is possible to create a rich semantic lexicon in a consistent and cost-effective way after inheriting these features through the hyponymy relations.

The transitivity of synonymy relations leads to another powerful organization into synsets [Mil90a], see section on Wordnets 3.6. A synset is a set of synonymous words, where synonymy is defined as: substitutability of words in *most* linguistic contexts. Synsets are said to represent a single concept, so that semantic descriptions as the above hyponymy relations and semantic features only have to be represented once for all the synset members, e.g.:

```
{coffee bar 1; cafe 1; coffee shop 1; coffeehouse 1}
        HAS-HYPERONYM
{restaurant 1, eating house 1}
```

In this example, the hyponymy relation only needs to be specified once for the synset as a whole and not 8 times for each synset member (the same holds for other conceptual information that is shared: the gloss, part-of-speech, domain information). Information sharing enforces consistency in lexical resources.

6.1.1 Lexical Resources

Hyponymy and synonymy are implemented in many resources. In Chapter 3, the following resources with hyponymy and synonymy information are listed:

GLDB a Swedish Machine Readable Dictionary (§3.5)

PrincetonWordNet and EuroWordNet databases with hyponymy and synonymy relations as the core information (§3.6)

Memodata lexical semantic database with French as the core language(§3.7)

EDR Japanese en English lexical semantic database (§3.8)

Cyc, MicroKosmos, Upper-Model and Sensus Ontologies with world-knowledge (§3.9)

Snomed and UMLS Medical ontologies (§3.10)

Corelex, Acquilex, ET-10/51 Experimental lexicons for NLP (§3.12)

Only WordNet and EuroWordNet also incorporate the notion of a synset, others provide cross-references to synonyms. However, both organizations are compatible and can be interchanged. Because the notion of a synset as the basic unit for semantic organization is rather novel, we will specify the information at the word sense level in these guidelines.

Whereas many of the above resources only contain the lexical semantic links, only the higher level ontologies and the experimental lexicons also use the hierarchies to inherit rich semantic specifications from the top-levels to the more specific concepts (possibly using multiple inheritance). The top nodes of these hierarchies play an important role. They determine the semantic classes of all words in the lexicon, which can guide further specification and differentiation of their meaning.

6.1.2 Usage

Synonymy and hyponymy relations are used in many different applications that deal with content. There are two powerful ways in which such a network can be used:

- to capture lexical variation by giving alternative or close words that can be used to express the same or similar meaning, and
- to make semantic inferences by inheriting semantic properties from hyperonyms.

The first usage is exploited in Information Retrieval applications (§4.2) to expand a user's query words to other words that may express the same content, or in Information Extraction (§4.3) to know what words may refer to the same content. A query with the keywords *bicycle* and *journey* can be expanded to the following synonyms and hyponyms:

```
journey 1 ->
  long haul 1; odyssey 1; trip 3; flight 2; domestic flight 1;
  international flight 1; nonstop 1; redeye 1; round trip 1; run 3, 4;
  trek 1; errand 1; service call 1; passage 2; lockage 1; drive 2;
  lift 2; joyride 1; expedition 1; scouting trip 1; campaign 1;
  exploration 1; digression 1; trek 2; circuit 2; walkabout 3; grand
  tour 1; pub crawl 1; pilgrim's journey 1; hadj 1; excursion 2; air
  2; voyage 1; crossing 2; space travel 1; way 2
bicycle 1 ->
  bike 1; velocipede 1; exercise bike 1; all-terrain bike 1;
  ordinary bike 2; push-bike 1; safety bicycle 1
```

Such an expanded query will improve the recall of documents considerably, when applied to a large set of documents. The same relations can be exploited to find alternative wordings for Machine Translation (§4.1)or Language Generation Tools (§4.5).

The second usage we see for verifying selection restrictions, anaphora resolution, term recognition, PP-attachment in parsing (Machine Translation, Information Extraction (§4.2), Language Generation, Tex Summarization (§4.4)). In all these cases it is necessary to make a semantic inference, e.g.:

- The doctor was young but she brilliantly transplanted the organ with her skilled hands.
- The musician was young but she brilliantly played the organ with the new keyboard.

The selectional restriction for the object of *to transplant* is *animate tissue*. This information can be derived from the top nodes in the hierarchy: *organ*(as a *musical instrument*) is a sub-hyponym of *inanimate object* and *organ*(as a *body part*) is sub-hyponym of *animate*. It is also possible to inherit this information by assigning features to the top nodes (as described above) and inherit these to all (sub-)hyponyms. In the latter case it is easier to control the features that can be used to make inferences or specify selectional restrictions. In the former case, the restrictions can be made much more specific: i.e. there can be as many restrictions as nodes in the hierarchy. The same example also shows that the reference of the pronoun *she* can be resolved by inferring that *doctor* and *musician* are *human*. Similar inferences from hyponymy classifications and meronymy relations (see below) are needed to extract ! that *her skilled hands* belong to the *doctor* and *the new keyboard* to the *organ*.

Finally, we see that hyponymy/synonymy are used for component tasks such as word-sense disambiguation, semantic tagging and parsing. Word-sense disambiguation (§5.3) and semantic tagging are important subtasks for Information Retrieval, Information Extraction or Machine Translation. Nevertheless, they can involve sophisticated processing. The previous example shows how selectional restrictions can resolve the ambiguity of *organ*. Parsing can also benefit from this information because the same selectional restrictions can exclude possible verb complements or PP-attachments. Other techniques use the hyponymy structure to calculate the conceptual distance of words for word-sense-disambiguation or word clustering.

6.1.3 Guidelines

In many resources the notions *same meaning* or *similar meaning* are not further defined and just expressed in different relations. As a result we see that many thesauri cluster words that are simply related. For example, we find the following classification in the standardized medical thesaurus MESH:

```
Transportation
... Aviation
... ... Aircraft
... ... ... Air Ambulances
... ... Space Flight
... ... ... Extravehicular Activity
... ... ... Spacecraft
... Motor Vehicles
... Railroads
... Ships
```

The terms *Space Flight* and *Extravehicular Activity* do not represent subclasses of *vehicles* but are types of *activities* related to certain *vehicles*. This structure can only be used to globally extract words that are *related*, but it cannot be used to make inferences such as: all the things that can be used to transport people, goods. Another unclarity of the above structure is that the related categories are not differentiated as word meanings. If one of the target meanings is polysemous (e.g. *organ*) it is not clear which if the different meanings is related here.

A good specification of a lexical semantic resource thus requires that the links and the targets are explicitly encoded. The next example shows how this is done in EuroWordNet:

```
0 WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
```

```
  2 LITERAL "Hammond organ"
    3 SENSE 1
    3 DEFINITION "an electronic simulation of a pipe organ"
  2 LITERAL "electric organ"
    3 SENSE 1
  2 LITERAL "electronic organ"
    3 SENSE 1
  2 LITERAL "organ"
    3 SENSE 1
1 INTERNAL_LINKS
  2 RELATION "HAS_HYPERONYM"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "n"
      4 LITERAL "musical instrument"
        5 SENSE 1
```

Here we see that synonymy is stored by listing explicit references to word senses as VARIANTS and that other semantic relations are made explicit by specifying the relation type HAS_HYPERONYM and the relation target sense.

Another issue in networks of synonymy and hyponymy relations is the set of concepts between which relations are expressed. In some approaches, the resources are limited to lexicalized words and phrases that are actually used in languages (traditional dictionaries, EuroWordNet), in other resources (the Princeton WordNet, EDR, Cyc) we see that artificial classes are introduced to structure the ontology:

- hyponymic relations between concepts and the features that belong to the concept, e.g. *instrument - instrumentality.*
- hyponymic relations to non-lexicalized word classes, e.g. *leg - external body part.*

The difference is clearly illustrated by the next example, which contain the most frequent classes from the public part of the Cyc ontology (3000 concepts):

```
10 #$Action)
10 #$BiologicalEvent)
10 #$BiologicalTaxon)
10 #$BodyMovementEvent)
10 #$LandTransportationDevice)
10 #$PropositionalInformationThing)
10 #$QualitativeTimeOfDay)
11 #$CompositePhysicalAndMentalEvent)
11 #$HandlingAnObject)
11 #$InanimateThing)
11 #$Movement-TranslationEvent)
11 #$Person)
12 #$AbstractInformation)
12 #$AnimalActivity)
12 #$Business)
```

```
12 #$CalendarMonth)
12 #$Collection)
12 #$IntangibleIndividual)
12 #$SheetOfSomeStuff)
13 #$SingleDoerAction)
13 #$TimeInterval)
13 #$TransportationDevice-Vehicle)
14 #$Animal)
15 #$Artifact)
15 #$ContainerProduct)
15 #$DestructionEvent)
15 #$ExistingObjectType)
15 #$Individual)
15 #$SocialOccurrence)
17 #$AnimalBodyPart)
18 #$Ablation)
18 #$Predicate)
19 #$BinaryPredicate)
20 #$IntrinsicStateChangeEvent)
20 #$PhysicalEvent)
21 #$BiologicalLivingObject)
21 #$PhysicalDevice)
22 #$PartiallyTangible)
23 #$Organization)
24 #$Professional)
24 #$TangibleThing)
25 #$SolidTangibleThing)
29 #$FeelingAttribute)
31 #$UnitOfMeasure)
32 #$AttributeValue)
32 #$HumanActivity)
33 #$PurposefulAction)
42 #$PhysicalAttribute)
56 #$ScalarInterval)
```

Artificial levels may be helpful to capture important inferences or to provide a more clear structure of the hierarchy, but they also blur the distinction between legitimate and unlikely expressions in a language. A Language Generation tool should not generate expressions with artificial terms, and word clustering based on semantic distance or lexical density in the hierarchy will be affected by artificial levels. A good practice, as is followed in EDR, is therefore to mark unnatural classes and expressions from the actual lexicalizations in a language.

At a more general level, we can state that the confusion is the result of the fact that the purpose of the information is not clear. Is the network created for capturing world-knowledge inferences, for capturing lexicalization patterns in languages, for a thematic organization of information, for building a translation resource? For standardizing resources, merging existing resources or evaluating resources it is extremely important that one is aware of these differences and that they are clarified where possible.

Synonymy

Many scholars believe that real synonyms do not exist in languages. Most available resources do not provide a clear specification either what differences are in- or excluded from their definition of synonymy. In the case of Wordnet1.5, we see for example that a very weak definition of synonymy is given: two words are synonymous if they can be interchanged in most contexts. According to this definition it is sufficient to find some contexts, while other contexts (which may reveal the above differences) block substitution. This definition for example also includes pragmatic specialization, where a more general word (e.g. *machine*) is used to refer to a more specific concept (e.g. *car*). Consequently, the synsets in Wordnet1.5 contain word meanings that are synonyms in a very broad sense.

In practice, it is a good thing to exclude differences in style, register and dialect from the definition of synonymy: words are synonyms if they have the same meaning regardless of differences in style, register, or dialect. In the case of morpho-syntactic properties things are more complicated. Often there is a semantic difference that corresponds with a morpho-syntactic difference, but this is not always the case:

- water - waters
- museums - musea
- person(s) - people
- eat (transitive) - eat (intransitive)
- move (transitive) - move (intransitive)

We see here that differences in plural form can be meaningful (water - waters), but sometimes there is clearly no difference (museums - musea), and in yet other cases the difference in meaning is unclear (person(s) - people). The same holds for valency differences, which either just reflect vagueness (*eat*) or a difference in causation (something *moves* without explicit cause, or somebody *moves* something). As a rule of thumb, it is a good practice to exclude morpho-syntactic differences that are not functional: i.e. do not correlate with a semantic difference that can be expressed in the lexical semantic specification. If a causation relation can be expressed in a database, it makes sense to keep the two syntactic patterns of *move* separate.

We also see the reverse effect that conceptual synonymy is broadened over syntactically distinct classes. In EuroWordNet, pseudo-synonymy relations are expressed between words that belong to different parts-of-speech, e.g. *to adorn - adornment*, *death - dead*. These synonymy relations are differentiated as a specific subtype of synonymy. This is very useful to map for example nominalizations to sentences in information retrieval tasks, e.g.:

- The departure of the train is at 5 pm.
- The train leaves at 5 pm.

The verb *leave* is in the same synset as the verb *depart*, which has a cross-part-of-speech synonymy relation with the noun *departure*.

Finally, the notion of synonymy directly relates to polysemy. If different meaning aspects are spread over different meanings of a word, we get a different clustering of synonyms than we would get for resources where meaning aspects are derived from condensed core meanings (Corelex, Generative Lexicon, Acquilex). For example, the compound *biological mother* can be synonymous with *mother*, if a specific sense for *mother* is restricted to biological motherhood, otherwise it will only be a hyponym.

More general, we can say that all the semantic information expressed for word meanings, such as the glosses or definitions, the hyperonyms and hyponyms, semantic features, should hold for all words that are considered to be synonyms. This makes the synset a powerful organizational principle that enforces consistency of encoding. It also suggests that encoding other semantic information first may make it easier to decide on what word meanings are synonyms.

In the case of doubt, it is still possible to distinguish a separate near-synonym relation, when all information is shared but intuitions are still not clear. Close meanings can then at least be kept separate from other co-hyponyms that are really different. For example, the words *apparatus*, *device*, *machine*, *tool*, *instrument* may be very close in meaning but are not real synonyms because they group different clusters of hyponyms. If one decides to store them as separate concepts, they will share the hyperonym *artifact* among them but also with other co-hyponyms such as *construction*. By encoding an explicit near-synonym relation it is possible to create a subgroup of co-hyponyms below *artifact* that are very close and can be differentiated from other more distant co-hyponyms:

```
0 WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "apparatus"
      3 SENSE 1
  1 INTERNAL_LINKS
    2 RELATION "HAS_HYPERONYM"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "artifact"
          5 SENSE 1
    2 RELATION "HAS_NEAR_SYNONYM"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "tool"
          5 SENSE 1
    2 RELATION "HAS_NEAR_SYNONYM"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "machine"
          5 SENSE 1
    2 RELATION "HAS_NEAR_SYNONYM"
      3 TARGET_CONCEPT
```

```
      4 PART_OF_SPEECH "n"
      4 LITERAL "device"
        5 SENSE 1
  2 RELATION "HAS_NEAR_SYNONYM"
    3 TARGET_CONCEPT
      4 PART_OF_SPEECH "n"
      4 LITERAL "instrument"
        5 SENSE 1
```

Near synonyms can partly overlap and substitute each other, other co-hyponyms cannot. This is important information for tasks such as Information Retrieval, co-reference resolution.

Hyponymy

Hyponymy is traditionally used for classifying nouns and things in the world, creating ontologies, thesauri or taxonomies. Recent attempts show that the relation is less useful for verbs and adjectives/adverbs. Whereas many hierarchical levels can be created for nouns, hierarchies for the other parts of speech are more shallow and therefore less informative. A componentional specification, combined with a specification of the argument structure appears more fruitful for verbs and adjectives/adverbs.

Currently, most systems allow for multiple hyperonyms, which results in complex tangled hierarchies or directed graphs rather than tree-like taxonomies. The use of multiple hyperonyms makes it possible to build more consistent and comprehensive hierarchies. However, there are several issues related to these more complex-graph structures which are often not clear.

- multiple hyperonyms can be exclusive, conjunctive or non-exclusive
- co-hyponyms can be disjunct or non-exclusive
- hyperonyms may be domain or context dependent

Three types of hyperonym combinations occur:

exclusive hyperonyms *albino* is either an *animal* or a *human*

conjunctive hyperonyms *spoon* is both *cutlery* and *container*

non-exclusive hyperonyms *knife* can be *cutlery*, a *weapon* or both

Exclusive hyperonyms often occur in traditional definitions as the coordinated genus words of a single sense (*a human or animal with ...*) but can also be represented as different meanings of the same word. In the former case, they lead to multiple hyperonyms that should be distinguished from cases such as *spoon* and *knife*, where both hyperonyms can apply simultaneously. In the latter cases there may be inheritance from multiple paths, in the case of *albino*, either of the two paths has to be chosen to derive more properties. Finally, the difference between *spoon* and *knife* is the optionality and relevance of their hyperonyms. It is save to say that non-exclusive combinations of classifications are the default (most word meanings can be considered from multiple perspectives) and conjunction and disjunction needs to be expressed explicitly.

Disjunction also plays a crucial role to differentiate between different types of co-hyponyms. In the case of hyponyms, we see a difference between hyponymic classes which are disjunct and therefore complementary, e.g. *horse*, *camel*, *cat*, *dog*, etc., and classes that may overlap and intersect, e.g. *draught animal*, *pet*, *breeder*, *racing animals*, etc. Overlapping hyponyms may not only overlap with each other but also with disjunct classes. This information is crucial for information retrieval and for making inferences. Exclusive classes cannot be interchanged or substituted in text (you cannor refer to a *dog* by *cat*), but non-exclusive classes can (you can refer to a *dog* with *pet*). On the other hand, exclusive classes represent rich classification schemes for inheriting properties, whereas the overlapping classes do not predict more than just the role or concept they express. This is inherent to their nature of ! appl ying to diverse types of concepts.

Finally, not all hyponymic relations are equally relevant. In many cases there are different classification schemes possible for a concept but these relate to very different domains and contexts. We see for example in Wordnet1.5 very elaborate biological classifications for species that are not relevant outside a very specific and limited domain:

```
horse 1
  -> equid 1 -> odd-toed ungulate 1 -> hoofed mammal 1
  -> eutherian 1 -> mammal 1 -> chordate 1 -> animal 1
```

To generate alternative expressions for *horse* or to measure the conceptual distance to other words such *cat* and *dog*, the long chain of hyperonyms may be totally irrelevant outside the domain of biology. In that case, it should be possible to skip these levels and directly go to *animal*. Yet other classifications below animal, such as *pet*, *draught animal*, seem to be too circumstantial to serve as a general classification for *horse*.

Although the above differentiation of the status of the hyponymy relations is hardly ever explicitly coded, it is extremely important for making inferences and tailoring hierarchies for information retrieval. Furthermore, it is necessary to consider the status of the relations to evaluate differences in hyponymy relations across resources. In current resources, one often sees that the specification of hyperonyms is very inconsistent. Sometimes, expert classifications are used, sometimes they are skipped for a related concept, sometimes in one and the same resource. A more consistent and comprehensive encoding of hyponymy relations is to prefer but in that case it is advisable to explicitly specify the different status of these relations as explained.

An explicit encoding of disjunctive classes can be done by adding features to the relations, as is done in EuroWordNet ([Vos98]):

```
being
 HYPONYM (disjunctive) plant
 HYPONYM (disjunctive) human
 HYPONYM (disjunctive) animal
 HYPONYM (non-exclusive) parent
 HYPONYM (non-exclusive) child
animal
 HYPERONYM being
 HYPONYM (disjunctive) horse
 HYPONYM (disjunctive) cat
```

```
 HYPONYM (disjunctive) dog
 HYPONYM (disjunctive)(biology) chordate
 HYPONYM (non-exclusive) pet
 HYPONYM (non-exclusive) draught animal
horse
 HYPERONYM (conjunctive)animal
 HYPERONYM (conjunctive)(biology) equid
 HYPERONYM (non-exclusive)pet
 HYPERONYM (non-exclusive)draught animal
 HYPONYM (disjunctive) mare
 HYPONYM (disjunctive) stallion
dog
 HYPERONYM (conjunctive) animal
 HYPERONYM (conjunctive) (biology) canine
 HYPERONYM (non-exclusive)pet
```

From this specification we can infer that a *horses* can be classified as a *pets* or *draught animals* if the circumstances allow it, but are always considered as *animals* and in the domain of biology as *equids*. However, the class *horse* itself is a disjunct type which can never overlap with other disjunctive co-hyponyms such as *cat* or *dog*. At the next level, we can also infer disjunctivenes with *plant* and *human* but overlap with classes such as *parent* and *child*.

A further test for building a more consistent hierarchy is the Principle of Economy [Dik78], which states that a word should not be defined in terms of more general words if there are more specific words that can classify it:

- If a word W1 (animal) is the hyperonym of W2 (mammal) and W2 is the hyperonym of W3 (dog) then W3 should not be linked to W1 but to W2.

This principle should prevent that senses are linked too high up in the hierarchy and that intermediate levels are skipped.

Finally, to achieve consistency in encoding hyponymy relations, the best approach is to build the hierarchy top down starting from a limited set of tops or unique beginners or, at least to have such a structured set available. One of the major shortcomings of most hierarchies is not that they wrongly classify words but that the classifications are incomplete. Having an overview of the classes, even at a very high level, makes it possible to more systematically check the possible classes. Furthermore, a systematized top level makes it easier to compare and merge different ontologies.

Base Types

In EuroWordNet, a standardized list of such important top concepts has been established for a series of languages. These concepts play an important role in wordnets for English (represented by WordNet1.5), French, German, Spanish, Italian, Dutch, Czech and Estonian and a lexicon for Swedish. If a word meaning has many relations with other meanings (mostly many hyponyms) and/or high positions in the hierarchy it has been selected. These words represent the most basic classes and semantic components in each language and are carried over to many other meanings via the hyponymy relations. They are therefore called the EuroWordNet Base Concepts. The Base Concepts in EuroWordNet can be compared with the unique beginners or taxonomy tops in WordNet1.5, but also include concepts at deeper levels that play a crucial role. The selections in the individual languages have been translated to the closest WordNet1.5 equivalents and these translations have been compared. This has resulted in a set of 1310 Common Base Concepts that occurred in at least 2 selections. Each group then tried to represent them in their language-specific wordnet. The construction of each wordnet started with the representatives of the set of Common Base Concepts, possibly extended with Base Concepts in their local languages which have not been part of the common set. For all the Base Concepts, the hyperonyms have been specified which results in a core wordnet for the most basic classes. Next the wordnets have been extended top-down. The purpose of the Common Base Concepts in EuroWordNet is to have a common starting point, and, at the same time, to make it possi ble to develop the wordnets relatively independently and allow for differences in the lexicalization of classes.

The first point for standardizing hyponymy classifications, which is relevant here, is also to agree on a set of top-level distinctions. The set of Common Base Concepts is however a rough empirical set of concepts derived from 8 languages and equally many different resources. Furthermore, the goal of the EuroWordNet selection has not been to develop a minimal unified ontology but to have the most representative set.[1] It was therefore more important to make sure that no important concepts have been missed rather than reducing it to the smallest set of primitives. It is thus possible to further minimalize the set of Common Base Concepts to smaller set by applying the following procedure:

- further limit the set to concepts selected by 6 out of the 9 languages (this resulted in a set of 164 WordNet1.5 synsets).
- remove specific concepts if a clear hyperonym is also selected and other co-hyponyms are not. For example, both a hyperonym *human* and a hyponym *worker* are present but not other co-hyponyms such as *tourist*, *ruler*. We have therefore removed *worker* from the list.
- remove close meanings that cannot be discriminated and represent them by a single concept, e.g. the synsets (*material 5*; *stuff 7*) and (*matter 1*; *substance 1*) can be represented by a single concept SUBSTANCE.

In some cases, we also added new concepts which are not represented by any of the 164 synsets. This was necessary to get a more clear structure. For example, there was no synset for *Change of Possession* but only synsets for *give*, *get* and *remove*. By inserting the concept *Change of Possession* these 3 perspectives can be grouped.

The output of these measures is a set of 74 concepts, listed below. There are several reasons why we can no longer refer to this list as a list of word meanings. First of all we have inserted new concepts to represents groups of related word meanings, and, secondly, we have inserted artificial (and therefore new) concepts to make the list more balanced. More important is however the role that these concepts play in the lexical semantic specification. As a standardized list of top-level distinctions it should be possible to incorporate them in the lexical resource of any language, regardless of the set of lexicalized items in that language. It may be that some concepts cannot be represented by a word meaning in a language or that

[1]It should be noted that EuroWordNet has only dealt with nouns and verbs. The selections are not representative for adjectives and adverbs.

there are word meanings in a languages that can only be represented by a combination of concepts. For example, in Dutch there is no word meaning for the concept *Artifact Object* but there is a word meaning *kunststof* for *Artifact Substa! nce* . In English we see the opposite lexicalization, where *artifact* refers to *objects* only, and there is no word for *artificial substances*. It therefore makes more sense to represent the standardized concepts as a list of language-independent classes rather than as word meanings. Since we can no longer speak of word meanings or synsets, the concepts are called Base Types and represented in capital letters to stress their artificial nature.

The first level of the 74 Base Types is divided into:

ENTITY things, mostly concrete

CONCEPT concepts, ideas in mind

EVENT happenings involving change

STATE static situations

Each of these is further subdivided, in some cases 5 levels deep.

```
1. ENTITY
1.1. ORIGIN
1.1.1. NATURAL
1.1.1.1. ANIMATE
1.1.1.1.1. HUMAN
1.1.1.1.2. ANIMAL
1.1.1.1.3. PLANT
1.1.1.2. INANIMATE
1.1.2 ARTIFACT

1.2. FORM
1.2.1. OBJECT
1.2.2. SUBSTANCE

1.3. COMPOSITION
1.3.1. WHOLE
1.3.2. GROUP
1.3.3. PART

1.4. ROLES
1.4.1. AGENT
1.4.4. COVER
1.4.5. DRINK
1.4.6. FOOD
1.4.7. FURNITURE
1.4.8. GOODS
1.4.9. INSTRUMENT
1.4.10. REPRESENTATION
1.4.11. LANGUAGE
1.4.12. MONEY
1.4.13. IMAGE
1.4.15. ORNAMENT
1.4.16. PLACE
1.4.16.1. AREA
1.4.16.2. POINT
1.4.16.3. WAY
```

```
1.4.17. POSSESSION

2. CONCEPT
2.1. CONTENT
2.2. KNOWLEDGE
2.3. BELIEF

3. EVENT
3.1. CAUSE
3.2. PHENOMENON

3.3. CHANGE
3.3.1. CHANGE OF QUANTITY
3.3.1.1. DECREASE
3.3.1.2. INCREASE
3.3.2. CHANGE OF QUALITY
3.3.2.1. IMPROVE
3.3.2.2. WORSEN
3.3.3. CHANGE OF POSSESSION
3.3.3.1. GET
3.3.3.2. REMOVE
3.3.3.3. GIVE
3.3.4. CHANGE OF EXISTENCE
3.3.4.1. MAKE
3.3.4.2. DESTROY
3.3.5. CHANGE OF LOCATION

3.4. DO
3.4.1. COMMUNICATE
3.4.2. THINK

3.5. EXPERIENCE

3.7. TIME

4. STATE
4.1. PROPERTY
4.1.1. PHYSICAL PROPERTY
4.1.2. MODAL PROPERTY
4.1.3.  QUALITY
4.1.4.  MENTAL PROPERTY

4.2. RELATION
4.2.1. SOCIAL RELATION
4.2.2. POSSESSION
4.2.3. SPATIAL RELATION
4.2.3.1 AMOUNT
4.2.3.2. DIRECTION
4.2.3.3. PHYSICAL HAVE
```

There may or may not be a word meaning in a language that corresponds with a Base Type and there may be several Base Concepts that map to a single Base Type. To see how Base Types map to synsets we have provided another list where we added to each Base Type the set of Base Concepts (of the set of 164) that it represents. The mapping from Base Types to the 164 WordNet1.5 synsets is represented as a lexicalization relation. There can be different lexicalization relations from a Base Type to synsets:

LX-SYNONYM Base Type is directly lexicalized by a synset

LX-NEAR-SYNONYM Base Type is lexicalized by a number of very close synsets

LX-HYPONYM Base Type is lexicalized by a more specific synset

LX-HYPERONYM Base Type is lexicalized by a more general synset

LX-HOLONYM Base Type is lexicalized by a synset that contains it

LX-MERONYM Base Type is lexicalized by a synset that is contained by it

LX-SUBEVENT Base Type is lexicalized by a synset that is contained by it

The lexicalizations given here are all taken from WordNet1.5. They are identified by the synset, the part-of-speech and the file off-set position in WordNet1.5: e.g. `LX-HYPONYM {natural object 1} n 9919`. Other wordnets will have other lexicalizations. If we take the example of *artifact*, discussed above, we see that it is more specific in its meaning than the Base Type Artifact, because it applies only to objects, whereas the Base Type Artifact can also apply to substances. The word meaning *artifact* is thus represented as a lexicalized hyponym (LX-HYPONYM) of the Base Type Artifact. Consequently, we can list the same word meaning *artifact* below the Base Type Object as well. Such multiple classifications have been included occasionally to illustrate the possibility, e.g. natural object 1 is listed below NATURAL and OBJECT. However, in most cases we have listed the Base Concept below its most typical Base Type, to keep the list short and readable. It will be c! lear that multiple assignments o f Base Types to word meanings makes it possible to classify these meanings in a flexible way, and to still unify different lexicalizations of classifications.

```
List of Base Types with WordNet1.5 lexicalizations
1. ENTITY
   LX-SYNONYM {entity 1}n 2403
1.1. ORIGIN
1.1.1. NATURAL
  LX-HYPONYM {natural object 1} n 9919
1.1.1.1. ANIMATE
  LX-SYNONYM {being 1; life form 1; living thing 1; organism 1}n 2728
  LX-WHOLE {body 3; organic structure 1; physical structure 1}n 3607347
1.1.1.1.1. HUMAN
  LX-SYNONYM {human 1; individual 1; mortal 1; person 1; someone 1; soul 1} n 4865
  LX-HYPONYM {worker 2} n 5856677
1.1.1.1.2. ANIMAL
  LX-SYNONYM {animal 1; animate being 1; beast 1; brute 1; creature 1; fauna 1} n 8030
1.1.1.1.3. PLANT
  LX-SYNONYM {flora 1; plant 1; plant life 1} n 8894
1.1.1.2. INANIMATE
1.1.2. ARTIFACT
  LX-HYPONYM {artefact 1; artifact 1} n 11607
  LX-HYPONYM {piece of work 1; work 4} n 2932267
  LX-NEAR-SYNONYM {product 2; production 2} n 2929839
  LX-HYPONYM {creation 3} n 1992919
  LX-HYPONYM {construction 4; structure 1} n 2034531
  LX-HYPONYM {building 3; edifice 1} n 2207842
1.2. FORM
1.2.1. OBJECT
```

```
  LX-NEAR-SYNONYM {inanimate object 1; object 1; physical object 1} n 9469
  LX-HYPONYM {natural object 1} n 9919
1.2.2. SUBSTANCE
  LX-SYNONYM {material 5; stuff 7} n 8781633
  LX-SYNONYM {matter 1; substance 1} n 10368
  LX-HYPONYM {fluid 2} n 8976164
  LX-HYPONYM {liquid 4} n 8976498
  LX-HYPONYM {chemical compound 1; compound 4} n 8907331
  LX-HYPONYM {chemical element 1; element 6} n 8805286
  LX-HYPONYM {mixture 5} n 8783090
1.3. COMPOSITION
1.3.1. WHOLE
  LX-HYPONYM {body 3; organic structure 1; physical structure 1}n 3607347
1.3.2. GROUP
  LX-SYNONYM {group 1; grouping 1} n 17008
1.3.2. PART
  LX-SYNONYM {part 3; portion 2} n 2855539
  LX-HYPONYM {bound 2; boundary 2; bounds 2} n 5383364
  LX-HYPONYM {part 12; portion 5} n 8450839
  LX-HYPONYM {extremity 3} n 5413816
  LX-HYPONYM {amount 1; measure 1; quantity 1; quantum 1} n 18966
1.4. ROLES
1.4.1. AGENT
  LX-SYNONYM {causal agency 1; causal agent 1; cause 1} n 4473
1.4.4. COVER
  LX-SYNONYM {covering 4} n 1991765
  LX-HYPONYM {cloth 1; fabric 1; material 1; textile 1} n 1965302
  LX-HYPONYM {apparel 1; clothes 1; clothing 1; vesture 1; wear 2; wearing apparel 1} n 2307680
  LX-HYPONYM {garment 1} n 2309624
1.4.5. DRINK
  LX-SYNONYM {beverage 1; drink 2; potable 1} n 5074818
1.4.6. FOOD
  LX-SYNONYM {food 1; nutrient 1} n 11263
1.4.7. FURNITURE
  LX-SYNONYM {article of furniture 1; furniture 1; piece of furniture 1} n 2008299
  LX-GROUP {furnishings 2} n 2043015
1.4.8. GOODS
  LX-SYNONYM {commodity 1; goods 1} n 2329807
  LX-HYPONYM {consumer goods 1} n 2344541
1.4.9. INSTRUMENT
  LX-NEAR-SYNONYM {device 2} n 2001731
  LX-NEAR-SYNONYM {instrument 2} n 2657448
  LX-NEAR-SYNONYM {instrumentality 1; instrumentation 2} n 2009476
1.4.10. REPRESENTATION
  LX-NEAR-SYNONYM {representation 3} n 2354709
  LX-NEAR-SYNONYM {symbol 2} n 4434881
1.4.11. LANGUAGE
  LX-SYNONYM {language unit 1; linguistic unit 1} n 4156286
  LX-HYPONYM {document 2; papers 1; written document 1} n 4242515
  LX-HYPONYM {writing 4; written material 1} n 4195435
  LX-HYPONYM {written communication 1; written language 1} n 4187642
  LX-HYPONYM {word 1} n 4157535
1.4.12. MONEY
  LX-NEAR-SYNONYM {medium of exchange 1; monetary system 1} n 8207032
  LX-NEAR-SYNONYM {money 2} n 8214427
1.4.13. IMAGE
  LX-HYPONYM {line 26} n 8484352
```

```
1.4.15. ORNAMENT
  LX-SYNONYM {decoration 2; ornament 1} n 2029323
1.4.16. PLACE
  LX-SYNONYM {location 1} n 14314
1.4.16.1. AREA
  LX-SYNONYM {part 9; region 2} n 5449837
  LX-HYPONYM {dry land 1; earth 3; ground 7; land 6; solid ground 1; terra firma 1} n 5720524
  LX-HYPONYM {opening 4} n 2028879
  LX-HYPONYM {surface 1} n 2486678
  LX-HYPONYM {surface 4} n 5467731
  LX-HYPONYM {line 21} n 5432072
1.4.16.2. POINT
  LX-SYNONYM {point 12} n 5443777
  LX-HYPONYM {place 13; spot 10; topographic point 1} n 5469653
1.4.16.3. WAY
  LX-SYNONYM {way 4} n 2031514
  LX-HYPONYM {passage 6} n 2857000
1.4.17. POSSESSION
  LX-SYNONYM {possession 1} n 17394
  LX-HYPONYM {asset 2} n 8179398

2. CONCEPT
  LX-NEAR-SYNONYM {concept 1; conception 3} n 3954891
  LX-NEAR-SYNONYM {abstraction 1} n 12670
  LX-NEAR-SYNONYM {attribute 2; dimension 3; property 3} n 3963400
  LX-NEAR-SYNONYM {cognitive content 1; content 2; mental object 1} n 3940357
  LX-NEAR-SYNONYM {idea 2; thought 2} n 3953834
2.1. CONTENT
  LX-NEAR-SYNONYM {communication 1} n 18599
  LX-NEAR-SYNONYM {message 2; content 3; subject matter 1; substance 4} n 4313427
2.2. KNOWLEDGE
  LX-SYNONYM {cognition 1; knowledge 1} n 12878
  LX-HYPONYM {information 1} n 3944302
  LX-HYPONYM {know-how 1; knowhow 1} n 3841532
  LX-HYPONYM {method 2} n 3863261
  LX-HYPONYM {structure 4} n 3898550
2.3. BELIEF
  LX-SYNONYM {attitude 3; mental attitude 1} n 4111788

3. EVENT
  LX-SYNOMYM {event 1} n 16459
  LX-HYPONYM {happening 1; natural event 1; occurrence 1} n 4690182
3.1. CAUSE
  LX-SYNONYM {cause 7; do 5; give rise to 1; make 17} v 941367
  LX-HYPONYM {cause 6; get 9; have 7; induce 2; make 12; stimulate 3} v 432532
  LX-HYPONYM {cease 3; discontinue 2; give up 12; lay off 2; quit 5; stop 20} v 1515268
3.2. PHENOMENON
  LX-SYNONYM {phenomenon 1} n 19295
  LX-HYPONYM {consequence 3; effect 4; outcome 2; result 3; upshot 1} n 6465491
3.3. CHANGE
  LX-NEAR-SYNONYM {change 1} n 108829
  LX-NEAR-SYNONYM {alter 2; change 12; vary 1} v 71241
  LX-NEAR-SYNONYM {change 11} v 64108
  LX-HYPONYM {change of state 1} n 113334
  LX-HYPONYM {change magnitude 1; change size 1} v 101800
  LX-HYPONYM {development 1} n 139142
3.3.1. CHANGE OF QUANTITY
```

```
3.3.1.1. DECREASE
  LX-SYNONYM {decrease 5; diminish 1; fall 11; lessen 1} v 90574
3.3.1.2. INCREASE
  LX-SYNONYM {increase 7} v 93597
3.3.2. CHANGE OF QUALITY
3.3.2.1. IMPROVE
  LX-SYNONYM {improvement 1} n 138272
3.3.2.2. WORSEN
  LX-SYNONYM {worsening 1} n
3.3.3. CHANGE OF POSSESSION
3.3.3.1. GET
  LX-SYNONYM {get hold of 2; take 17} v 691086
  LX-HYPONYM {consume 2; have 8; ingest 2; take 16} v 656714
3.3.3.2. REMOVE
  LX-SYNONYM {remove 2; take 4; take away 1} v 104355
3.3.3.3. GIVE
  LX-SYNONYM {give 16} v 1254390
  LX-HYPONYM {furnish 1; provide 3; render 12; supply 6} v 1323715
  LX-HYPONYM {cover 16} v 763269
3.3.4. CHANGE OF EXISTENCE
3.3.4.1. MAKE
  LX-SYNONYM {create 2; make 13} v 926361
  LX-HYPONYM {production 1} n 507790
  LX-HYPONYM {emit 2; express audibly 1; let loose 1; utter 3} v 554586
  LX-HYPONYM {represent 3} v 556972
3.3.4.2. DESTROY
  LX-SYNONYM {kill 5} v 758542
3.3.5. CHANGE OF LOCATION
  LX-NEAR-SYNONYM {change of location 1; motion 1; move 4; movement 1} n
157028
  LX-NEAR-SYNONYM {change of position 1; motion 2; move 5; movement 2} n
186555
  LX-HYPONYM {locomotion 1; travel 1} n 159178
  LX-HYPONYM {motion 5; movement 6} n 4704743
  LX-HYPONYM {go 14; locomote 1; move 15; travel 4} v 1046072
3.4. DO
  LX-NEAR-SYNONYM {act 12; do something1; move 19; perform an action 1; take
                  a step 2; take action 1; take measures 1;
                  take steps 1} v 1341700
  LX-NEAR-SYNONYM {act 1; human action 1; human activity 1} n 16649
  LX-NEAR-SYNONYM {action 1} n 21098
  LX-HYPONYM {activity 1} n 228990
  LX-HYPONYM {act together 2; act towards others 1; interact 1} v 1346535
  LX-HYPONYM {allow 6; let 7; permit 5} v 1371393
3.4.1. COMMUNICATE
  LX-SYNONYM {communicate 1; intercommunicate 1; transmit feelings 1;
             transmit thoughts 1} v 416793
  LX-SUBEVENT {sign 3;signal 1; signaling 1} n 4425761
  LX-SUBEVENT {convey 1; impart 1} v 522332
  LX-SUBEVENT {evince 1; express 6; show 10} v 531321
  LX-SUBEVENT {express 5; give tongue to 1; utter 1} v 529407
  LX-SUBEVENT {say 8; state 7; tell 7} v 569629
3.4.2. THINK
  LX-SYNONYM {cerebrate 1; cogitate 1; think 4} v 354465
  LX-HYPONYM {remember 2; think of 1} v 342479
3.5. EXPERIENCE
  LX-SYNONYM {experience 7; get 18; have 11; receive 8; undergo 2} v 1203891
```

```
  LX-HYPONYM {feeling 1} n 13522
3.7. TIME
  LX-SYNONYM {time 1} n 14882
  LX-HYPONYM {amount of time 1; period 3; period of time 1; time period 1} n 9065837

4. STATE
  LX-SYNONYM {situation 4; state of affairs 1} n 8522741
  LX-SYNONYM {be 4; have the quality of being 1} v 1472320
  LX-SYNONYM {state 1} n 15437
4.1. PROPERTY
  LX-SYNONYM {attribute 1} n 17586
  LX-HYPONYM {character 2; lineament 2; quality 4} n 3963513
  LX-SYNONYM {property 2} n 3444246
4.1.1. PHYSICAL PROPERTY
  LX-HYPONYM {color 2; coloring 2; colour 2; colouring 2} n 3463765
  LX-HYPONYM {form 6; pattern 5; shape 5} n 4003083
  LX-HYPONYM {visual property 1} n 3460270
  LX-HYPONYM {form 1; shape 1} n 14558
4.1.2. MODAL PROPERTY
  LX-SYNONYM {ability 2; power 3} n 3841132
4.1.3.  QUALITY
  LX-SYNONYM {quality 1} n 3338771
  LX-HYPONYM {condition 5; status 2} n 8520394
  LX-HYPONYM {disorder 1} n 8550427
4.1.4.  MENTAL PROPERTY
  LX-SYNONYM {psychological feature 1} n 12517
  LX-HYPONYM {need 5; require 3; want 5} v 675532
  LX-HYPONYM {need 6} v 675686
4.2. RELATION
  LX-SYNONYM {relation 1} n 17862
  LX-HYPONYM {relationship 1} n 8436181
  LX-HYPONYM {ratio 1} n 8457189
  LX-HYPONYM {unit 6; unit of measurement 1} n 8313335
4.2.1. SOCIAL RELATION
  LX-SYNONYM {social relation 1} n 18392
  LX-HYPONYM {relationship 3} n 8523567
4.2.2. POSSESSION
  LX-HYPERONYM {have 12; have got 1; hold 19} v 1256853
4.2.3. SPATIAL RELATION
  LX-NEAR-SYNONYM {spatial property 1; spatiality 1} n 3524985
  LX-NEAR-SYNONYM {be 9; occupy a certain area 1; occupy a certain position 1} v 1501697
  LX-NEAR-SYNONYM {space 1} n 15245
  LX-NEAR-SYNONYM {spacing 1; spatial arrangement 1} n 3535737
4.2.3.1 AMOUNT
  LX-SYNONYM {definite quantity 1} n 8310215
  LX-HYPONYM {distance 1} n 3536009
  LX-HYPONYM {magnitude relation 1} n 8454813
4.2.3.2. DIRECTION
  LX-SYNONYM {direction 7; way 8} n 5477069
  LX-HYPONYM {aim 4; bearing 5; heading 2} n 5477280
  LX-HYPONYM {course 7; trend 3} n 5477560
  LX-HYPONYM {path 3; route 2} n 5441398
4.2.3.3. PHYSICAL HAVE
  LX-HYPERONYM {have 12; have got 1; hold 19} v 1256853
```

A rich and powerful hierarchy can be built in a systematic way by extending these nodes top-down while combining classes with multiple hyponymy relations, as shown above. It is not necessary to encode all word meanings in a lexicon but only the most important ones at crucial points in the hierarchy of hyponymy relations (compare the Base Concepts in EuroWordNet).

A couple of things need to be said about the list of Base Types. First of all the number of Types can easily be doubled or halved. Instead of distinguishing subtypes of Roles we could have used one Base Type for Role in general, and similarly, our current enumeration of Roles is not complete, but just represents the selection of 164 Base Concepts. Something similar can be said for Change which can be minimalized or extended. The differentiation of a list of Base Types can thus be adapted to the needs of the application (e.g. the distinctions that play a role in selectional restrictions). The current differentiation is only based on a specific empirical procedure.

Furthermore, there are in principle two extreme ways in which an ontology with top-level distinctions, such as the Base Types, can be organized:

- tree of disjoint types
- flat lattice with explicit opposition relations

In the case of a classical tree, all the subdivisions are unified in a top, which branches between disjoint, opposite types. If applied to the first levels below Entity in the Base Types above, this would mean that the division at each level is fully disjoint and that distinctions cannot be introduced at multiple places of the hierarchy:

```
ENTITY
....NATURAL
........ANIMATE
............ANIMATE OBJECT
................ANIMATE OBJECT WHOLE
................ANIMATE OBJECT GROUP
................ANIMATE OBJECT PART
............ANIMATE SUBSTANCE
........INANIMATE
............INANIMATE OBJECT
................INANIMATE OBJECT WHOLE
................INANIMATE OBJECT GROUP
................INANIMATE OBJECT PART
............INANIMATE SUBSTANCE
....ARTIFACT
........ARTIFACT OBJECT
........ARTIFACT SUBSTANCE
```

Here we see that OBJECT/SUBSTANCE and WHOLE/PART/GROUP are not given as independent distinctions but only relative to other distinctions. The advantage is a simpler tree structure and implicit opposition of the types by disjointness of the branches. The latter means that word meanings cannot occur at multiple places in the hierarchy.

A disadvantage of this system is that it can only account for one structuring of semantics. If there is a word meaning for OBJECT but it can be applied to both ARTIFACT and NATURAL object, then this word meaning is too general to be put in this tree-structure. Another disadvantage is that it does not account for the generalization that ANIMATE

OBJECT and ARTIFACT OBJECT have *objecthood* in common. If there turns out to be little redundancy in the meanings of words (which is mostly the case when we are dealing with a general lexicon of a language), in the sense that for example being an OBJECT or PART cannot be predicted from any of the other distinctions such as NATURAL, ARTIFACT, INANIMATE or ANIMATE, the advantages of the clear tree structure will get lost because the distinction OBJECT and PART will have to be introduced at many other places.

An alternative way of organizing the Base Types is a lattice. In its most extreme form, all distinctions are given as a flat list of features and any concept can score for any combination of them. In that case we have a very flexible classification with as many subclasses as there are combinations of features. The disadvantage is that there is no redundancy at all and that combinations that are not allowed are not automatically excluded. However, it is up to the application to decide what is more important.

In practice we see combinations of a pure lattice structure and tree-structures. The above Base Types are presented as a relatively flat list, with some redundant subhierarchies. This specification can be elaborated by explicitly indicating what distinctions are disjoint and what combinations are typically combined. Typically, we can say that first level distinctions are disjoint or exclusive, as well as for example NATURAL versus ARTIFACT, OBJECT versus SUBSTANCE, and ANIMATE versus INANIMATE. However, we also see that many concrete word meanings represent combinations of classifications in terms of ORIGIN, FORM, COMPOSITION and ROLE, as is proposed in the Generative Lexicon approach ([Pus95a]). Furthermore, our conceptualization of the world may change over time, making redundant orderings of distinctions obsolete. In the future it may become impossible to maintain the redundancy that ANIMATE entities are also NATURAL. Artificial Intelligence can be seen as a form of ANIMATE ARTIFACT and genetic manip ulation may lead to instances of ARTIFACT ANIMATE. This illust rates the danger of building too much redundancy into the ontology.

Systems that have incorporated top-level ontologies like the above in their lexical specification are: EDR (§3.8), Mikrokosmos (§3.9.3), EuroWordNet (§3.6.3), Simple, Corelex (§3.12.2), and Acquilex (§3.12.3). In all these systems, a limited set of ontological distinctions is represented as a semi-lattice and each distinction is defined by means of Typed Feature Structures, opposition relations or definitions. There are then different ways in which these classes or distinctions can be assigned to the lexical items, in which case the (formalized) definition or specification can used for verification of the word meaning or for further specification.

6.1.4 Examples

To summarize, we present the following principles for building or evaluating a semantic network of hyponymy and synonymy relations:

- determine the purpose of the hierarchy: information retrieval, semantic inferences;
- determine the set of relevant semantic classes (given the purpose);
- determine the need for multiple class-assignment and encode the types of combinations of multiple classes;
- build the hierarchies top-down, starting from a limited set of basic distinctions, without skipping relevant levels or classes;

- differentiate co-hyponyms as disjoint or overlapping clusters when the semantic clusters grow;
- try to organize close and synonymous meanings into synsets to make sure that the semantic specifications hold for each synonym;

Below are some examples of word senses discussed above as they can be encoded using the EAGLES guidelines for lexical semantics standards (§6.9).

```
[ -ORTHOGRAPHY : being
  -WORD-SENSE-ID : being_1
  -BASE-TYPE-INFO : [ BASE-TYPE: ANIMATE
                      LX-RELATION: LX-SYNONYM]
                    [ BASE-TYPE: AGENT
                      LX-RELATION: LX-HYPONYM]
   SYNONYMS : life form_1; living thing_1; organism_1
   HYPONYMS : [HYP-TYPE: disjunctive
               HYP-ID: plant_1]
              [HYP-TYPE: disjunctive
               HYP-ID: human_1]
              [HYP-TYPE: disjunctive
               HYP-ID: animal_1]
              [HYP-TYPE: non-exclusive
               HYP-ID: parent_1]
              [HYP-TYPE: non-exclusive
               HYP-ID: child_1]]

[ -ORTHOGRAPHY : animal
  -WORD-SENSE-ID : animal_1
  -BASE-TYPE-INFO : [ BASE-TYPE: ANIMAL
                      LX-RELATION: LX-SYNONYM]
                     [ BASE-TYPE: OBJECT
                       LX-RELATION: LX-HYPONYM]
   SYNONYMS : animate being_1; beast_1;
   NEAR-SYNONYMS : creature_1; fauna_1; brute_1;
   HYPERONYMS : [HYP-TYPE: ?
                 HYP-ID: being 1]
   HYPONYMS : [HYP-TYPE: disjunctive
               HYP-ID: horse_1]
              [HYP-TYPE: disjunctive
               HYP-ID: cat_1]
              [HYP-TYPE: disjunctive
               HYP-ID: dog_1]
              [HYP-TYPE: non-exclusive
               HYP-ID: pet_1]
              [HYP-TYPE: non-exclusive
               HYP-ID: draught animal_1]]
```

```
[ -ORTHOGRAPHY : horse
  -WORD-SENSE-ID : horse_1
  -BASE-TYPE-INFO : [ BASE-TYPE: ANIMAL
                      LX-RELATION: LX-HYPONYM]
                    [ BASE-TYPE: OBJECT
                      LX-RELATION: LX-HYPONYM]
   SYNONYMS : Equus_caballus_1
   HYPERONYMS : [HYP-TYPE: conjunctive
                 HYP-ID: animal_1]
                [HYP-TYPE: conjunctive
                 HYP-ID: equid_1]
                [HYP-TYPE: non-exclusive
                 HYP-ID: pet_1]
                [HYP-TYPE: non-exclusive
                 HYP-ID: draught_animal_1]
   HYPONYMS : [HYP-TYPE: disjunctive
               HYP-ID: mare_1]
              [HYP-TYPE: disjunctive
               HYP-ID: stallion_1]]

[ -ORTHOGRAPHY : cell
  -WORD-SENSE-ID : cell_1
  -BASE-TYPE-INFO : [ BASE-TYPE: ANIMATE
                      LX-RELATION: LX-HYPONYM]
                    [ BASE-TYPE: SUBSTANCE
                      LX-RELATION: LX-HYPONYM]
                    [ BASE-TYPE: PART
                      LX-RELATION: LX-HYPONYM]
   HYPERONYMS : [HYP-TYPE: conjunctive
                 HYP-ID: entity_1]
                [HYP-TYPE: conjunctive
                 HYP-ID: part_1]]

[ -ORTHOGRAPHY : milk
  -WORD-SENSE-ID : milk _1
  -BASE-TYPE-INFO : [ BASE-TYPE: NATURAL
                      LX-RELATION: LX-HYPONYM]
                    [ BASE-TYPE: SUBSTANCE
                      LX-RELATION: LX-HYPONYM]
                    [ BASE-TYPE: DRINK
                      LX-RELATION: LX-HYPONYM]
   HYPERONYMS : [HYP-TYPE: conjunctive
                 HYP-ID: liquid_1]
                [HYP-TYPE: conjunctive
                 HYP-ID: part_1]]

[ -ORTHOGRAPHY : wheel
```

```
 -WORD-SENSE-ID : wheel_1
 -BASE-TYPE-INFO : [ BASE-TYPE: ARTIFACT
                     LX-RELATION: LX-HYPONYM]
                   [ BASE-TYPE: OBECT
                     LX-RELATION: LX-HYPONYM]
                   [ BASE-TYPE: PART
                     LX-RELATION: LX-HYPONYM]
                   [ BASE-TYPE: INSTRUMENT
                     LX-RELATION: LX-HYPONYM]
  SYNONYMS : steering_wheel_1
  HYPERONYMS : [HYP-TYPE: conjunctive
                HYP-ID: artifact_1]
               [HYP-TYPE: conjunctive
                HYP-ID: part_1]]
```

6.2 Meronyms

Meronymy is defined as a lexical part-whole relationship between elements. A meronomy is a lexical hierarchy whose relation of dominance is the lexical relation of meronymy. Prototypical examples are given by body parts and car parts. “Finger” is a meronym of ”hand” which is a meronym of “arm” which is a meronym of “body”. The ”inverse relation” is called *holonymy*. “Body” is an holonym of ”arm” which is the holonym of “hand” which is the holonym of ”finger”. The *co-meronymy* relationship is one between lexical items designing sister parts (arm, leg, head are co-meronyms of body). Meronymy is different from taxonymy because it does not classify elements by class. That is to say, the hierarchical structuring of meronymy does not originate in a hierarchy of classes (toes, fingers, heads, legs, etc. are not hierarchically related).

Linked to the concept of meronymy are the notions of whole, piece, part and generality. It is important to draw the line bewteen a *piece* and a *part*. A part implies a whole independent unit which can be linked to another part via attachment means. For instance, the hand is linked to the arm via the wrist. All parts are necessary to form the whole. For instance all body parts are necesserary to form a whole body. However, as this is too restrictive Cruse ([Cru86]) introduces the notion of *facultative meronym* (”handle” is a facultative meronym of “door”). He also tries to give criteria to decide whether a lexical item is the meronym of another one. A possible test is:

> X is a meronym of Y if and only if sentences of the form *Y has X* and *X is a part of Y* are normal when the noun phrases X and Y are interpretated generically.

Lexical items which satisfy the two frames are good candidates: “a hand has fingers” and “fingers are part of a hand” opposed to “a husband has a wife” and “a wife is a part of a husband”.

Note that, although we are in an inclusion/overlap relation it can be that meronyms are more general than their holonyms. For instance, ”nail” is more general than its holonym “toes” as it can also be part of a finger as well. This is why Cruse introduces the notions of *super meronym* (”nail” is a super-meronym of “toes”) and *hypo holonym* (”toes” is a hypo-holonym of “nail”) Questioning the meronym definition deeper and deeper, Cruse suggests other distinctions such as, for instance the one of *holo-meronymy* ("blade" is holo-meronym of "leaf").

As far as lexical resources are concerned, WordNet gives meronym/ holonym relations. There are 5 types of information related to the notion of meronyms.

- member of holonyms
- part of holonyms
- has part meronym
- all meronyms
- all holonyms

For the word "car', Wordnet provides the following information for each of the above categories:

```
Member Holonyms of noun car: 1 of 5 senses of car
     Sense 2 - car, railcar, railway car, railroad car
                MEMBER OF: train, railroad train

Part Holonyms of noun car: 3 of 5 senses of car
     Sense 3 - car, gondola
                PART OF: airship, dirigible
     Sense 4 - car, elevator car
                PART OF: elevator, lift
     Sense 5 - cable car, car
                PART OF: cable railway, funicular, funicular railway

Part Meronyms of noun car: 2 of 5 senses of car
     Sense 1 - car, auto, automobile, machine, motorcar
                HAS PART: accelerator, throttle, throttle valve
                HAS PART: accelerator, accelerator pedal, gas pedal,
                          gas, hand throttle, gun
                HAS PART: auto accessory
                HAS PART: automobile engine
                HAS PART: automobile horn, car horn, motor horn
                HAS PART: boot, luggage compartment, trunk
                HAS PART: buffer, fender
                HAS PART: bumper
                HAS PART: car door
                HAS PART: car mirror
                HAS PART: car window
                HAS PART: door
                HAS PART: fender, wing, mudguard
                HAS PART: first gear, first, low gear, low
                HAS PART: floorboard
                HAS PART: glove compartment
                HAS PART: grille, radiator grille
                HAS PART: high gear, high
                HAS PART: hood, bonnet
                HAS PART: rear window
                HAS PART: reverse
```

```
             HAS PART: roof
             HAS PART: running board
             HAS PART: second gear, second
             HAS PART: sunroof, sunshine-roof
             HAS PART: suspension, suspension system
             HAS PART: tail pipe
             HAS PART: third gear, third
             HAS PART: transmission
             HAS PART: window
             HAS PART: blinker, turn signal, turn indicator,
                       trafficator
  Sense 2 - car, railcar, railway car, railroad car
             HAS PART: suspension, suspension system
```

Meronyms can be found in most off-the-shelf dictionaries (monolingual as well as bilingual) but they are usually not given in a systematic way. For instance in "le Petit Robert" one can find "fallenge" at the entry for "doigt" and "humerus, biceps, triceps" at the entry for "bras". Some dictionaries are more consistent than others in providing a meronym list but criteria on deciding when meronyms are given remain unclear.

We recommend the following encoding for meronyms and holonyms expressed using the EAGLES guidelines for lexical semantics standards (§6.9).

```
meronym : [ MER-TYPE: (member | substance | part)
            HOLS: [HOL-ID: word-sense-id
                   REL-TYPE : (exclus | conjunct | non-exclus)]+
            MER-ID : word-sense-id ].

holonym : [ HOL-TYPE: (member | substance | part)
            MERS: [MER-ID: word-sense-id
                   REL-TYPE : (exclus | conjunct | non-exclus)]+
            HOL-ID : word-sense-id ].
```

6.3 Antonyms

The definition of antonym offerred in section 2.3 says that "W1 and W2 are antonyms if they have most semantic characteristics in common, but also differ significantly on at least one essential semantic dimension". While [Cru86], on which the discussion in section 2.3 is based, contains an extended discussion of antonyms, including the definition of a considerable number of subtypes, the high-level message is that "in spite of the robustness of the ordinary speaker's intuitions concerning opposites, the overall class is not a well-defined one, and its adequate characterisation is far from easy". Given this lack of theoretical clarity about the notion, it is not surprising that it finds little expression in existing lexical resources and even less in current applications. It follows that any guidelines proposed here must be extremely tentaive.

6.3.1 Resources Including Antonyms

The review of lexical semantic resources discussed in chapter 3 of this report reveals that the only resources examined which record antonymical lexical relations are WordNet and EuroWordNet 3.6. In these resources antonymy is represented as a binary relation between synsets.

6.3.2 Applications Using Antonyms

The review of NL and IS areas of application and component technology discussed in chapters 4 and 5 reveals that *no* current application or component technology appears to be making use of antonymical lexical semantic relations. However, a number of areas clearly could make use of such information:

- **NLG** The lexical choice component of a natural language generation system (section 4.5) will frequently need to decide between using a negation or an appropriate lexical item which expresses the negated meaning. Clearly knowledge of antonyms is essential here for generating unstilted text, especially where multiple negatives may be involved.

- **WSD/WC** Certain word sense disambiguation algorithms (section 5.3) should be able to make use of antonymical information by relying on the fact that antonyms exhibit paradigmatic similarities. If a word W1 which has senses S1 and S2 appears in a context C then in the absence of any collocational information about W1 in this context, it may prove useful to look at antonyms for each of W1's senses. If an antonymical sense of sense S2 can be realised as W2 and there is strong collocational evidence for W2 in context C (and not for any realisation of an antonym of S1) then S2 is probably the appropriate sense. For example, to disambiguate *hot* in the expression *hot news*, between the sense of *having a relatively or noticeably high temperature* and *fresh, recent*, it may prove useful to rely on information about occurrences of expressions like *old news*, *stale news* etc. across a corpus, together with the absence of expressions such as *cold news*, in order to pick the second sense. Similar techiques could also be used in support of word clustering (section 5.1).

  Of course, other applications (e.g. machine translation) or component technologies (e.g. parsing) whose performance can be enhanced by improved word sense disambiguation will benefit indirectly from this use of antonyms.

- **IR/IE** Information retrieval (section 4.2) and information extraction systems (section 4.3) may well benefit from being able to perform inferences based on knowledge of antonymical lexical relations. For example consider the query *Tell me about cities in the north of England*. A document about Bristol which contains the assertion that *Bristol is a city in the south of England* is clearly not relevant, but this knowledge relies upon knowing in turn that north and south are opposites. Clearly retrieval and extraction performance should be enhanced if use is made of such knowledge.

6.3.3 Recommendations Pertaining to Antonyms

As indicated above, given the difficulty in adequately characterising antonyms or opposites, any recommendations must at this time be very tentative.

A minimal recommendation would be that any lexical semantic standard should record a simple binary relation of antonymy where possible between word senses (as done in WordNet and EuroWordNet) as it is clear from the discussion in the preceding section that this information may well be of use in a variety of applications and component technologies.

In addition, it may prove useful to distinguish subtypes of the opposite or antonym lexical relation of the sorts discussed by Cruse [Cru86]. These subtypes include:

- **complementaries** Complementaries are opposites which divide partition a conceptual space, leaving no room for intermediates (e.g. *true:false*, *dead:alive*, as opposed to, say *good:bad*).

- **gradables** Adjectives, typically, which can take modifiers which express degrees (contrast *very hot* with *? very dead*).

- **antonyms** Cruse uses *antonym* to refer to lexical opposites which are gradable, not complementary (i.e. do not bisect domain), express a variable degree of some property (speed, length, etc.) and which have the characteristic that when intensified each member of the pair moves in opposite direction with respect to the underlying property (e.g. *very heavy* and *very light*).

  Antonyms are further divided into **pseudo-comparatives** and **true comparatives** depending on whether they assert possession of the base term to a greater degree, or only assert possession of a greater degree of the property of which the base term represents one pole. For example, *heavier* is a pseudo-comparative since something can fail to be heavy but still be heavier than something else; *hotter* is a true comparative since it is unnatural to say that something is, e.g. cold, but hotter than something else.

- **directional opposites** In their simplest form these terms indicate opposite directions along a straight-line path – e.g. *north:south, up:down.*

  Various subcategories of directional opposites can be identified such as **antipodals** (terms indicating the extremes of directional opposition – *top:bottom*); **reversives** (verbs denoting motion or change in opposite directions – *ascend:descend*); **relational opposites** (terms indicating opposed relations between pairs of entities – *above: below*).

These distinctions form only a part of Cruse's account, and serve only to begin to indicate the wider range of detail that could be recorded concerning antonym relations in a lexical semantic standard.

6.4 Subject Domains

Subject domain codes are available in a number of machine readable dictionaries and lexical databases, e.g. LDOCE, CIDE (*Cambridge International Dictionary of English*), *Chambers 20th century Dictionary*, Oxford-Hachette French-English bilingual and the GLDB (The Göteborg Lexical DataBase). They tend to differ both in granularity and classification. For example, in the GLDB (§3.5) there are unstructured 95 codes, in LDOCE there are 100 main fields and 246 subdivisions (see §3.2.2), and in CIDE (§3.4) there are about 900 codes organized into 4 main levels as shown in Table 6.2. Such differences present a potential problem for standardization. However, sufficient progress can be made by choosing the richest system of classification as reference point — the CIDE scheme in our case — and selecting those domain codes which occurr most often across dictionaries. Tables 6.4 and 6.5 show the results of choosing subject domain codes which occurr in at least three of the dictionaries and lexical databases mentioned earlier: LDOCE, CIDE, Chambers, Oxford-Hachette French-English bilingual and the GLDB. In presenting the results of this merger, the hierarchical relations among selected domain codes which are encoded in CIDE are maintained; nodes subsuming selected domain codes have been included to maintain the hierarchy backbone.

```
A - Agriculture, Animal Husbandry & Forestry
   AG - Agriculture & Arable Farming
   AH - Animal Farming & Husbandry
      BEE - Beekeeping & Apiculture
      BLK - Blacksmithing
      BRE - Breeds&Breeding
         BREC - Breeds of cat
         BRED - Breeds of dog
   FI - Commercial Fishing (not whaling [WHAL])
   FR - Forestry (not tree names [TRE])
      WOO - Types of Wood
   HO - Horticulture&Gardening
   SS - Soil Science
   VE - Veterinary Medicine
```

Table 6.2: Example of subject domain codes in CIDE

Another issue for standardization in this area concerns the relevance of thesaurus-like semantic classifications such as the one given in LLOCE (§3.2.3) which at a coarse granularity level are often reminiscent of subject domain codes, as shown in Table 6.3. No attempt has been made so far to integrate thesaurus-like top level classifications into the "standard" hierarchy for subject domains presented in Tables 6.4 and 6.5.

6.5 Word Co-occurrence relations

This section illustrates importance and role of word co-occurrence relations in a standard for lexical encoding, as inferred from their use in NLP applications and their availability in lexical resources. It then provides some preliminary recommendations for their encoding.

6.5.1 Word co-occurrence relations: a typology

By 'word co-occurrence relations' we mean any relation holding between two lexical items which are simultaneously present in a single structure such as the verb–complement relation or the relation holding between constituents of a compound. From such a broad definition, it appears rather clearly that this area selected for the standard of lexical encoding covers in principle a variety of linguistic constructions ranging from 'collocations', 'multi–word expressions' and 'compounds' to 'selectional restrictions' whenever expressed extensionally.

In this chapter, selectional restrictions are already dealt with as information typically associated with arguments in predicate frames (see §6.8) which is usually expressed in terms of semantic categories, possibly specified at different degrees of granularity (whether in terms of base types or of synonym clusters), and extended through use of functional restrictions. However, given the widely acknowledged difficulty of appropriately characterising selectional restrictions imposed by predicates on their arguments through use of semantic categories, another representational option offered by the standard for lexical encoding is in extensional terms, i.e. by listing typical lexical fillers of a given argument position. This kind of option

```
<A> Life and living things
<B> The body, its functions and welfare
<C> People and the family
<D> Buildings, houses, the home, clothes, belongings, and personal care
<E> Food, drink, and farming
<F> Feelings, emotions, attitudes, and sensations
<G> Thought and communication, language and grammar
<H> H Substances, materials, objects, and equipment
<I> Arts and crafts, sciences and technology, industry and education
<J> Numbers, measurement, money, and commerce
<K> Entertainment, sports, and games
<L> Space and time
<M> Movement, location, travel, and transport
<N> General and abstract terms
```

Table 6.3: Major codes in LLOCE.

may turn out to be useful when an appropriate semantic category cannot be found, or when this information has to be used by example–based NLP systems. In the former case, the representation of selectional restrictions in extensional terms could be seen as a preliminary step towards the creation of a new semantic category. Last but not least, word co-occurrence patterns are also very useful as illustrative examples of general selectional restrictions (i.e. which can be conveyed through semantic categories); in fact, they can ease the process of manual checking and inspection of lexical resources and also their updating.

Due to this overlapping with other areas selected for the standard of lexical encoding, in this specific context the expression 'word co-occurrence relations' will mainly – but not only – refer to idiosyncratic patterns of word co-occurrence, where the selection involves specific lexical items rather than a general semantic class. A typical example of word co-occurrence pattern is represented by the English example *break the rules* (Benson 1986) where the selection operates at the lexical level: despite *rule* and *regulation* can be classified as synonyms, *rules* can be broken, but *regulations* normally cannot.

As pointed out above, word co-occurrence relations can also be used to represent selectional restrictions in terms of typical patterns of use when there are no appropriate semantic categories available. This is particularly useful with very specific selectional restrictions, e.g. *diagonalize* which is done only to *matrices*. Another case in point is represented by the lack of available semantic categories capturing the dimension of semantic similarity which is relevant in the specific context: for instance, one would normally talk of *toasting bread* and *grilling meat*, and not vice versa, in spite of the fact that the actual operations may not be otherwise distinguishable. To account for the different selectional restrictions of the two verbs, in principle features such as the state of the food when the process starts ([+cooked]/[+raw]), or the kind of result ([+dry]/[+juicy]) could be resorted to. A simpler and less arbitrary way to deal with them is by treating them as lexical relations operating on individual words, (e.g. between *grill* and *meat*, *toast* and *bread* etc).

Applications based on word co-occurrence information range over both natural language analysis and generation. In analysis, word co-occurrence information is used for instance

```
Agriculture, Animal Husbandry & Forestry
    Agriculture & Arable Farming
    Forestry
    Horticulture & Gardening
    Veterinary Medicine
Building & Civil Engineering
    Building
       Building Tools
       Interior Decoration
    Buildings
       Archaeology
       Architecture
    Civil Engineering
    Surveying
Arts & Crafts
   Drawing, Painting & Fine Arts
   Colours
   Knitting, Sewing, embroidery, needlework
   Photography
   Sculpture
Food & Drink
    Food, Cooking & Baking
       Cookery & Cooking Terms
Earth Science & Outer Space
    Astronomy
    Calendar
       Periods of Time
    Geology
       Mineralogy
    Geography
    Weather, Climate & Meteorology
    Countries, Nationalities and Continents
Finance & Business
    Business & Commerce
       Accounting & Bookkeeping
       Marketing & Merchandising, Commerce
       Postal System, Terms and Service
       Work: Staff and the Workforce
    Economics & Finance
    Insurance
    Manufacturing
       Textiles
    Taxation
Sports, Games & Pastimes
    Games
    Horsemanship
    Hunting & Fishing
       Fishing
       Hunting
    Motor Racing
    Pastimes
       Numismatics
History & Heraldry
   Heraldry
   History
Media & Publishing
   Media
       Advertising
       Broadcasting
          Radio
          Television
       Publishing
          Books & Book Publishing
          Newspapers & Newspaper Publishing
          Printing
Crime and the Law
Language & Literature
    Language
       Dialect
       Linguistics
       Phonology & Phonetics
       Tropes, e.g. metaphor, metonymy, onomatopoeia
    Literature
       Poetry
```

Table 6.4: Subject Domain codes from more than 2 dictionaries - Part 1 (sources: CIDE, Chambers, Oxford-Hachette, LDOCE, GLDB)

```
Life Sciences
   Biology
       Anatomy
       Ecology
       Physiology
    Zoology
       Entomology
    Botany
Maths, Physics & Chemistry
    Chemistry
    Mathematics & Arithmetic
       Geometry
       Statistics
       Trigonometry
    Weights & Measures
    Physics
       Optics
Entertainment & Music
    Dance & Choreography
    Motion Pictures, Cinema & the Film Industry
    Music
       Musical Instrument
    Theatre
Medicine, Health & Psychology
    Health Issues
      Hygiene
    Medicine
       Dentistry
       Pharmacy, Pharmacology & Pharmaceuticals
       Surgery
    Psychology
Names of Plants and Animals
Education
    Schools
    University
Religion & Theology
Society
    Sociology
    Titles & Forms of Address
       Professional Titles
    Anthropology & Ethnology
    Culture: Relationships, Customs and Lifestyle
       Kinship & Familial Terms
       Social Security & State Benefits
Technology
    Communications
       Telephony
       Telegraphy
    Data Processing & Computer Technology
    Electricity, Electronics & Electrical Engineering
    Engineering
       Mechanical Engineering
       Mining Engineering & Quarrying
    Metallurgy
Politics, Diplomacy & Government
    Overseas Politics & International Relations
       Diplomacy & Mediation
Travel and Transport
    Traffic
    Aviation, Aeronautics & Aircraft
    Aerospace
    Automotive Vehicles
    Nautical Terms & Seafaring
       Navigation
War & the Military
    Military (the armed forces)
       Navy
Mythology, Occult & Philosophy
    Mythology & Legend
    Occult
       Alchemy
       Astrology
    Philosophy
       Logic
       Metaphysics
Clothes, Fashion & Beauticulture
    Fashion Industry & Modelling
    Clothing
```

Table 6.5: Subject domain codes from more than 2 dictionaries - Part 2 (sources: CIDE, Chambers, Oxford-Hachette, LDOCE, GLDB)

to: i) resolve syntactic ambiguities such as PP attachment or, as far as Italian is concerned, disambiguate functional relations; ii) for word sense disambiguation and anaphora resolution. In the generation process, word co-occurrence information, especially collocations, plays a crucial role.

6.5.2 Towards a standard representation of word co-occurrence relations

In this section, we present some preliminary recommendations for a standard representation of word co-occurrence relations, which build on the typology which emerges from available lexical resources and from NLP applications requirements. In what follows, we will briefly illustrate these guidelines, and discuss some of the issues at stake when dealing with word co-occurrence relations.

Word co-occurrence information in the proposed EAGLES word sense entry

COLLOCATIONS is the attribute describing the typical co-occurrence relations involving the word being described. Within the standard proposed for lexical semantic encoding (§6.9, this attribute enters into the definition of both a word sense entry and of an argument in a predicate frame as shown in the feature structures below:

```
word-sense-entry ->
 [ -ORTHOGRAPHY : string
   -WORD-SENSE-ID : word-sense-id
   -BASE-TYPE-INFO : base-type-info*
   SUBJECT-DOMAIN : subject-domain*
   SYNONYMS : word-sense-id*
   NEAR-SYNONYMS : word-sense-id*
   HYPONYMS : hyponym*
   ANTONYMS : antonym*
   MERONYMS : meronym*
   QUANTIFICATION : quantification
   COLLOCATIONS : collocation*
   SEMANTIC-FRAME : sem-frame
   ACTIONALITY : actionality ]

arg -> [ SELECTIONAL-RESTR : (base-type* | word-sense-id*)
         COLLOCATIONS : collocation*
         THEMATIC-ROLE : th-role ]
```

How

The value of the COLLOCATIONS attribute is a list of 0 to n 'collocation' elements. Each element of the list is represented as a conjunction of attribute-value pairs as described below:

```
 collocation ->
  [ COLLOCATE : word-sense-id+
     DIRECTION : (left | right)
     WORD-DISTANCE : [LOWER-LIMIT : integer
```

```
                  UPPER-LIMIT : integer]
DEPENDENCY : (h2d | d2h | d2d | h2h)
DEP_TYPE : dependency_type
PROBABILITY : probability-value* ]
```

specifying respectively:

- the attribute COLLOCATE records the typical collocates of the item being described, expressed in terms of sets of word senses;
- the attribute DIRECTION specifies the right or left location of the collocates with respect to the word being defined;
- the attribute WORD-DISTANCE specifies the distance range between the word being defined and the collocates: two different attributes – LOWER–LIMIT and UPPER–LIMIT – specify respectively the minimal and maximal distance between the co-occurring words;
- the attribute DEPENDENCY describes the relevant dependency configuration;
- the attribute DEP_TYPE carries further information about the dependency type holding between the co-occurring words;
- the attribute PROBABILITY expresses the strength of the association of the co-occurring words in one or more reference corpora, possibly specialised with respect to a given domain; mutual information could for instance be used as a value of this attribute.

Some remarks are in order here. The DEPENDENCY attribute gives information about the dependency configuration being described, in particular about the relationship holding between the word sense entry or the argument (henceforth, the selection source) and the items in the COLLOCATIONS list (henceforth, the selection target). Four different dependency configurations are identified, following the lexical specifications elaborated within the SPARKLE project (LE-2111, Work Package 2, http://www.ilc.pi.cnr.it/sparkle.html):

h2d : from head to dependent, i.e. the selection source is the head and the selection target syntactically depends on it;

d2h : from dependent to head, i.e. the selection source syntactically depends on the selection target;

d2d : from dependent to dependent, i.e. both the selection source and the selection target syntactically depend on some other element;

h2h : from head to head, i.e. both the selection source and the selection target are heads of different but typically co-occurring constructions.

Note that this type of definition of the DEPENDENCY attribute permits the simultaneous encoding within the same COLLOCATIONS list of both the elements subcategorised for by the word being described and the elements subcategorising for it.

DEP_TYPE expresses the grammatical function with respect to the head:

- if the head is the word being defined, then DEP_TYPE specifies the grammatical function of the item in the COLLOCATIONS list relative to the head;
- if the head is one element of the COLLOCATIONS list then:
  - DEP_TYPE of elements other than the head indicates the grammatical function with respect to the head (see example below);
  - DEP_TYPE of the element specified as the head indicates the grammatical function of the word being defined with respect to it (see example below).

Whether the element in the COLLOCATIONS list is the head or not is expressed at the level of the DEPENDENCY attribute where the first element of the relationship label (e.g. "h" in "h2d") always refers to the selection source, namely the word sense entry or the argument; "h" in first position means that the selection source is the head of the described construction, "d" that the selection source is a dependent on the selection target.

The DEP_TYPE attribute further specifies the basic functional information already conveyed by DEPENDENCY. The range of proposed values of the DEP_TYPE attribute follows the general recommendations put forward in the framework of the EAGLES Lexicon–Syntax group (see http://www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html), refined and articulated in a practical and coherent proposal in the framework of the European SPARKLE project (LE-2111). This hierarchical scheme of grammatical relations is summarised below:

dependency_type –> arg | mod | arg_mod
arg –> subj | comp
subj –> ncsubj | xsubj | csubj
comp –> obj | clausal
obj –> dobj | obj2 | iobj
clausal -> xcomp | ccomp
mod –> ncmod | xmod | cmod

It should be noted that in order to avoid redundancy between standards of lexical encoding pertaining to different levels of linguistic description (namely, syntax and semantics), functional information in the proposed standard is restricted to the COLLOCATIONS attribute only, since grammatical function is already included in the standard proposed by EAGLES for what concerns syntactic subcategorisation.

Note that not every information contained in the feature structure 'collocation' is to be obligatorily specified. A necessary specification is constituted by the COLLOCATE attribute through which the basic word co-occurrence information can be conveyed. Recommended specifications concern the location (DIRECTION and WORD–DISTANCE attributes) and the dependency type (DEPENDENCY) of the item in the COLLOCATIONS list with respect to the selection source. Finally, the DEP_TYPE and PROBABILITY attributes convey useful specifications: the former further refines the functional information already carried by the DEPENDENCY attribute, the latter gives information about the association strength of the co-occurring words, also relative to different corpora taken as representative, e.g., of different domains.

What for

These preliminary recommendations provide the user with flexible means to represent in a lean way the entire typology of word co-occurrence information sketched in section 6.5.1 above. Both idiosyncratic word co-occurrence patterns and typical patterns of use representative of the selectional restrictions of a given predicate can be couched in this representation format straightforwardly. Idiosyncratic collocational patterns can only be encoded through the COLLOCATIONS attributes. By contrast, when dealing with selectional restrictions, the user is presented with different options. (S)he can decide, on the basis of the application requirements and the available linguistic information, whether:

- to encode selectional restrictions imposed by predicates on their arguments through use of semantic categories (as suggested in §6.8);
- or to encode them in extensional terms, namely by listing the typical lexical fillers of a given argument position with the related linguistic information (this option is particularly useful when the appropriate semantic category is missing);
- or to combine the different encodings in order to integrate selectional information, conveyed through semantic categories, with illustrative examples (this apparent redundancy may ease the process of manual checking and inspection of the lexicon).

As illustrated above, the COLLOCATIONS attribute enters into the definition of both the word sense entry and of the arguments in the predicate frame.

At the level of the word sense entry, all kinds of co-occurrence patterns involving the item under definition can be encoded, regardless of the fact that the word being defined is the head or not.

Consider an Italian idiomatic expression such as *tagliare la corda* 'run away' lit. 'cut the rope'. It can be encoded either under the headword *corda* or under the verb *tagliare*; both options are foreseen within the proposed standard. Assume that we want to encode this information under *corda*, which could make things easier for dictionary lookup. In such a case, the COLLOCATIONS list will contain two items:

```
[ COLLOCATE : tagliare_1
   DIRECTION : left
   WORD-DISTANCE : [LOWER-LIMIT : 1
                    UPPER-LIMIT : 2]
   DEPENDENCY : d2h
   DEP_TYPE : obj ]

[ COLLOCATE : la_1
   DIRECTION : left
   WORD-DISTANCE : [LOWER-LIMIT : 0
                    UPPER-LIMIT : 1]
   DEPENDENCY : h2d
   DEP_TYPE : mod ]
```

The first item represents *tagliare*, the head of the entire construction, while the second one describes the definite article *la* immediately preceding the entry headword, to be obligatorily realised.

The specification of the COLLOCATIONS attribute within the ARG slot of a predicate frame can be used to convey information about possible restrictions on the argument collocates. This is the case, in Italian, of the sense 'reach a certain value' of the verb *arrivare* which is typically associated with the co-occurrence of the indirect argument with a numeral modifier as in the case of *è arrivato a 100 libri / a 100 voti / a 3 case* 'he reached 100 books / 100 votes / 3 houses' where the selectional restrictions associated with the argument cannot be of any help to discriminate this verb sense from other senses of the same verb.

6.6 Tense/Time/Aspect

As a starting point, let us recall the important, though often neglected, distinction beween **aspect** and **actionality** (**aktionsarten**). In §2.2.2 we suggested that aspect, that is the perfective-imperfective distinction, be better seen as intimately bound to the notion of complete vs. non complete events. Furthermore, aspect is made available by means of verbal morphemes, either bound or free. In this respect it patterns together with tense in that the relevant information is, or should be, provided to NLP systems by morphological analysers. The latters usually provide for some kind of simple features-values pairs. In the case at hand we might expect something like *aspect=perfective*, or even just *perfective*. The NLP system must be able to cope with such information, but we do not regard them as part of lexical semantics. Therefore, we do not expect to find them listed in LKBs, unless they directly encodes inflected forms, and we will not consider tense and aspect information in this section any more.[2] For our purposes, and given the discussion in §2.2.1, we will simply assume that the NLP system exploiting our hypothetical lexicon has access to aspectual information concerning verbs and predicates, in the form of a simple $\pm$*perfective* feature opposition.

On the other hand, verbal actionality refers to the internal constitution of events in the extension of the verbal predicate. For the purposes of the discussion here, we will simplify both the theoretical debate on actionality, and the way such a debate has affected computational lexical semantics, by distinguishing the LKBs encoding actionality into two basic types: those proposing a static view of actionality, and those enforcing a somewhat more dynamic encoding. A static LKB will talk basically try to assign each verb to one of a predetermined number of actional classes, usually reducible to the Vendlerian ones. They can do this either by exploiting taxonomic resources (e.g. lexical or ontological hierarchies), or by means of suitable features. A LKB taking a dynamic view of actionality, on the other hand, will provide verbal entries with information that are to interact with other knowledge coming from, say, direct objects to determine the actional behaviour of the complex predicate so formed. Therefore, dynamic LKB take the compositional nature of actionality seriously, following the leads of such authors as Verkuyl, [Ver93], and Krifka, [Kri89] and [Kri90].

It is important to notice that the difference between the two kinds of LKBs should not be reduced to the type of notational device they use, e.g., to the fact that static LKBs uses taxonomic encodings whereas dynamic ones don't. The real differences lie in the purported use of such information. In a static LKB, the fact that a verb belongs to a given class, e.g. *activities*, rather than to another one, e.g. *accomplishments*, fully determines and exhaustively characterises the grammatical and interpretive behaviour of that verb. This is very clear in the typical uses of Vendlerian distinctions: *statives* never combine with *in*-adverbials, resist the progressive, etc... On the hand, the information a dynamic LKB attaches to verbs contribute

[2] But see the Eagles recommendation on standards for morphological and grammatical information.

to specify their behaviour only in conjunction with information coming from other sources: structural information, lexical information from arguments, and so on, this way, at least partially, reconstructing the properties of the traditional Vendlerian classes.

The distinction between static and dynamic ways to encode actionality can greatly affect the behaviour and performances of actual NLP systems. Simplifying, static classifications obviously tend to involve a lesser amount of computational burden (they purport to directly capture the relevant facts). However, their empirical inappropriateness results in a severe limitation as to their capability of covering relevant phenomena. Thus, an NLP system using a statical LKB will not be able to distinguish the following two occurrences of the verb to eat:

(74) a. John ate apples (for one hour / *in one hour).

b. John ate an apple (*for one hour / in one hour).

Most probably, the LKB will classify *eat* as an accomplishment verb, and the NLP system will expect that one such a verb be telic, irrespectively of the direct objects it combines with. Thus, it will accept the ungrammatical option when the object is a bare plural so that the NLP system as a whole will assign the verbal predicates in (74a) the wrong temporal properties. On the other hand, dynamic LKBs are more computationally demanding, but also much more capable of capturing phenomena such as the one exemplified above.

6.6.1 Verbal Actionality in Lexical Resources

6.6.2 Static LKBs

In this section we review some LKBs exploiting what we have called a static encoding schema for actionality.

UNITRANS [Dor93]. This system exploits a set of three features: $\pm$*d(ynamic)*, $\pm$*t(elic)*, and $\pm$*atomic*. The $\pm d$ feature distinguishes between events and states; $\pm t$ distinguishes between culminated and non-culminated events; $\pm a$ distinguishes between atomic (durationaless) and non-atomic events. Such a classification basically captures Vendlerian categories. For instance, *[+d, -t, -a]* can be taken to identify an activity, *[+d, +t, -a]* a classical accomplishment, as in *John ate an apple*, and so on.

DIONYSUS [Nir93]. This systems provides information about phasal properties (e.g. inchoativity), durational properties (e.g. $\pm$*prolonged)* and iteration. Such information can be correlated with actionality, as discussed here, by, e.g., associating *-prolonged* to achievements. Less relevant, to our purposes, are the phasal and iteration properties which are better seen as belonging to the domain of aspect, as defined above, rather than to actionality.

(Generalised) Upper Model [Bat90], [Bat94]. The GUM features a hierarchy which permits, at least to a certain extent, to reconstruct actional properties as ordinarily conceived. For instance, a verb such as *to eat* is classified as a **Dispositive Material Action**, which is a subtype of **Directed Action**. From such an information it can be gained that the direct object (the entity playing the actee role) is affected by the event/action. This, in turn, permitts to classify such predicates as *eat an apple* as Vendlerian accomplishments. Similarly, *to prepare*, as in *John prepared a cake*, is a **Creative Material Action**, this way entailing that once the process has come to its natural end an object of type cake has come to existence. Again, this permits to classify prepare as an accomplishment. In general, it seems possible to gather from the (G)UM enough information to (automatically, or semi-automatically) derive the actional properties of verbs (and complex predicates).

6.6.3 Dynamic LKBs

ACQUILEX, [San91], attempts at a more dynamic classification of aktionsarten which also tries to address the interactions with thematic roles. All this is done within a TFS formalism. Thus, the traditional vendlerian distinction of verbal aktionsarten is reconstructed in the type hierarchy by means of the two types **stative** and **dynamic**. These are then combined with the two krifkean types **cumulative** and **quantized** to obtain a quadripartite classification of verbal predicates. Such a classification is then connected with the properties of thematic (proto-)roles to allow for the transfer of properties from the objectual to the eventive domain.

6.6.4 Verbal Actionality in Applications

Machine Translation. Most interlingua based MT systems use actional information. This is the case of such systems as Kant and Tamerlan (exploiting the DIONYSOUS LKB).

Information Extraction. Most current IE systems do not seem to use actional information.

Natural Language Generation. Actional information are used by many deep generation systems, especially in the initial planning phase (strategic generation). This is clear in NLG systems exploiting the Pennman system (which, in turn, adopt the UM as the higher level for the LKB). The availability of actional information is even more important in multilingual NLG, e.g. GIST [], given the different ways languages express the dependencies between the actional properties of predicates, the properties of the the arguments and the verbal aspect.

6.6.5 Suggestions and Guidelines

6.6.6 Static Encoding

The simplest and more direct way to perform static encoding of actionality remains the resort to vendlerian categories. Therefore, we suggest that verbs be assigned to one of the following: *statives*, *activities*, *accomplishments*, *achievements*. This can be done either by directly having such categories in the basic linguistic ontology, or by exploiting a feature *actionality* with the labels above as values.

6.6.7 Dynamic Encoding

Given the discussion in §2.2.1 and in the introduction to this section, we suggest that the purpose of a dynamic encoding schema for actionality be that of contributing to the detection, at processing (parsing) time, of the status of complex predicates as to the telic/non-telic distinction. Such a distinction, beyond being important for interpretive purposes, also affects the kind of temporal adverbials the complex predicate can combine with (*for*- vs. *in*-adverbials). Thus it has sufficiently many consequences to deserve standards and guidelines.

Our starting point is a bipartite distinction of predicates by means of the feature $\pm process$ (equivalently, $\pm dynamic$), together with a bipartition of NPs (DPs) according to the $\pm SQA$ feature. As discussed in §2.2.1, the term *+process* applies to vendlerian activities and accomplishments, and, in a sense, to achievements as well (but see below), this way encoding the *stative* vs. *eventive* distinction as well.

(75) a. to love, to seem, to be happy, ... -process

b. to run, ro eat, to go, to read .... +process

On the other hand, $\pm SQA$ (Specified Quantity of substance A) distinguishes such NPs (DPs) as *the/an apple* from bare plurals such as *apples* and bare mass nouns such as *meat, beer, wine, beef.*[3]

(76) a. an/the apple +SQA

b. apples -SQA

c. meat -SQA

d. three pieces/kilos of meat +SQA

Such a general scheme should be coupled with information specifying the distinguished argument/constituent which, in construction with the verb, forms the complex predicate whose actionality properties we are trying to characterise. [4] In most cases such an argument is the direct object:

(77) a. eat an apple/two pears/every cake, write a book

b. drink beer/water, write papers

However, it can also be another constituent, e.g., the directional argument, as is the case with motion verbs:

(78) a. John ran home in/*for three minutes. (telic)

b. John ran *in/for three minutes. (atelic)

As said at the beginning of this section, the purpose of all such information is to dynamically derive whether the complex predicate is *telic* or not. Recall that the notion of telicity refers to the possibility of a predicate to produce true telic readings once combined with a suitable aspectual value, that is perfectivity. On the other hand, atelicity refers to the impossibiity of such a reading. Such information can be computed (by the parser), adopting the following schema:

(79) a. +process & +SQA $\rightarrow$ telic

b. +process & -SQA $\rightarrow$ -telic

This means that whenever the verb is a *process* which combines with a *+SQA* distinguished argument, then the resulting complex predicate is *telic.*[5]

(80) a. John ran(+process) home(+SQA). $\rightarrow$ +telic

b. John ran(+process) (-SQA). $\rightarrow$ -telic

Actuality of the telos can only be determined by means of aspectual information. Generally, a *telic* predicate gives raise to an actual telos only if its aspectual values is *perfective.*

[3]The information captured by $\pm SQA$ is basically compositional, in that it it is a property of the whole NP. It is determined by: the status of the basic noun as to the *mass/count* distinction, and the kind of determiner/quantifier/classifier introducing the NP. For more on these matter see §6.7.

[4]The proper place for such an information is the subcategorisation frame, thematic role structure, or whatever other means is available for specifying the kind of arguments the verb combines with. Here, as for other cases of information entering the determination of compositional actionality, we simply assume that such an information be available.

[5]Notice that in (80b) we proposed that a missing directional argument be understood as *-SQA*.

This kind of information is, again, to be computed at parse time, by putting together other bits of basic information. To this end, the following schema is suggested:

(81) a. +telic & +perfective $\rightarrow$ +telos

b. +telic & -perfective $\rightarrow$ -telos

That is, if the complex predicate is telic (according to the schema above) and its aspectual value is *+perfective*, then the telos is actual (the event has culminated, reaching its natural end-point). If the aspectual value of the predicate is *-perfective*, then there is no actual telos.[6] On the other hand, if the complex predicate is *atelic* (no potential telos/culmination), the aspectual value does not change such a situation:

(82) -telic $\rightarrow$ -telos

This way we encode the fact that vendlerian activities, or certain occurrences of accomplishments, remain atelic even when perfective:

(83) a. John ran *in/for three hours.

b. +process & -SQA $\rightarrow$ -telic
-telic & +perfective $\rightarrow$ -telos

Finally, we suggested that vendlerian achievements are such that they always give raise to actual teloses. Therefore they should be lexically notated as *+telos*.

To sum up, the proposal exploits four features, $\pm$*process*, $\pm$*telic*, $\pm$*telos* and $\pm$*perfective*, for verbs. Furthermore, it hypotheses a fifth feature for NPs, namely $\pm SQA$. The only proper lexical feature is $\pm$*process*. $\pm$*telic* and $\pm$*telos* can both be used in the lexicon (for achievements), or being the result of syntactic-semantic computations. $\pm$*perfective* is the only aspectual feature we need at present. Together with the nominal feature $\pm SQA$, it is provided by modules other than the lexicon (e.g., the morphological analyser and the parser, respectively).

The advantages of the present proposal are:

- it is simple (using only a handful of binary features). Yet it is powerful enough to permit fine distinctions among the actional behaviour of both simple and complex predicates;
- in many cases, a lexicon with the proper annotations can be automatically produced from a lexicon providing for some sort of static encoding of actionality. For instance, if the source lexicon provides for vendlerian-like distinctions (e.g., by exploiting the suggestions of the previous section) then all is needed is to map *stative* into *-process*, *activity* and *accomplishment* into *+process*, and *achievement* into *+telos*;
- it can easily be extended to deal with the interaction of actionality with other aspectual phenomena, e.g. the progressive. Exploiting the intensionality of such verbal forms, the actuality of the telos in such cases as *John is eating an apple* is in the scope of the intensional progressive operator, thereby turned into a potential telos;

[6]Notice that we used the feature $\pm$*telos* to indicate the **actual** presence of a telos/culmination in the semantics of the sentence, distinguishing such a case from the one in which the telos is simply **potential**, notated by means of the $\pm$*telic* feature.

- it highlights the connections between the eventive and the objectual domain, and specifies the division of labour between the various modules of an NLP system.

The main disadvantage is:

- it might be difficult be difficult to extend the schema to aspectual values not captured by the $\pm$*perfective* distinction, e.g., inchoative and continuative aspect.

6.7 Quantification

6.7.1 Criteria

Quantification has been addressed in section 2.7 because this phenomena affects the interpretation of nouns and nominal expresions. The range of phenomena included in this heading have been mostly treated by *Formal Models*, because it addresses the semantic of terms as referring constructions. The set of *quantifiers*, i.e. expressions that quantify over a set of entities, includes elements from differents categories. Thus, we have to prevent the reader that in EAGLES - Lexicon - Morphosyntactic Phenomena there is no reference to this class of elements as categorical or part of speech distinction. What in grammar description have been called *indefinite pronouns*, *indefinite determiners*, *adverbs* and *articles* can also be grouped toghether in what is called the group of 'Quantifiers' because of its semantic characteristics. In a logical sense, these particles have universal o partitive meaning or can be said to be selected from the Noun meaning a 'quantified' reading (see §2.7).

(84) Some animals fly

(85) All animals fly

(86) Most of the animals fly

(87) Every animal flies

(88) No animals fly

All these expressions can be said to select a certain quantity of what is referred as 'animal' to predicate over. Because of its implications with respect to 'true conditions' (i.e. some of these sentences can be said to be false) the notion of 'quantifier', has been carefully studied in the framework of Formal Semantics and Model Theoretic interpretation.

Without entering into the details of the formalization of its characteristics in a model-theoretic framework, we must nevertheless acknowledge that studies on this ground have been very successful in describing certain lexical characteristics of these elements. Some of these explanations have contributed to explain puzzles never treated before by other linguistic approaches.

Referential interpretation of quantified NP have special denotations which might be taken into account. Some properties relating to the cardinality of the denotations of quantified NP's have consequences on the interpretation of other phenomena, for instance in Aktionsart calculation, the cardinality of the internal argument has to be taken into account for aspectual characterisation of full sentences:

(89) I smoke several/two/many cigarrets in an hour

(90) I smoke cigarrets *in an hour

Their contribution to the referential nature of the noun can be explained in terms of 'weak' vs. 'strong' quantifiers, as to explain certain phenomena such as:

(91) *There were most men in the room

(92) There is a man in the room

We can say that quantifiers, that is the 'determiners' or 'specifiers' of a given NP, can affect the referential nature of the N they are specifying.

Strong determiners/quantified expression are those that force a pressuposition of existence, or in other words, that their true conditions are independent of the model we are considering. Let's see an example:

(93) Every linguist in the room had a book

(94) Many linguists in the room had a book

In the first case, 'every linguist' has a pressuposed existence, because we can say that if there are linguists in the room the sentence will be true, but if there is none it will also be true. 'every' is said to be a 'strong' quantifier. In comparison with sentence 2 we can see better the difference. Because we have to look at the linguists in the room to see which are the ones that 'had a book', and in the case there are no linguists in the room the sentence cannot be said to be true. 'many' is said to be a 'weak determiner'. This characterisation of quantified expressions is what can be used to understand why a sentence such as (95) is good in comparison with (96):

(95) There are many linguists in the room

(96) ?There is every linguist in the room

We can say that 'strong' quantified expressions are not informative, because there is this pressuposition behind, and thus they sound 'odd'. There are other consequences of the occurrence of such determiners. To take these into account we must first make a distinction between KIND-denoting noun phrases and all other noun phrases (cf. Carlson 1977). A KIND-denoting NP is an object which cannot be said to be a phisycally individuated entity, but rather it is conceived as -possibly- what is common to all the individuals to which a given name can apply. Hence, simplifying very much, we can say they have no referential nature. For all the rest, for the non KIND-denoting noun phrases, we can establish the following classification:

- weak determiners trigger existential or non-pressupositional reference (i.e. introduce a discourse entity), and
- strong determiners trigger non-exitential or pressupositional reference, which can be *definite* (i.e. it refers to an existing discourse entity) or a *proportional* reference (i.e. it introduce a subset within an existing set).

With reference to strong determiners, we must say that many noun phrases are ambiguous between the existential and the proportional reading, but the crucial differences associated can be stated as indicated (following Loebner 1987).

Existential

1. head noun referentially new
2. partitive paraphrase impossible
3. determiner counts the whole denotation of the head noun
4. immediate subsequent pronominal anaphora refer to the whole denotation of the head noun
5. head noun stressed, obligatory
6. determiner unstressed, omissible
7. VP stressed or unstressed

Proportional

1. head noun referentially given
2. partitive phrase possible
3. determiner counts part of the denotation of the head noun
4. immediate subsequent pronominal anaphora refer to part of the denotation of the head noun
5. head noun unstressed, omissible
6. determiner stressed, obligatory
7. VP stressed

Another property of quantifiers has been studied in relation with the inferences drawn from the meaning of a verb. It is important to see how the possible entailments of sentences and the relations between the sets of entities involved in interpreting sentences containing quantified expressions are blocked by means of quantifiers.

(97) Every fish swims

(98) Every fish moves

If we consider that the set of entities that *swim* are contained in the set of entities that *move*, it must follow that if (97) is true, (98) must also be true. We say that the quantified expression denoted by *every fish* is *upward entailing*. *some, many* and *every* are upward entailing. But *no, few* and *two* are not, as we can see in the following sentences, they can be said to be *downward entailing* as the entailments are in the other direction. If (100) is to be true, (99) should also be true:

(99) No fish walks

(100) No fish moves

This property of quantifers is very helpful to explain the behavior of some quantifiers, such as that of English: 'ever', 'any', which are contextually restricted to occur in negative contexts, and are said to have *negative polarity* items, but it is easy to see that it is not jus negative elements that provide the right environment for the proper use of these elements, as we can see in the following examples.

(101) We did not read *any* books

(102) *We read *any* books

(103) I have not *ever* been to Pisa

(104) *I have *ever* been to Pisa

(105) No student who knows *anything* about phonology would *ever* say that

(106) Every student who know *anything* about phonology would (**ever*) say that

For these negative polarity items we can say that they can appear in expressions that denote downward-entailing functions. For instance in sentence (101) we can say that if we did not read any books, then we did not read romance books, old books, linguistics books, etc.

6.7.2 Quantifiers in applications

A proper theory of quantification has been mostly used in order to handle access to databases. Beyond that, information about how these phenomena have been treated in current applications is scarce. Probably, this little implementation of quantification issues is due to the relatively few systems that are ready to handle semantics on a Formal Semantic basis. Thus, most of the systems have been limited to implement grammar rules that handle combinatorial information of each type of quantifier and 'countable-mass' nouns. Nevertheless, this information is useful in applications such MT and Language generation.

6.7.3 Guidelines

We recommend that the lexical properties of determiners concerning quantification be minimally encoded in terms of quantification strength and direction of entailment, as indicated below (see also §6.9).

```
quantification : [ Q-STRENGTH : (weak | strong)
                   Q-ENTAILMENT : (upward | downward) ]
```

6.8 Predicate Frames

This section surveys the use of predicate frames in NLP applications and their availability in lexical resources, either experimental or related to precise applications. It then gives a few structured simple recommendations.

Predicate frames mainly include verbs, but also predicative nouns, adjectives and prepositions. Adverbs are not included here since they are often higher-order predicates whose analysis and encoding is delicate (similarly for some adjectives). The structures reviewed here are: argument structure and verb subcategorization frames (see also the Eagles report on syntax), selectional restrictions (as a natural complement to subcategorization frames), thematic roles, semantic templates, and, at a macro-level, verb semantic classifications.

6.8.1 Predicate frame elements

This section describes the elements usually found in predicate frames, and reviewed in the above mentioned report.

Predicate frames are an important element in applications and in lexical resources. It incorporates most of the lexical semantics elements, since predicates are often the 'kernel' of propositions.

A distinction is made here between real-world applications and experimental ones, which propose solutions for the near future. Importance (I), availability (A) and common encoding (CE) on the one hand and feasability (FEA) if not used or if resources are not available, on the other hand, are given.

The following elements are taken into account in this document.

Argument structure (AS) This notion, which is usually very simple, often includes the number and the syntactic nature of the arguments (the subcategorization frame). This notion is not necessarily made explicit in lexical data bases, but can straightforwardly be infered from more complex descriptions. It is therefore easily available, and its encoding is quite straightforward, since it minimally amounts to the number of arguments the predicate has. The number of arguments of a predicate is however not a very crucial information per se. It should however be noted that determining the arity of a predicate is not a trivial matter, it depends on the approach (minimalist or maximalist, see below under recommendations) and on the underlying syntactic theory (e.g. lexical redundancy rules dealing with syntactic alternations usually cut down the number of subcategorization frames, reducing therefore the proliferation of arities for a given verb-sense).

Selectional restrictions (SR) Selectional restriction systems can be classified on the basis of their granularity. Verbs make different demands on the properties of their arguments. Sometimes such demands are quite general, while other times the demand is very specific. At one extreme, we have verbs like *eat*, *drink*, *pass away*, which show wide collocational ranges for their subjects and/or objects (corresponding to semantically coherent sets of nouns denoting food, liquids, animates or humans). At the other extreme, we find verbs like *bark* which, in its literal meaning, may be only said of *dogs*; *devein* which is used only in reference to *shrimps*; *diagonalize* which is done only to *matrices*. On this basis, 'general' and 'specific' selectional restrictions can be distinguished; note that both restriction types are required by NLP applications such as MT, syntactic and semantic disambiguation or information retrieval. From this it follows that the representation of selectional restrictions, to be really accurate (e.g. in natural language generation, to avoid overgeneration) needs to cope with ranges of details from coarse to fine grained. However, systems of rules dealing with metaphoric or metonymic usages contribute to limit this complexity.

Selectional restrictions are represented in different ways ranging from a small set of general labels to sophisticated ontologies (see section on ontologies). The real problem is to identify the set of relevant features to be used to further specify the general semantic categories, set which is virtually open-ended. A different representational option for selectional restrictions

is in extensional terms, by specifying the collocational set associated with a given argument position; this kind of representation is particularly useful for example-based NLP systems. At the level of encoding, selectional restrictions are often included into subcategorization frames, e.g.:
GO: [NP(+human), PP(+localization)].
Selectional restrictions are sometimes specialized when associated with application domains (and corresponding thesauri), which makes their re-usability somewhat delicate. Most, if not all, systems make an intensive use of selectional restrictions. Verbs and prepositions should systematically be associated with selectional restrictions whereas this is much more delicate for adjectives and for predicative nouns.

Thematic roles, thematic grids (TR) As underlined in chapter 2, roles are very diverse, and have very diverse names and definitions. They may also be more or less fine-grained. They are often used as a relatively simple level of semantic representation, when no other representation is available. They establish a relation between arguments and their related predicate such as agent or location. Thematic roles are not used as a syntactic criterion, or to determine syntactic functions in language generation. They constitute a kind of general representation of the world (or of a domain) rather than a linguistic representation. From that point of view they can be viewed as 'macros', encoding various properties (proto-roles or other approaches).

Thematic roles are often available in lexical resources, and used in e.g. IR and IE systems, their encoding is quite simple, but the fuzyness of definitions, may make their re-use problematic. Roles may be coded separately for a few arguments or may appear as a grid. This is particularly the case for verbs, whereas for prepositions only the 'object' is encoded (the other element of the relation being outside the PP). Adjective do not have in general thematic roles associated since they can be too diverse (in particular when used in small clauses). Deverbal nouns (or nominalizations) usually inherit the grid of the verb they are 'derived from', but arguments are all optional.

In some descriptions, a predicate-sense gets as many grids as it may have different syntactic configurations (e.g. alternations) while in other approaches lexical redundancy rules cut down this rather unmotivated proliferation.

Finally, note that assigning thematic roles to arguments tends to be quite subjective, there is indeed e.g. a kind of continuum between some roles. Also for some verbs, it is not possible to find appropriate thematic roles. This point is addressed in more depth in the recommendation-guidelines section.

Semantic Templates, Frame Semantics (ST) This notion covers a large number of notions and structures in applications. They may be sometimes a reformulation of thematic roles and of selectional restrictions, but they may also, in experimental systems, be fragments of semantic networks, lexical functions, scripts or primitive-based representations (e.g. the Lexical Conceptual Structure or lexical semantic templates), or frame nets (Fillmore's approach). Qualia structures may also be viewed as semantic templates, playing a specialized function. These are exemplified in the recommendation section.

Semantic templates are not often available in applications, nor in large lexical resources. In experimental resources, their encodings is based on different assumptions and is subject to large variations, which makes re-usability quite delicate. Their use in applications is however often desirable (e.g. in MT, automatic summary construction, lexicalization in NLG, and IE).

The description of semantic templates can be made much easier if one structures them in terms of sets of semantically close verbs (e.g. directed motion verbs, causal verbs, transfer of possession verbs) because of a number of regularities which have been observed. Encodings can be made from the most general to the most specific elements, using underspecified descriptions. Base types, ontologies and verb semantic classes are extremely useful from that point of view.

Semantic classes (VCL) Classification methods have been developed to structure knowledge according to different paradigms: alternations, ontology top nodes or base-word types, general semantic classifications (e.g. motion, possession), etc., they are rarely explicitly used so far, but this is desirable. A particular attention has been devoted to verbs, and, to a certain extend, to adjectives.

Verb classes mainly serve to better organize verb-senses in order to improve the coherence and the homogeneity of their semantic descriptions (verbs of a given class share a lot of semantic aspects, which should be encoded in a similar way). The most accessible and re-usable classes are those defined in WordNet and EuroWordNet for English, other classes have been defined from syntactic criteria (alternations), but seem to be less usable and contain a quite large number of exceptional behaviours. Classes for other languages may be substantially different, depending on the classification assumptions and the language properties. Encodings are relatively simple and availability is good.

The belonging of a predicate to a given class allows predictions to be made, which facilitates acquisition of new lexical items.

6.8.2 Predicate frames in applications and in lexical resources

Below, we first review if and how the above different elements are used in applications and in component technologies. We then survey their existence, encodings and availability in lexical resources. Verb frames will be the main focus of this synthesis since verbs are the category which received the largest attention.

Predicate frames in applications

Machine Translation We cover here section 4.1 of the report. It turns out that verb frames are mainly used in experimental MT systems based on interlingual forms:

Project name	Verb frame elements used
Kant	TR, SR
Principtran	AS, TR, SR ST (mainly the LCS) and VCL

In Principtran, the semantics of a large set of verbs is represented by the Lexical Conceptual Structure. Verb semantic classes (VCL) organize this complex description into classes.

The other MT systems not based on interlingua forms basically use subcategorization frames and selectional restrictions.

Information retrieval and extraction These applications are reported in sections 4.2 and 4.3 of the report. Very little information about verb frames is used for these two types of applications. While badger and Snomed-UMLS use a kind of case-frame to enhance constrained retrievals, information extraction systems mainly use subcat frames with some relatively coarse-grained ST and SR. It would however be much desirable for these applications to use thematic roles for example to label arguments when predicate-argument forms - describing states or actions - -are extracted. Similarly, VCL would be useful to somewhat normalize the information extracted: the prototypical element of the class could be the norm, avoiding a proliferation of predicates with close meanings. This could be alternatively solved by means of semantic relations describing notions of semantic proximity.

Natural Language Generation These systems are described in section 4.5 of the report. It seems that very few elements of verb frames are used, besides ST and SR, in real applications (which aren't very numerous), but they are almost systematically used in experimental systems. TR are not used in conjunction with linking rules, but some advanced frames (scripts and frames, and Melc'uk functions) are used in experimental systems, often as the starting point of the generation procedure. Note also the development of large-size systems integrating a large number of NL aspects specific to the Systemic Grammar approach.

Component technologies Component technologies seem to rely quite heavily on statistical observations, which, quite often, entail modes of knowledge representation that differ from classical linguistic representations. Nevertheless, some applications require quite a lot of elements from verb frames. Word clustering uses SR quite extensively, in association with statistical co-occurences. Word sense disambiguation is the component that makes the most extensive use of verb frame information since the elements in relation with the word to disambiguate are very relevant for determining the right sense.

There are however some knowledge-based approaches that rely heavily on SR and ST, also on WordNet and the LDOCE categories:

Project name	Verb frame elements used
Locolex	AS, subcat frames, extraction of SR and ST from the OUP dictionary
Eurotra	AS, SR and VCL (via ontological classes)
ET10-75	AS, ST

Lexical Resources

Let us now review different lexical resources (Chapter 5 of the report). We have grouped them into 3 sets: (1) general purpose resources, which are quite diverse, (2) Lexicons for MT, which turn out to be very rich and also quite homogeneous, and (3) experimental lexicons, which are more innovative at the level of ST, which are much more elaborated, and where some introduce Qualias of the Generative lexicon.

As can be seen, lexical resources offer much more than the applications effectively use so far. This is obviously not surprising. Here is now a synthesis of the elements the reader can find in 3:

General purpose lexical resources

Resource name	Verb frame elements used
LDOCE, LLOCE	AS, SR, marginal uses of VCL
GLDB	AS, SR, VCL by domains, description of modifiers
WordNet	AS, global VCL from major ontological cat., some ST
EuroWordNet	AS, SR, TR, some ST
EDR	some weak forms of TR and ST

Lexicons for Machine Translation

Resource name	Verb frame elements used
Eurotra-Spain	AS, SR, TR, ST
CAT2	AS, SR, TR
METAL	some forms of TR, some restricted SR
LOGOS	AS, SR, TR, ST
Systran	AS, SR, some forms of argument typing

Experimental lexicons

Resource name	Verb frame elements used
Corelex	AS, SR, Qualias of the Generative Lexicon
Acquilex	AS, SR, ST, VCL (on a semantic basis), TR, some forms of Qualias
ET10-51	AS, SR (associated with Cobuild), ST
Delis	AS, TR advanced ST

6.8.3 Guidelines

In this section, we present a few preliminary recommendations with some elements of discussion whenever necessary. It should be kept in mind that these recommendations remain very basic and simple.

A. Argument structure and subcategorization frames

(i) Issues As discussed in this section and in §2.6, defining the arity and the subcategorization frame of a predicate (see Eagles report on syntax) is both trivial and complex. For ease of exposition, selectional restrictions are discussed separately.

The issues at stake are the following:

- It is sometimes hard to make a distinction between arguments and some modifiers which are quite close to the verb. Some approaches prefer to speak in terms of 'distance' of an argument w.r.t. the predicator. In this category fall arguments denoting e.g. instruments and means.

- Some approaches tend to develop a subcat frame per syntactic configuration (e.g. passive, middle, etc.) while others develop a single subcat frame and lexical redundancy rules that generate derived subcat frames. Note that lexical redundancy rules can be directly constructed from alternations (see below section on verb semantic classes). A third approach is to have a single subcat frame per sense and to encode in the grammar the alternations the predicator may undergo.

- argument structure is often associated with the subcat frame, where syntactic categories are given. In a number of cases, that category is not unique (e.g. NP and S), then, a disjunction of possibilities in the frame must be introduced. Similarly, the order of categories in the frame may be more of less constrained (e.g. in languages with case marks, where the constituent order is less constrained).

- in the same range of ideas, it is also important to mention those arguments which are compulsory and those which can be ommited. Incorporated arguments are not explicitly part of the argument structure (they can be viewed as arguments in a pre-lexical system).

- Finally, we think it is worth introducing some information about prepositions for PPs. Since prepositions are highly polysemic, we suggest to introduce a general characterization of prepositions such as: location, source, dative, accompaniement, etc. These labels have in fact a lot in common with thematic roles. Prepositions are also quite often used metaphorically.

(ii) Scheme An simple argument structure can have the following form (see e.g. HPSG for more complex content organizations):

verb: 'eat': [NP(obligatory), NP()].
verb: 'chat': [NP(obligatory), PP(prep=accompaniment), PP(prep=about)]. (A chats with B about C).

B. Selectional restrictions

This is a much debated topic in Eagles. Since selectional restrictions are directly related to ontologies, we suggest the reader to consult this section. There are however a few points to discuss at this level.

The different forms of selectional restrictions can be hierarchically organized as follows:

1. general base types (about 70),

2. subject domains (about 150), technical domains can be introduced at this level,

3. Wordnet synsets labels (or clusters of synset labels),

4. functional restrictions (e.g. container-of, vehicule-for).

It is possible to only use one of these forms, or they can be mixed. WordNet synsets are by far the most precise ones and implicitly include the 2 first items.

An argument may be associated with several selection restrictions, in that case a logical language is necessary to express the different restrictions. It turns out that well-known and widely-used operators such as AND, OR, NOT, EXCEPT and By-DEFAULT are sufficient. Preferences (via statistical observations) can also be stated. Expressions are seldom complex.

(i) Issues

- the granularity level of selectional restrictions depends on the type of application being developed and on its desired robustness. Roughly, there is a kind of compromise to find between accuracy and expressivity on the one hand and efficiency on the other hand. For language generation, it is clear that these restrictions must be very accurate, if at all possible.

- at a more theoretical level, the granularity and the nature of selectional restrictions depends on the theory of meaning considered. If one wants to define narrow senses as in WordNet, then restrictions must be quite refined. If one wants restrictions in order to make a few checkings and to solve a majority of ambiguities, then simple restrictions should suffice in most cases. Finally if one wants to develop generative devices and rules to handle metaphors and metonymies, then selectional restrictions must be quite precise and type shifting rules must be developed to account for various changes (e.g. food → 'entity with intellectual content' as in *devour meat* and *devour books*).

- Coarse-grained restrictions overgenerate, but they also permit the taking into account of sense variations, and thus introduce some flexibility in parsing.

- In general, very generic terms are associated with top level restrictions whereas more specialized terms get more specialized restrictions.

(ii) Scheme An simple argument structure with selectional restrictions can have the following form:

verb: 'eat': [NP(obligatory, +human), NP(+edible-entity)].
verb: 'chat': [NP(obligatory, +human), PP(+human, prep=accompaniement), PP(+event, prep=about)]. (A chats with B about C).

For example, in Mikrokosmos, these informations are encoded as follows (from the Mikrokosmos database) for the verb *cut*:

```
cut  - V1
        syn: root: tag{0}
                   cat:  V
                   sem:  tag{00}

             subj: tag{1}
                    cat:NP
                    sem: tag{11}

             obj:  tag{2}
                    cat: NP
                    sem: tag{21}

             pp-adjunct:  tag{3}
                    root: with,using, ...
                       obj: tag{3}
```

```
                          cat: NP,opt+
                          sem: tag{31}

 sem: tag{00} CUT
   agent: tag{11} human
   theme: tag{21} object
   instrument: tag{31} object
```

Subcategorization frames (44 different forms) considered in Mikrokosmos are, e.g.:

```
[np,v].
[np,v,pp(at)].
[np,v,poss,np].
[np,v,self,pp(on)].
[np,v,np]).
[np,v,np,pp(with)].
[np,v,adv(easily)].
[np,v,np(and)].
[np(and),v].
[np,v,np,pp([off,of])].
[np,v,np,pp(off)].
[np,v,pp([off,of])].
```

The subcat frames are extracted from B. Levin's descriptions ([Lev93]) with a few filters to avoid redundancies. Note that adjuncts have been added whenever required by alternations. Tags are just labels as in type feature structure systems.

C. Thematic roles

While argument structure, subcategorization frames and selectional restrictions are constraints on proposition structures, thematic roles and also semantic templates are usually used as semantic representations. Thematic roles are a very coarse-grained representation of meaning where arguments (and also modifiers) of predicates are labelled with thematic roles. This approach may be sufficient for some applications, in particular in IR. In connection with argument structure, thematic roles are grouped into sets of 2 or 3 elements, where, roughly, a thematic role is associated with each argument. These sets are often called thematic grids.

(i) Issues

- Thematic grids correspond to subcategorization frames, therefore, the same approach must be used w.r.t. the number of grids associated with each word-sense (see above).

- It is often postulated that a given word-sense has a unique thematic grid, this is a good principle to start constructing grids, but counter-examples abound. In the same ragne of ideas, several roles can be associated with a single argument position to express its different facets (visible in more accurate semantic representations), e.g. the giver in *give* is the agent and the source.

- Although some theories have associated hierarchies of thematic roles with grammatical functions (e.g. subj, obj1), this is not used in applications because of the instability of hierarchies.

- Thematic roles do not cover all situations: some argument positions don't have corresponding roles.

In some applications, the standard thematic roles have not been used because it was felt that they did not totally correspond to what was necessary. In that case, new generalizations (i.e. new roles) have been derived from sets of verbs where a given argument was felt to play an identical 'role'. New roles are then discovered step by step, upon examination of verbs. This is particularly useful for restricted application domains where one may want to elaborate various forms of e.g. patients, agents or instruments.

(ii) Scheme We now suggest lists of thematic roles. To make this document readable we propose 3 sets of roles, the first is simple and basic (but may be useless), the second has a much wider coverage and has been used in experimental applications and the third one views roles as sets of properties (proto-roles and the like).

1. **A basic list of thematic roles** A simple list of roles, as the one given in §2.4 and 2.6 can be: *agent, patient (or theme), localization, experiencer and instrument.*

2. **A middle-size list of roles** The list below, from J. Allen's book (Natural Language Understanding) seems to be a good compromise between a short list of roles, insufficiently expressive, and a long, detailed one, which may be more difficult to use (ex. 3 below):

   - Causal agent or causal instrument,
   - Theme or patient,
   - Experiencer: the person involved in perception or a physical or psychological state,
   - Beneficiary,
   - At: for location, possession, time, etc. (i.e. ontological domains),
   - To: location, possession, etc.
   - From: with the same subfields as above,
   - Path,
   - Co-agent: secondary agent in an action,
   - Co-theme: secondary theme in an exchange.

Similar-sise systems can be found in e.g. Pundit, developed at Unisys in the last decade.

3. An extended list of thematic roles This hierarchically organized set can be used as a whole, or parts can be extracted (e.g. subtrees, or top nodes).

- the agent (ag): a general notion of agentivity, the degree of agentivity may be more or less strong at this level. Subtle agentive distinctions follow (abbreviations for grids are suggested between parenthesis):

  - effective agent (ae): the agent that makes the action

  - initiative agent (ai): the agent that takes the initiative of an action
  - volitive agent (av): the agent that wants the action to be done.

- general theme (tg): themes may be any type of entity, including humans;
  - holistic theme (th): the theme not affected in its integrity.
  - incremental theme (ti)
    * incremental beneficiary theme (tib)
    * incremental victim theme (tiv), tib and tiv are used when there is a subtancial gain or loss for the theme.
  - causal theme (tc)
  - consequence theme (tcons)
  - theme experiencer (exp) someone who enjoys or suffers from something or from someone. (somewhat close to tib, tiv for humans).

- localisation (loc) : may be subtyped spatial (spat), temporal (temp) or abstract (abs).
  - source (src)
  - position (pos) : a fixed location.
    * absolute position (pa)
    * relative position (pr)
  - goal (but)
  - destination (dest): the destination is effectively reached (as opposed to goal which is a direction, not necessarily reached),

- means (moy)
  - instrumental means (mi)
  - non-instrumental means (or indirect means) (mni)

- manner (ma)

- identification (id)

- accompaniement (acp), indicates accompaniement, co-occurence and plural arguments (for verbs like mix or agglomerate).

- amount or quantity (amount).

3. Roles as clusters of properties Thematic roles can also be viewed as clusters of semantic properties, including inferential schemas. They contribute to the definition of verb semantic classes. The most well-known roles (or proto-roles) are the proto-agent and the proto-patient. Other roles are: causation, change, locative (cause of movement), formal (creation or destruction of objects), emotive (changes in emotional attitudes) and matter (changes in shape, size, matter, etc.).

D. Semantic templates

Semantic templates are designed to represent the meaning of lexical items and, by composition of several templates, to represent the meaning of entire propositions. Semantic templates cover a large spectrum of representations, from very coarse-grained ones to very refined ones, including a number of pragmatic elements. Therefore, it is not possible to present here a set of recommendations that takes into account all the aspects of semantic templates. We simply present recommendations related to a few, relatively well-known, semantic templates which are used in a number of applications or easily available in resources.

Semantic templates are often designed for specific tasks (i.e. certain forms of MT, for certain pairs of languages), therefore, if availability is 'average', re-usability may be much lower. Below is a set of templates, organized from the simplest to the most complex one. For illustrative purposes, we have chosen the verb *cut* which is neither trivial nor too complex. It also has a corresponding noun (a cut) and an adjective (in Romance languages (Fr: coupant, Sp: cortante)).

(a) Semantic labels: from thematic grids to semantic networks

For the verb cut, we have the following grid:
CUT: [agent, theme] (or incremental victim theme).

A simple semantic network has labels which are almost identical, it includes, in addition: (1) variables to represent NPs, PPs, and Ss (represented in what follows by capital letters), (2) labels for modifiers and (3) inter-argument relations of various forms (often paired with selectional restrictions, not included here since, but presented above). Note that the label names are often close to thematic role names, but labels in networks play a very different role, in particular, arguments as well as adjuncts are labeled. This is why, e.g., that 'cut' below has a goal label, not related to an argument but to a modifier. For cut, we would get:

```
CUT:
   agent: X
   patient - theme: Y
   goal: G
   instrument: W

   X has-goal G
   W acts-on Y.
```

As we said above, a thematic grid, where arguments and modifiers of a predicate are labelled by roles is a simple, but efficient, semantic representation, useful e.g. in IR and IE.

(b) Primitive-based representations: from LST to LCS LST (lexical semantic templates) is a very basic, but of much interest, form of semantic representation, a kind of precursor of the LCS (lexical conceptual structure). Between these 2 representations, we find a large number of primitive-based representations, often coming from artificial intelligence. Primitives have the advantage of introducing a quite high level of genericity, allowing therefore several forms of inferences, if required. Note that thematic labels are also primitives, but of a very different nature (but note that thematic roles can be derived from some LCS templates).

About primitives, both LST and LCS introduce general purpose primitives for concrete predicates. These primitives are often derived by analogy from movement verbs, but it is clear that they are not sufficiently adequate for a number of other verbs (e.g. verbs of communication, psychological verbs) and adjectives. Therefore, additional primitives need to be introduced for these cases. In applications, domain specific primitives can also be added in a small number. LST and LCS can be viewed as a framework, a methodology and a formal language that users can slightly enhance depending on their needs. These forms are also of much interest because of their link to syntax. Finally, notice that primitives related to preposition form a closed set of about 50 primitives, these latter primitives need not be enhanced, but only general purpose primitives such as GO, CAUSE, etc. need so. This seriously limits the risk of errors and misconceptions.

LST for cut (primitives are also in capital letters) may have different forms, here are a few, among the most common:
X CAUSE Y BECOME SEPARATED.
X USE Instrument to CAUSE Y BECOME SEPARATED.
X CAUSE Y GO AT part-of(Y).
where part-of is a function. Alternatively TOWARD can be used instead of AT. (Note that the verb cut has an interesting analysis in WordNet)

LCS for cut (others are given in the section below devoted to verb semantic classes):
$\lambda\ I, J\ [_{event}\ CAUSE([_{thing}\ I\ ],$
$[_{event}\ GO_{+loc}([_{thing}\ PART{-}OT(J)\ ],\ [_{path}\ AWAY{-}FROM_{+loc}([_{place}\ LOCATION{-}OF(J)])\ ])])]$.
Note that for metaphorical uses of cut (e.g. cut funds), only a few elements of the representation need to be changed.

The most important structures of the LCS are the following, where semantic fields have not been included for the sake of readability since there are several possible combinations for each (e.g. loc, poss, temp, char-ident, psy, epist):

1. $PLACE\ \rightarrow\ [_{place}\ PLACE - FUNCTION([_{thing}\ ])]$

2. $PATH\ \rightarrow\ [_{path}\ TO/\ FROM/\ TOWARD/\ AWAY - FROM/$
$VIA([_{thing/place}\ ])]$

3. $EVENT\ \rightarrow\ [_{event}\ GO([_{thing}\ ],[_{path}\ ])]\ /$
$[_{event}\ STAY([_{thing}\ ],[_{place}\ ])]\ /$
$[_{cause}\ CAUSE([_{thing/event}\ ],[_{event}\ ])]$

4. $STATE\ \rightarrow\ [_{state}\ BE([_{thing}\ ],\ [_{place}\ ])]\ /$
$[_{state}\ ORIENT([_{thing}\ ],\ [_{path}\ ])]\ /$
$[_{state}\ EXT([_{thing}\ ],\ [_{path}\ ])]$

PLACE-FUNCTIONS are symbols such as ON, UNDER, ABOVE, related to the expression of localization. Note that STAY is treated as a form of event.

It is important to note that we show here an example of semantic template under the form of LCS. It is clear that many variants are possible, with different primitives and different notations. However, in general, semantic representation principles remain the same.

(c) Scripts and frames Again, this is a very open field, with a quite long tradition in cognitive science. Scripts are used in specific applications where a lot of pragmatic knowledge is involved, e.g. inferences describing the consequences of an action on the current (application) world.

Here is an 'abstract' sample summarizing several facets of scripts, for the verb cut, variables are called 'slots', must be filled in by dedicated procedures:

```
CUT: action:   agent X, patient Y 'must be separable'.
mods: place P, manner M, cicumstance C, instrument I 'must be sharp'.
basic acts:
   X moves I towards Y.
   X makes pressure on Y.
   Y is separated in pieces P.
consequences: Y no longer exist.
              parts-of Y are in current world.
```

(d) Qualias Qualias of the Generative Lexicon are also a kind of semantic template, representing the environment, the parts, the use and the way the entity denoted by the word has been created. However, it seems that Qualias are better appropriate for arguments than for predicates. Some elements of the Qualia structure of an arguemt may then be incorporated into the semantic representation of the predicate it is the argument of, to account for the interactions between the semantics of the predicate and that of the argument. Availability is average (see Corelex at Brandeis, and the ongoing SIMPLE EEC project).

In SIMPLE, roles are structured in terms of templates, so as to make more explicit the role played by each element in a Qualia role. For example, the constitutive role contains e.g. the following templates: has-part, is-in, is-made-of, state.

E. Verb semantic classes

Verb semantic classes can be used for several purposes. The main purpose is to have verbs grouped by families which share a lot of semantic features, making semantic descriptions more economical and much more homogeneous and reliable, a crucial advantage in semantics. In terms of classes, we recommend to use either WordNet or EuroWordNet classifications which are very well adapted to this task (see section on WN).

Verb semantic classes based on alternations (from Beth Levin, most notably), can also be used but their semantic classification power is lower and they include a number of exceptions. However, they can be used as a very comprehensive set of lexical redundancy rules, generating different forms from a more basic one. For example, thematic grids can be generated for different syntactic configurations from the original grid, adding modifiers when they are required in the alternation. The following example was generated from alternations (and cross-classified with WordNet classes to identify senses) by Bonnie Dorr:

```
ag, goal(at).
ag, th, mod-loc, instr(with).
instr, th, mod-loc.
ag, th, manner(apart).
ag, th, src(off, off of).
```

`ag, th, pred(into), benef(for). etc.`

(Third line: grammatically: subject is instrument, then obj1 is th and iobj is mod-loc).

Similarly, LCS representations can be defined for each syntactic configuration. Here is a sample for the verb cut from Bonnie Dorr's database:

```
;;; [21.1.a]  Cut Verbs - Change of State
;;; WORDS: "chip, cut, saw, scrape, slash, scratch"
;;; THETA ROLES: _th
;;; SENTENCES:  "The bread !! easily"
(
:DEF_WORD "!!"
:LCS (go ident (* thing 2)
         (toward ident (thing 2) (at ident (thing 2) (!!-ed 9))))
:VAR_SPEC ((9 :conflated))
)

;;; [21.1.b]  Cut Verbs - Change of State / -on
;;; WORDS: "chip, clip, cut, hack, hew, saw, scrape, scratch, slash, snip"
;;; THETA ROLES: _instr_th,mod-loc()   ;; pred removed 11/12
;;; SENTENCES:  "The stick !!-ed him (on the arm)"
(
:DEF_WORD "!!"
:LCS (cause (* thing 1)
       (go ident (* thing 2)
        (toward ident (thing 2) (at ident (thing 2) (!!-ed 9))))
       ((* [on] 23) loc (*head*) (thing 24)))
:VAR_SPEC ((1 (UC (animate -))) (9 :conflated))
)

;;; [21.1.c]  Cut Verbs - Change of State / -on/with
;;; WORDS: "chip, clip, cut, hack, hew, saw, scrape, scratch, slash, snip"
;;; THETA ROLES: _ag_th,mod-loc(),instr(with) ;; pred removed 11/12
;;; SENTENCES:  "She !!-ed him (on the arm) (with a knife)"
(
:DEF_WORD "!!"
:LCS (cause (* thing 1)
       (go ident (* thing 2)
        (toward ident (thing 2) (at ident (thing 2) (!!-ed 9))))
       ((* [on] 23) loc (*head*) (thing 24))
       ((* with 19) instr (*head*) (thing 20)))
:VAR_SPEC ((1 (UC (animate +))) (9 :conflated))
)

;;; [21.1.d]  Cut Verbs - Locational
;;; WORDS: "chip, clip, cut, hack, hew, snip"
;;; THETA ROLES: _th
;;; SENTENCES:  "The knife !! well"
(
:DEF_WORD "!!"
:LCS (go loc (* thing 2)
        (toward loc (thing 2) (at loc (thing 2) (thing 6)))
        (!!-ingly 26))
:VAR_SPEC ((2 (UC (animate -))) (26 :conflated))
```

```
)

;;; [21.1.e]  Cut Verbs - Locational / -at
;;; WORDS: "chip, clip, cut, hack, hew, saw, scrape, scratch, slash, snip"
;;; THETA ROLES: _ag_goal(at)
;;; SENTENCES:  "I !!-ed at him"
(
:DEF_WORD "!!"
:LCS (cause (* thing 1)
         (go loc (!! 2)
             (toward loc (!! 2) ((* at 10) loc (!! 2) (thing 6)))))
:VAR_SPEC ((1 (UC (animate +))) (2 :conflated))
)
```

A few examples

We give below a few examples of the above recommendations, with the recommended notation given in this chapter. It is necessary to add elements from the other sections to make a comprehensive and useful lexical entry. Similarly, syntactic and morphological features need to be specified. We give below the semantic frame of the verbs give and walk, the preposition on and the adjective good.

Semantic frame for **give**:

```
[ARG :   arg1: [SELECTIONAL-RESTR : human
                THEMATIC-ROLE: agent, source ]
         arg2: [SELECTIONAL-RESTR : not(animate)
                THEMATIC-ROLE: theme ]
         arg3: [SELECTIONAL-RESTR : animate
                THEMATIC-ROLE: destination ]
 PRED-CLASS : transfer of possession
 SEM-REPT : ‘arg1 CAUSE arg2 GO TO arg3’ ]
```

Semantic frame for **walk**:

```
[ARG :   arg1: [SELECTIONAL-RESTR : animate
                THEMATIC-ROLE: agent ]
         arg2: [SELECTIONAL-RESTR : location
                THEMATIC-ROLE: destination ]
 PRED-CLASS : movement
 SEM-REPT : ‘arg1 CAUSE arg1 GO TO arg2 MANNER walkingl’ ]
```

The semantic representations given are a form of lexical semantic templates (LST).

Semantic frame for the preposition **on**:

```
[ARG :   arg1: [SELECTIONAL-RESTR : concrete-entity
                THEMATIC-ROLE: theme ]
         arg2: [SELECTIONAL-RESTR : location
                THEMATIC-ROLE: localization ]
 PRED-CLASS : spatial prep
 SEM-REPT : ‘arg1 ON arg2’ ]
```

Semantic frame for a sense of the adjective **good**, meaning that the object denoted by the argument arg1 performs well the action for which it has been designed:

```
[ARG :   arg1: [SELECTIONAL-RESTR : tool
               THEMATIC-ROLE: theme ]
 PRED-CLASS : evaluative
 SEM-REPT : ‘ABILITY-TO(‘pred from telic of arg1’) = good ’ ]
```

6.9 The EAGLES Guidelines for Lexical Semantic Standards

The basic information unit is a word sense. Obligatory attributes are preceded by a dash (-). The star and plus signs (`*`, +) are used in the usual way to indicate expansion of types (e.g. `subject-domain`), into $0, ..., n$ and $1, ..., n$ tokens (`music, dance`). The vertical bar (`|`) indicates disjunction.

```
word-sense-entry -->
  [ -ORTHOGRAPHY : string
    -WORD-SENSE-ID : word-sense-id
    -BASE-TYPE-INFO : base-type-info*
    SUBJECT-DOMAIN : subject-domain*
    SYNONYMS : word-sense-id*
    NEAR-SYNONYMS : word-sense-id*
    HYPONYMS : hyponym*
    HYPERONYMS : hyperonym*
    ANTONYMS : antonym*
    MERONYMS : meronym*
    HOLONYMS : holonym*
    QUANTIFICATION : quantification
    COLLOCATIONS : collocation*
    SEMANTIC-FRAME : sem-frame
    ACTIONALITY : actionality ].
```

A word sense identifier is an integer which refers to a WordNet synset 3.6.2.

```
word-sense-id --> integer.
```

Base type information provide a specification of the conceptual entities germane to the word sense in question, chosen from the list of base types given in §6.1.3, and the relation(s) that each conceptual entity bears to the word sense (e.g. LX-synonym, LX-near-synonym, see §6.1.3).

```
base-type-info -->
  [ BASE-TYPE :  base-type
    LX-RELATION : lx-relation+ ].
base-type --> (entity | animate | ...).
lx-relation --> (lx-synomym | lx-near-synonym | lx-hyponym
                | lx-hyperonym | lx-holonym | lx-meronym
                | lx-subevent).
```

Subject domain information is encoded in terms of the categories described in §6.4 and the subsumption relations among them.

```
subject-domain --> (sports-games-pastimes | history-heraldry | ...).
sports-games-pastimes --> hunting-fishing.
hunting-fishing --> (hunting | fishing).
...
```

Information about hyponymy, antonymy and meronymy involves specification of the specific type of lexical semantic relation involved (e.g. see §6.1.3). Note: "non-exclusive" is the default value for 'HYP-TYPE'.

```
hyponym : [ HYP-TYPE: (exclusive | conjunctive | non-exclusive)
            HYP-ID : word-sense-id ].

hyperonym : [HYP-TYPE: (exclusive | conjunctive | non-exclusive)
             HYP-ID : word-sense-id ].

antonym: [ ANT-TYPE: (complementary | gradable |
                      pseudo-comparative |true-comparative |
                      antipodal | reversive | relational)
           ANT-ID : word-sense-id ].

meronym : [ MER-TYPE: (member | substance | part)
            HOLS: [HOL-ID: word-sense-id
                   REL-TYPE : (exclus | conjunct | non-exclus)]+
            MER-ID : word-sense-id ].

holonym : [ HOL-TYPE: (member | substance | part)
            MERS: [MER-ID: word-sense-id
                   REL-TYPE : (exclus | conjunct | non-exclus)]+
            HOL-ID : word-sense-id ].
```

Quantification specifies lexical properties of determiners such as quantification strength and direction of entailment.

```
quantification : [ Q-STRENGTH : (weak | strong)
                   Q-ENTAILMENT : (upward | downward) ]
```

Collocation information includes reference to each collocate, expressed in terms of sets of word senses, its location and upper/lower distance as well as the relevant dependency configuration (e.g head/dependendent to head/dependendent),

```
collocation -->
     [ COLLOCATE : word-sense-id+
       DIRECTION : (left | right)
       DISTANCE : [LOWER-LIMIT : integer
                   UPPER-LIMIT : integer]
       DEPENDENCY : (h2d | d2h | d2d | h2h)
       DEP_TYPE : dependency_type
       PROBABILITY : probability-value]
```

Semantic frames include information about the semantic class of a predicate, expressed as either a base-type or a set of word senses, and its arguments.

```
sem-frame : [ -ARG : arg*
              PRED-CLASS : (base-type* | word-sense-id*)
              SEM-REPT : {Some form of semantic representation,
                           with links to arg*}
              QUALIA : {Some encoding of Qualia roles,
                         with links to args*} ]
```

Argument information includes reference to selectional restrictions, collocations and thematic role (see §6.8.3 for a complete list and alternative encodings of thematic role values).

No specific guidelines are given for semantic representation and qualia encoding; see §6.8.3 for suggestions.

```
arg --> [ SELECTIONAL-RESTR : (base-type* | word-sense-id*)
          COLLOCATIONS : collocation*
          THEMATIC-ROLE : th-role ]
```

Both static and dynamic options are give for the encoding of actionality information, as discussed in §6.6 (CUMULATIVE = +SQA, Quantized = −SQA)

```
actionality --> (static-actionality | dynamic-actionality).

static-actionality --> [ (STATE | PROCESS | ACHIEVEMENT | ACCOMPLISHMENT) ].

dynamic-actionality -->
     [ ACT-TYPE : (DYNAMIC | STATIVE)
       THEME-REF : (CUMULATIVE | QUANTIZED) ].
```

Bibliography

[Cyc97] Cyc Ontology Guide: Introduction. http://www.cyc.com/cyc-2-1/intro-public.html. Site visited 27/08/97.

[Gum97] The Generalized Upper Model 2.0. http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html. Site visited 28/08/97.

[EAG96] EAGLES: Preliminary Recommendations on Subcategorisation. http://www.ilc.pi.cnr.it/EAGLES96/synlex/node63.html. Site visited 6/2/98.

[Mik97] Mikrokosmos. http://crl.nmsu.edu/Research/Projects/mikro/. Site visited 28/08/97.

[Sen97] Ontology creation and use: Sensus. http://www.isi.edu/natural-language/resources/sensus.html. Site visited 28/08/97.

[UMLS97] *UMLS Knowledge Sources, NLM, U.S. Dept. of Health and Human Services.* `http://www.nlm.nih.gov/pubs/factsheets/umls.html`. Site visited 6/2/98.

[Agi96] Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of COLING'96.*

[Alo94a] Alonge, A. 'Lexicalisation Patterns in Italian: Analysis of Dictionaries and a Textual Corpus', in Alonge A. and N. Calzolari (eds.), 1994.

[Alo94b] Alonge, A. and N. Calzolari *First Deliverable on Lexicalisation*, Esprit BRA-7315 Acquilex-II, Pisa, 1994.

[Alo94c] Alonge, A. e N. Calzolari (eds.) 1994 First Deliverable on Lexicalisation, Esprit BRA-7315 Acquilex-II, Pisa.

[Alo95] Alonge, A. e N. Calzolari (eds.) 1995 Second Deliverable on Lexicalisation, Esprit BRA-7315 Acquilex-II, Pisa.

[AloFC] Alonge, A., L. Bloksma, N. Calzolari, I. Castellon, T. Marti, W. Peters, P. Vossen, 'The Linguistic Design of the EuroWordNet Database', *Computers and the Humanities*, forthcoming.

[Als87] Alshawi, Hiyan.1987. *Memory and Context for Language Interpretation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, England.

[Als91] Alshawi H., Arnold D.J., Backofen R., Carter D.M., Lindop J.,Netter K., Pulman S.G., Tsujii J., Uszkoreit H. (1991) *EurotraET6/1: Rule Formalism and Virtual Machine Study. Final Report*, Commission of the European Communities.

[Ama95] Amaral, Do & Satomura, Y. (1995) Associating semantic grammars with the SNOMED: processing medical language and representing clinical facts into a language-independent frame, Medinfo 95, pp.18-22.

[Ana88] Ananiadou, S., 1988. A Methodology for Automatic Term Recognition. Ph.D Thesis, University of Manchester Institute of Science and Technology.

[Arn89] Arnold D., *Theoretical and descriptive issues in machine Translation, Phd dissertation*, University of Essex, 1989.

[Ash95] Asher, N. and Lascarides A. (1995) Lexical Disambiguation in a Discourse Context. *Journal of Semantics.*

[Aus62] Austin J. (1962) *How to do Things with Words.* Oxford: Clarendon Press.

[BAD] BADGER

[Bac86] Bach, E., The Algebra of Events, in *Linguistics and Philosophy*, 9, pp. 5-16, 1986.

[Bad96] Badia, T. (1996) Prepositions in Catalan. Forthcoming in Balari S. and Dini L. (eds.) 1996 *HPSG in Romance*, Standford University.

[Bak88] Baker, M.C., *Incorporation: A Theory of Grammatical Function Changing*, Chicago University Press, 1988.

[Bal96] Balari S. and Dini L. (1996) "HPSG in Romance" Standford University.

[Bar64] Y. Bar-Hillel. *Language and Information.* Addison-Wesley, 1964.

[Bar97] Barzilay, R. and M. Elhadad (1997) Using Lexical Chains for Text Summarization. In I. Mani and M. Maybury (eds) *Intelligent Scalable Text Summarization, Proceedings of a Workshop Sponsored by the Association for Computational Linguistics*, Madrid, Spain.

[Bat90] Bateman J.A. Upper modeling: organizing knowledge for natural language processing. In *5th. International Workshop on Natural Language Generation*, Pittsburgh, PA, June 1990.

[Bat94] Bateman J.A., B. Magnini, and F. Rinaldi. The generalized *{Italian, German, English} upper model*,. In *Proceedings of the ECAI94 Workshop: Comparison of Implemented Ontologies*, Amsterdam, 1994.

[Bau92] Baud, R.H. and Rassinoux, A-M. and Scherrer, J-R., Natural Language processing and medical records, *Proc. of Medinfo '92*, 1992, 1362-1367

[Bau] Baudin, Catherine, Smadar Kedar and Barney Pell, Using Induction to Refine Information Retrieval Strategies. In Proceedings of AAAI-94, Seattle, 1994.

[Bau94] Bauer Daniel, Frédérique Segond and Annie Zaenen, *Enriching an SGML-Tagged Bilingual Dictionary for Machine-Aided Comprehension*, Rank Xerox Research Center Technical Report, MLTT, 11, 1994.

[Bau95] Bauer Daniel, Frédérique Segond, Annie Zaenen, *LOCOLEX : the translation rolls off your tongue*, ACH-ALLC95, Santa-Barbara, USA, July 1995.

[Bec90] Beckwith, R. and G.A. Miller (1990) Implementing a Lexical Network In International Journal of Lexicography, Vol 3, No.4 (winter 1990), 302-312. 1990.

[Bel88] Belletti, A., Rizzi, L. (1988). "Phych-Verbs and Q-Theory, *Natural Language and Linguistic Theory*, 6, pp. 291-352.

[Berg91] Bergsten, N. 1991. *A Study on Compound Substantives in English.* Almquist and Wiksell, Uppsala.

[Blo96] Bloksma, L., Díez-Orzas, P., and Vossen, P. (1996). User requirements and functional specification of the eurowordnet project. Technical report, Deliverable D001, EC-funded project LE # 4003, University of Amsterdam, Amsterdam.

[Bog90] Boguraev B., J. Pustejovsky "Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design", in H. Karlgren (ed.), *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, 1990.

[Bog97] Boguraev, B. & C. Kennedy (1997) Salience-based Content Characterization of Text Documents. In I. Mani and M. Maybury (eds) *Intelligent Scalable Text Summarization, Prooceedings of a Workshop Sponsored by the Association for Computational Linguistics*, Madrid, Spain.

[Bou92] Bourigault, D., 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of COLING*, 977–981.

[Bou97] Bouillon P., Polymorphie et sémantique lexicale, Thèse de troisième cycle, Université de Paris VII, 1997.

[Boui92] Bouillon, P, K. Bösefeldt, and Graham Russell. 1992. Compound Nouns in a Unification-Based MT System. In *Proceedings of the Third Conference on Applied Natural Language Processing* (p209-215). Trento, Italy.

[Bra79] Brachman, R. (1979) "On the Epistemological Status of Semantic Networks,"in N. Findler (ed.) Associative Networks: Representation and Use of Knowledge by Computers. New York: Academic Press.

[Bra85a] Brachman R. and H. Levesque (1985) *Readings in Knowlegde Representation*, California: Morgan Kaufmann Publishers Inc.

[Bra85b] Brachman, R. and J. Schmolze (1985). "An Overview of the KL-ONE Knowledge Representation System," in *Cognitive Science*, Volume 9.

[Bre89] Bresnan, J. and Kanerva, J. (1989) Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. LI, Vol. 20, 1-50.

[Bre95a] Breidt Lisa, Frédérique Segond '*IDAREX: formal description of German and French Multi-Word Expressions with Finite-State Technology*, MLTT-022, Novembre 1995.

[Bre95b] Breidt Lisa, Frédérique Segond, Giuseppe Valetto "Formal description of Multi-word Lexemes with the Finite State formalism: IDAREX", *Proceedings of COLING*, August 5-9, Copenhagen, 1995.

[Bre95c] Breidt Lisa, Frédérique Segond, "Compréhension automatique des expressions à mots multiples en fran=E7ais et en allemand", *Quatrièmes journées scientifiques de Lyon, Lexicomatique et Dictionnairiques*, Septembre 1995.

[Bre96] Breidt Lisa, Frédérique Segond, Giuseppe Valetto, "Multiword lexemes and their autom atic recognition in texts", *COMPLEX96*, Budapest, September 1996.

[Bri89] Briscoe, E.J. and B. Boguraev, (eds) (1989) Computational Lexicography for Natural Language Processing. London/New York: Longman.

[Bri92] Brill Eric, 'A simple Rule-Bases Part of Speech Tagger' in *Proceedings of ANLP-1992.*

[Bri93] Briscoe, E J., A. Copestake and V. de Paiva (eds.)(1993) Default Inheritance in Unification Based Approaches to the Lexicon. Cambridge UK: Cambridge University Press.

[Bri95] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–566, December 1995.

[Bro91] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91)*, pages 264–270, Berkley, C.A., 1991.

[Bui97] Paul Buitelaar and Federica Busa. ms. Brandeis University. Waltham, MA, 1997.

[Bui98] P. Buitelaar. CoreLex*: Systematic Polysemy and Underspecification*. PhD thesis, Brandeis University, Department of Computer Science, 1998.

[Bun81] Bunt H.C. (1981) *The formal semantics of mass terms*, Ph.D. Thesis, University of Amsterdam.

[Bus90] Busa, F., *Compositionality and teh Semantics of Nominals*, PhD. Dissertation, Brandeis University, MA, 1996.CEC (1990) Eurotra Reference Manual. Luxemburg

[Bus96a] Busa, F. (1996). *Compositionality and the Semantics of Nominals*, PhD Thesis, Brandeis Univerisity.

[Bus96b] Busa, F., and M. Johnston (1996). "Cross-Linguistic Semantics for Complex Nominals in the Generative Lexicon," in *Proceedings of the AISB96 Workshop: Multilinguality in the Lexicon*, Brighton, UK.

[Cal90] Calzolari N., Bindi R. (1990). 'Acquisition of lexical information from a large textual Italian corpus', in *Proceedings of COLING-90*, Helsinky, Finland.

[Cal93] Calzolari et al. 1993 Calzolari, N., Hagman, J., Marinai, E., Montemagni, S., Spanu, A., Zampolli,A. (1993). "Encoding lexicographic definitions as typed feature structures: advantages and disadvantages", in F. Beckmann and G. Heyer (eds.), *Theorie und Praxis des Lexikons*, Walter de Gruyter, Berlin, pp. 274-315.

[Cal96] Calzolari N., 'Lexicon and Corpus: A Multi-faceted Interaction', (Plenary Speech), in M. Gellerstam, J. Jrborg, S.-G. Malmgren, K. Norn, L. Rogstrm, C. Rjder Papmehl (eds.), *EURALEX '96 Proceedings* I-II, Gotheborg University, Department of Swedish, 1996.

[Car] Carbonnel & Tomita

[Car77] Carlson G. N. "A Unified Analysis of the English Bare Plural", in : *Linguistics and Philosophy*, 1977, pp. 413-455.

[Car84] Carlson, G. (1984) On the Role of Thematic Roles in Linguistic Theory. Linguistics, 22, 259–279.

[Car88] Carter, R. 'On Movement', in B. Levin and C. Tenny. (eds.) *On Linking: Papers by Richard Carter*, Cambridge University Press, MIT, 1988.

[Car90] Carlson L. and S. Nirenburg. World modeling for NLP. Technical Report CMU-CMT-90-121, Carnegie Mellon University, Center for Machine Translation, Pittsburgh, PA, 1990.

[Car94] Carenini, G. and Moore, J.D., Using the UMLS Semantic Network as a basis for constructing a terminological knowledge base: a preliminary report, *Proc. of SCAMC '94*, 1994, 725-729

[Car95] Carlson G. N. and F. J. Pelletier, *The Generic Book*, Chicago and Londres: Chicago University Press, 1995.

[Cat89] R. Catizone, G. Russell, and S. Warwick. Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop (AAAi-89)*, Detriot, Michigan, 1989.

[Cha70] Chafe, W. L., *Meaning and the structure of language*, University Press, Chicago, 1970.

[Cha88] Chaffin R., D. Hermann and M. Winston (1988) An empirical taxonomy of part-whole relations: effects of part-whole relation type on relation identification, In Language and Cognitive Processes, 3. Utrecht: VNU Science Press: 17-48.

[Cha97] Chai, J. and Bierman, A. (1997). The use of lexical semantics in information extraction. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

[Chi90] Chierchia G. et S. McConnell-Ginet *An Introduction to Semantics*, Cambridge:MIT, 1990.

[Cho86] Chomsky, N. (1986) *Knowledge of Language*: *Its Nature, Origin and Use.* New York: Praeger.

[Chu89] Church K.W., Gale W., Hanks P., Hindle D. (1989). 'Parsing, word associations and typical predicate-argument relations, in *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, Pittsburg, PA, pp. 103-112.

[Chu90] Church, K. and P. Hanks, Word Association, Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16:1, PP. 22-29

[Cim89] Cimino, J.J. and Hripcsak, G. and Johnson, S.B. and Clayton, P.D., Designing an introspective, controlled medical vocabulary., *Proc. of 14th Annual Symposium on Computer Applications in Medical Care*, Kingsland, L.W., 1989, IEEE Computer Society Press, 513-518

[Coo79] Cook, W. A., *Case Grammar: development of the Matrix Model (1979-1978)*, Georgetown University Press, 1979.

[Cop91b] Copestake A., Sanfilippo A., Briscoe T., de Paiva V. (1991) The Acquilex LKB: an Introduction, in Briscoe T., Copestake A.,de Paiva V. (Eds.) *Default Inheritance in Unification Based Approaches to the Lexicon*, Esprit BRA Acquilex Project (Action 3030), pp. 182-202.

[Cop91] Copestake, A. and T. Briscoe (1991). "Lexical Operations in a Unification-Based Framework," in J. Pustejovsky and S. Bergler, (eds.), *Lexical Semantic and Knowledge Representation*, Springler-Verlag.

[Cop92] Copestake, A. (1992). *The Representation of Lexical Semantics Information*, CSRP, University of Sussex.

[Cop93] Copestake, A. and A. Sanfilippo (1993) Multilingual lexical representation. In *Working Notes of the AAAI-93 Spring Symposium Series: Building Lexicons for Machine Translation*, Stanford, University, CA.

[Cop95b] A Copestake, T Briscoe, P. Vossen, A Ageno, I Castellon, F Ribas, G Rigau, H Rodriguez, A Sanmiotou, 1995 Acquisition of Lexical Translation Relations from MRDs, In: Journal of Machine Translation, Volume 9, issue 3, 1995.

[Cop95] Copestake, A. (1995). "The Representation of Group Denoting Nouns in a Lexical Knowledge Base," in Saint-Dizier P. and E. Viegas, (eds.) *Computational Lexical Semantics*, Cambridge University Press.

[Cop97] Copestake, Ann., and Alex Lascarides. 1997. Integrating Symbolic and Statistical Representations: The Lexicon Pragmatics Interface. Proceedings of 35th Annual Meeting of the Association for Computational Linguistics. ACL Press, New Jersey.

[Cos67] Coseriu, E. 'Lexikalische Solidaritaten', *Poetica* 1, 1967.

[Cow] Cowie et al.

[Cru73] Cruse, A., Some Thoughts on Agentivity, *Journal of Linguistics*, vol 9-1, 1973.

[Cru86] Cruse, A., *Lexical Semantics*, Cambridge university Press, 1986.

[Cru94] Cruse, A. (1994)

[Cut92] Cutting Doug, Julian Kupiec, Jan Pedersen, and Penelope Sibun, 'A Practical Part-of-speech Tagger', in *Proceedings of ANLP-92*, Trento, Italy, (1992).

[Dag94a] Dagan I. and Itai A., 'Word Sense Disambiguation Using a Second Language Monolingual Corpus', Computational Linguistics, 20, 563-596.

[Dag94b] Dagan, I. and Church, K., 1994. TERMIGHT: Identifying and Translating Technical Terminology. In *Proceedings of EACL*, 34–40.

[Dai94] Daille, B.; Gaussier, E. and Lange, J.M., 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of COLING 94*, 515–521.

[Dam93] Damerau, F. (1993) Generating and evaluating domain oriented multi-word terms from texts, Information Processing & Management, 29(4):433-447.

[Dav90] David, 1990

[Dav67] Davidson, D., The logical Form of Action Sentences, in *The Logic of Decision and Action*, N. Rescher (ed.), University of Pittsburgh Press, 1967.

[DED] DEDAL, at http://cdr.stanford.edu/html/GCDK/dedal/dedal-info.html.

[DEFI] DEFI project: http://engdep1.philo.ulg.ac.be/fontenelle/thframe.htm

[Die96a] Diez Orzas, P. , Louw M. and Forrest, Ph, (1996) High level design of the EuroWordNet Database. EuroWordNet Project LE2-4003, Deliverable D007. 1996

[Die96b] Diez-Orzas, P. L. (1996) WordNet1.5 Contents and Statistics. Novell LTD Internal Report, March 1996, Antwerp

[Dik78] Dik, S. *Stepwise Lexical Decomposition*, Lisse, Peter de Ridder Press, 1978.

[Dik80] Dik, S. *Studies in Functional Grammar*, New York, Academic Press, 1980.

[Dik89] Dik, S.C. (1989). *The Theory of Functional Grammar, Part I: The Structure of the Clause*, Foris Publications, Dordrecht.

[Dix91] Dixon R. M. W., A new approach to English Grammar, on Semantic Principles . Oxford:Clarendon Press, 1991.

[Dor90] Dorr, B. Solving Thematic Divergences in Machine Translation. In *Proceedings of the 28th Conference of the Association for Computational Linguistics*, 1990.

[Dor93] Dorr, B., *Machine Translation, a view from the lexicon*, MIT Press, 1993.

[Dor95a] Bonnie Dorr & J. Klavans (1995) (eds) Building Lexicons for Machine Translation II, Special Issue of Machine Translation, vol.10, nos 1-2.

[Dor95b] Dorr, B., Garman, J., Weinberg, A. (1995) From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT, in Machine Translation 9, pp.221-250.

[Dow72] Dowty, D. (1972) *Studies in the Logic of Verb Aspect and Time Reference.* PhD Thesis, University of Texas.

[Dow77] Downing, P. 1977. On the Creation and Use of English Compound Nouns. Language 53. 810-842.

[Dow79] Dowty, D. R. (1979). *Word Meaning and Montague Grammar*, Kluver Academic Publishers.

[Dow89] Dowty, D., On the Semantic Content of the Notion of Thematic Role, in G. Cherchia, B. Partee, R. Turner (eds), *Properties, Types and meaning*, Kluwer, 1989.

[Dow91] Dowty, D., Thematic Proto-roles and Argument Selection, Language, vol. 67-3, 1991.

[Dun93] Dunning (1993)

[Eng94] Enguehard, C. & Pantera (1994) automatic manual acquisisition of terminology In Journal of Quantitative Linguistics, 2(1) 27-32

[Fab96] Fabre, C, Recovering a Predicate-Argument structure for the automatic interpretation of english and French nominal compounds, *Workshop on Predicative Forms in Natural Language*, Toulouse, August 1996.

[Fan61] Fano, R. (1961) Transmission of Information: A statistical theory of communications, MIT press, MA.

[Fed96] Federici S., Montemagni S., Pirrelli V. (1996). 'Example-based word sense disambiguation: a paradigm-driven approach, in *Proceedings of the Seventh Euralex International Congress.*

[Fed97] Federici S., Montemagni S., Pirrelli V. (1997). 'Inferring semantic similarity from Distributional Evidence: an Analogy-based Approach to Word Sense Disambiguation, in *Proceedings of the ACL/EACL Workshop on 'Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.*

[Fel90] Fellbaum, C. (1990) English Verbs as a Semantic Net In International Journal of Lexicography, Vol 3, No.4 (winter 1990), 278-301. 1990.

[Fel93] Fellbaum, C., *English Verbs as Semantic Net*, Journal of Lexicography, vol. 6, Oxford University Press, 1993.

[Fel95] Fellbaum, C., Co-occurrence and Antonymy, *Journal of Lexicography*, vol. 8-2, Oxford University Press, 1995.

[Fel97] Fellbaum, C. (in press). A Semantic Network of English Verbs, in C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1997.

[Fil68] Fillmore, C., The Case for Case, in *Universals in Linguistic Theory*, E. Bach and R.T. Hams (eds.), Holt, Rinehart and Winston, New York, 1968.

[Fil77] Fillmore, C., The Case for Case Reopened. In P. Cole and J. Sadock *Syntax and Semantics 8: Grammatical Relations*, Academic Press, New York, pp. 59-82.

[Fil92] Fillmore C. and B. T. S. Atkins, 'Towards a Frame-Based Lexicon: the Case of RISK', in A. Lehrer and E. Kittay (eds.), *Frames and Fields*, Erlbaum Publ., NJ, 1992.

[Fil94] Fillmore C. and B. T. S. Atkins, 'Starting where the Dictionaries Stop: the Challenge of Corpus Lexicography' in B. T. S. Atkins and A. Zampolli (eds.) *Computational Approaches to the Lexicon*, Oxford University Press, Oxford UK, 1994.

[Fin80] Finin, Timothy. W. 1980. *The Semantic Interpretation of Compound Nominals.* Doctoral Dissertation. University of Illinois at Urbana-Champaign.

[Fol84] Foley, W. and Van Valin, R., *Functional syntax and Universal grammar*, CUP, 1984.

[Fon97] Fontenelle Thierry, "Turning a Bilingual Dictionary into a Lexical-Semantic Database" Lexicographica Series Maior, Vol. 79, Max Niemeyer Verlag, Tübingen, October 1997.

[For82] Ford M., Bresnan J., Kaplan R.M. (1982). 'A Competence-based Theory of Syntactic Closure, in J. Bresnan, (ed.), *The mental representation of grammatical relations*, MIT Press, Cambridge Massachusetts, pp. 727-796.

[Fra96a] Frantzi K. and Ananiadou S., 1996. A Hybrid Approach to Term Recognition. In *Proceedings of NLP+IA*, 93–98.

[Fra96b] Frantzi K. and Ananiadou S., 1996. Extracting Nested Collocations. In *Proceedings of COLING*, 41–46.

[Fur] Furuse, O. & Iida, H. (1992) An example based method for transfer driven MT, in Proceedings of the 4th international conference on theoretical and methodological issues in MT, Montreal.

[Fuj97] Fujii, A., Hasegawa, T., Tokunaga, T., and Tanaka, H. (1997). Integration of hand-crafted and statistical resources in measuring word similarity. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

[Gal91] Gale & Church (1991) Concordances for parallel texts In Proc. of the 7th annual conf of the UW centre for the new OED and text research using corpora, pp.40-62.

[Gal92a] Gale, Church and Yarowsky, 'A Method for Disambiguating Word Senses in a Corpus', *Computers and the Humanities*, 26, pp. 415-439, 1992.

[Gal92b] Gale, Church and Yarowsky, (1992) 'Using Bilingual Materials to Develop Word Sense Disambiguation Methods', in *Proceedings of TMI-92*, pp. 101-112.

[Gal92c] W. Gale, K. Church, and D. Yarowsky. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 233–237, Harriman, NY, February 1992.

[Gil97] Gilarranz, J., Gonzalo, J., and Verdejo, M. (1997). An approach to cross-language text retrieval with the eurowordnet semantic database. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.* AAAI Spring Symposium Series, to appear. [Gilarranz et al., fc.] Gonzalo, J., F. Verdejo, C. Peters, N. Calzolari., Applying EuroWordNet to Multilingual Text Retrieval, in Computer and the Humanities, Special Edition on EuroWordNet, fc.

[Gio97] Giorgi, A. & F. Pianesi, *Tense and Aspect: From Semantics to Morphosyntax*, OUP, Oxford, 1997.

[Gol94] Goldberg, A., *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, 1994.

[Goo56] Goodenough, W. H. 'Componential Analysis and the Study of Meaning', *Language*, 32, 1956.

[Gre66] Greimas, A. J. *Sémantique structurale*, Paris, 1966.

[Gre94] Grefenstette G., 1994. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers.

[Gre96] Grefenstette Gregory, Schulze Maximilian, Heid Uli, Fontenelle Thierry and Gérardy Claire, "The DECIDE project: Multilingual collocation extraction", *Proceedings of the European Association for Lexicography* (EURALEX), Gothenburg, 1996.

[Gri90] Grimshaw, J., *Argument Structure*, Linguistic Inquiry monograph no. 18, MIT Press, 1990.

[Gro81] Gross, M. , "Les bases empiriques de la notion de predicat semantique", Langages 63, Larousse, Paris, 1981.

[Gro90] Gross, D. and K.J. Miller (1990) Adjectives in Wordnet In International Journal of Lexicography, Vol 3, No.4 (winter 1990), 265-277. 1990

[Gru67] Gruber, J., *Studies in Lexical Relations*, MIT doctoral dissertation and in *Lexical Structures in Syntrax and Semantics*, North Holland (1976), 1967.

[Gut91] J. Guthrie, L. Guthrie, Y. Wilks and H. Aidinejad, Subject-Dependent Co-Occurrence and Word Sense Disambiguation, ACL-91, pp. 146-152.

[Har97] A. Harley and D. Glennon. Sense tagging in action: Combining different tests with additive weights. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*, pages 74–78. Association for Computational Linguistics, Washington, D.C., 1997.

[Hat] Hatzivassiloglou, V. & McKeown (1993) Towards the automatic identification of adjectival scales: clustering adjectives according to meaning, in Proceedings of 31st ACL, pp. 172-182.

[Hei94] Heid U. and K. Krueger, *On the DELIS Corpus Evidence Encoding Schema (CEES)*, Delis Deliverable D-III-0, 1994.

[Hei95] Heid U., A. Alonge, S. Atkins, G. Bs, N. Calzolari, O. Corrazzari, C. Fillmore, K. Krueger, S. Schwenger, M. Vliegen, *A Lexicographic and Formal Description of the Lexical Classes of Perception and Speech Act Verbs*, Delis Deliverable D-III-1, 1995.

[Hen93] Henschel R., Merging the english and the german upper model. *Arbeitspapiere der GMD* 848, GMD/Institut für Integrierte Publikations-und Informationssysteme, Darmstadt, Germany, 1993.

[Hig85] Higginbotham, J., On Semantics, in *Linguistic Inquiry*, 16, pp. 547-593, 1985.

[Hig97] Higginbotham, J., The Role of Events in Linguistic Semantics, ms., Oxford University, 1997.

[Hin93] Hindle D., Rooth M. (1993). 'Structural Ambiguity and Lexical Relations, *Computational Linguistics*, vol. 19, n.1, March 1993, pp. 103-120.

[Hir97] Hirst, G. and D. St-Onge (1997) Lexical Chains as Representation of context for the detection and correction of malapropism. In C. Fellbaum (ed) *WordNet: An electronic lexical database and some of its applications*. MIT Press, Cambridge, Mass.

[Hje61] Hjelmslev L. *Prolegomena to a Theory of Language*, Bloomington, Indiana, 1961.

[Hob91] Hobbs, J. (1991). "SRI International's TACITUS System: MUC-3 Test Results and Analysis," in *Proceedings of the Third MUC Conference.*

[Hob93] Hobbs, Jerry R., Martin. E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as Abduction. In Fernando C.N. Pereira and Barbara Grosz (eds.) *Natural Language Processing.* MIT Press, Cambridge, Massachusetts.

[Hoe91] Hoey M. (1991) Patterns of Lexis in Text. OUP, Oford, UK. Σ

[Hov90] Eduard H. Hovy. Approaches to the planning of coherent text. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics.* Kluwer Academic Publishers, July 1991. Presented at the Fourth International Workshop on Natural Language Generation. Santa Catalina Island, California, July, 1988.

[HovFC] E. Hovy, fc. What would it Mean to Measure an Ontology? ISI, University of Southern California.

[Hut92] J. Hutchins and H. Sommers. *Introduction to Machine Translation.* Academic Press, 1992.

[Hwe96] Hwee Tou Ng and Hian Beng Lee 'Integrating Multiple Knowledge sources to disambiguate word sense', *ACL* 1996.

[Isa84] Isabelle, P. 1984. Another Look at Nominal Compounds. In *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Meeting of the ACL.* (pp. 509-516).

[Jac72] Jackendoff, R., *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge, 1972.

[Jac76] Jackendoff, R.S., Toward an Explanatory Semantic Representation, *Linguistic Inquiry 7*, 1976.

[Jac83] Jackendoff, R. S. (1983) Semantics and Cognition, Cambridge, Mass., The MIT Press.

[Jac87] Jackendoff, R., The Status of Thematic Relations in Linguistic Theory, Linguistic Inquiry, vol. 18, 1987.

[Jac90] Jackendoff, R. S. (1990) Semantic Structures, Cambridge, Mass., The MIT Press.

[Jac92] Jackendoff R.(1992) Parts and boundaries In B. Levin and S. Pinker (eds.) Lexical & conceptual semantics. Cambridge MA: Blackwell: 9 - 45.

[Jac96] Jackendoff, R. *The Architecture of the Language Faculty*, MIT Press, 1996.

[Jes42] Jespersen, Otto. 1942. *A Modern English Grammar on Historical Principles, IV.* Munksgaard, Copenhagen.

[Jin] Jing, H, et al. (1997) Investigating Complementary Methods for verb sense pruning, in proce. of ACL SIGLEX, workshop on tagging text with lexical semantics, April 1997 Washington.

[Job95] Jones, Bernard. 1995. Nominal Compounds and Lexical Rules. *Working Notes of the Acquilex Workshop on Lexical Rules.* Cambridge, England, August 1995.

[Joh95] Johnston, Michael, Branimir Boguraev, and James Pustejovsky. 1995. The Acquisition and Interpretation of Complex Nominals. *Working Notes of AAAI Spring Symposium on the Representation and Acquisition of Lexical Knowledge*, Stanford University, Palo Alto, California.

[Joh98] Johnston, M. and F. Busa. "Qualia Structure and the Compositional Interpretation of Compounds", in Proceedings of SIGLEX96, and in E. Viegas, (ed.), *Depth and Breadth of Semantic Lexicons*, Kluver, 1998.

[Jon] Jones, D. (1996) Analogical NLP, UCL Press, London.

[Jos87] A. Joshi. The relevance of tree adjoining grammar to generation. In Gerard Kempen, editor, *Natural Language Generation: Recent Advances in Artificial Intelligence, Psychology, and Linguistics.* Kluwer Academic Publishers, Boston/Dordrecht, 1987. Paper presented at the Third International Workshop on Natural Language Generation, August 1986, Nijmegen, The Netherlands.

[Jou94] Joubert, M. and Fieschi, M. and Robert, J-J., A conceptual model for retrieval with UMLS, *Proc. of JAMIA '94*, 1994

[Jus95] Justeson, J.S. and Katz, S.M., 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. In *Natural Language Engineering*, 1:9–27.

[Kam93] Kamp, H. & U. Reyle, *From Discourse to Logic*, Kluwer, Dordrecht, 1993.

[Kap82] Kaplan, R., Bresnan, J., Lexical Functional Grammar: A Formal System for Grammatical Representation, in *The Mantal Representation of Grammatical Relations*, J. Bresnan (ed.), MIT Press, Cambridge, 1982.

[Kar92] Karttunen Lauri, Beesley Kenneth, Two level rules compiler *Technical Report ISTL-92-2*, Xerox Palo Alto Research Center, 1992.

[Kar93] Karttunen Lauri, Yampol Todd, Interactive Finite-State Calculus, *Technical Report ISTL-NLTT-1993-04-01*, Xerox Palo Alto Research Center, 1993.

[Kat63] Katz , J.J. and J. A. Fodor 'The Structure of a Semantic Theory', *Language*, 39, 1963.

[Ken63] Kenny, A., *Actions, Emotion, and Will*, Humanities Press, 1963.

[Kip70] Kiparsky, P., C. Kiparsky 1970 'Fact', in Bierwisch, M. and K. E. Heidolph (eds.), Progress in Linguistics, The Hague, Mouton.

[Kit86] Kittredge, R., A. Polguère and E. Goldberg 25-29 August, 1986. "Synthesizing Weather Forecasts from Formatted Data." Proceedings of COLING-86, The 11th International Conference on Computational Linguistics, University of Bonn, West Germany, 1986. 563-565.

[Kni94] Knight K. and S. Luk. Building a large knowledge base for machine translation. In *Proceedings of the American Association of Artificial Intelligence Conference AAAI-94*, Seattle, WA, 1994.

[Kri87] Krifka, M. (1987) Nominal Reference and Temporal Constitution: Towards a Semantics of Quantity. FNS Bericht 17, Forshungstellefür natürlich-sprachliche Systeme, Universität Tübingen.

[Kri89] Krifka, M., Nominal Reference, Temporal Constitution and Quantification in Event Domains, in R. Bartsch, J. van Benthem and P. van Emde Boas (eds.) *Semantics and Contextual Expressions*, Forsi, Dordrecht, 1989.

[Kri90] Krifka, M. (1990) Thematic Relations as Links between Nominal Reference and Temporal Constitution. In Sag, I. & Sabolcsi, A. (eds.) Lexical Matters, Chigaco University Press.

[Kro92] Krovetz, R. and Croft, W. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

[Kro97] R. Krovetz. Homonomy and polysemy in information retrieval. In *35th Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics (ACL/EACL-97)*, pages 72–78, Madrid, Spain, 1997.

[Kur94] Kurohashi, S. and Nagao, M. (1994). A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. *IEICE TRANSACTIONS on Information and Systems*, E77-D(2):227–239.

[Lah89] Lahav R., Against compositionality: the case of adjectives, in: *Philosophical studies*, 57, 1989.

[Lak70] Lakoff, G. /it Irregularity in Syntax, New York, Holt, Rinehart, & Winston, 1970.

[Lak80] Lakoff, G., Johnson, M., *Metaphors we Live By*, Chicago University Press, 1980.

[Lan87] Landman F. (1987) Groups, plurals, individuals and intentionality, in J. Groenendijk and M. Stokhof (eds.) *Proceedings of the 6th Amsterdam Colloquium, University of Amsterdam* Amsterdam, pp. 197-217.

[Lan89] Landman F.(1989a) "Groups," in *Linguistics and Philosophy*, 12. Dordrecht: Reidel.

[Lan92] Landman, F., The Progressive, in *Natural Language Semantics*, 1, pp. 1-32, 1992.

[Lau] Lauriston, A. (1994) Automatic recognition of complex terms: problems and the TERMINO solution, Terminology 1(1):147-170.

[Lee70] Lees, Robert. 1970. Problems in the Grammatical Analysis of English Nominal Compounds. In Bierwisch and Heidolph (eds.) *Progress in Linguistics*. Mouton, The Hague.

[Lee93] Lee J.H., Kim M.H., Lee Y.J. (1993). 'Information Retrieval based on conceptual distance in IS-A hierarchies, *Journal of Documentation*, 49(2), June 1993, pp. 188-207.

[Leh81] Lehnert W.G., "Plot Units and Narrative Summarization" in Cognitive Science, 4, pp 293-331. 1981.

[Lev78] Levi, J.N. *The syntax and semantics of complex nominals*, New York:Academic Press, 1978.

[Lev93] Levin, B. (1993) English verb classes and alternations: a preliminary investigation, Chicago, The University of Chicago Press.

[Lev95] Levin, B., Rappaport Hovav, M., *Unaccusativity: At the Syntax-Lexical Semantics Interface*, Linguistic Inquiry monograph no. 26, MIT Press, 1995.

[Levi78] Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

[Lew75] Lewis, D., Adverbs of Quantification, in E.L. Keenan (ed.), *Formal Semantics of Natural Language*, CUP, Cambridge, 1975.

[LiA95] [Li and Abe, 1995]Li-95 Li, H. and Abe, N. (1995). Generalizing case frames using a thesaurus and the mdl principle. In *Proceedings of Recent Advances in Natural Language Processing*, pages 239–248.

[Lin83] Link, G., The Logical Analysis of Plurals and Mass Terms, in R. Baeuerle, C. Schwarze and A. von Stechow (eds.), *Meaning, Use and Interpretation of Language*, Walter de Gruyter, Berlin, 1983.

[Lin87] Link, G., Algebraic Semantics for Event Structures, in J. Groenendijk, M. Stokhof and F. Veltman (eds.), *Proceedings of the Sixth Amsterdam Colloquium*, ITLI, University of Amsterdam, 1987.

[Lin88] Link G. (1988) "Generalized quantifiers and plurals," in Garderfors (ed.), *Generalized Quantifiers*. Dordrecht: Reidel.

[Lin97] Lindstromberg, S., Everything you wanted to know about prepositions, John Benjamins, 1997.

[Lon76] Longacre, R. E., *An anatomy of speech notions*, Peter de Ridder Press, 1976.

[Lon95] Lonsdale, D., Mitamura, T., & Nyberg E. (1995) Acquisition of Large Lexicons for Practical Knoweledge based MT, in Machine Translation 9, pp.251-283.

[Lou56] Lounsbury, F. G. 'A Semantic Analysis of the Pawnee Kinship Usage', *Language*, 32, 1956.

[Luk95] A. Luk. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Meetings of the Association for Computational Linguistics (ACL-95)*, pages 181–188, Cambridge, M.A., 1995.

[Lyo77] Lyons J. (1977) Semantics Cambridge University Press, Cambridge. 1977

[Mah95a] Mahesh K. and S. Nirenburg. A situated ontology for practical nlp. In *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada, August 1995.

[Mah95b] Mahesh K. and S. Nirenburg. Semantic classification for practical natural language processing. In *Proc. Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting*, Chicago, IL, October 1995.

[Mak95] Maks, I. and W. Martin (1995) Multitale: linking medical concepts by means of frames. In Proceedings of the 16th international conference on computational linguistics, COLING-96: 746-751, Copenhagen, Denmark.

[Man83] W. Mann and C. Matthiessen. Nigel: A systemic grammar for text generation. Technical Report ISI/RR-83-105, Information Sciences Institute, February 1983. 4676 Admiralty Way, Marina del Rey, California 90292-6695.

[Mar69] Marchand, Hans. 1969. *The Categories and Types of Present Day English Word Formation.* C.H Becksche, Munich.

[Mar80] Marcus M. (1980). *A Theory of Syntactic Recognition for Natural Language*, MIT Press, Cambridge Massachusetts.

[Mar83] Marx W., "The meaning-confining function of the adjective", in: Rickheit et Bock (éds), 1983.

[Mar85] Marantz, A. P., *On the Nature of Grammatical Relations*, MIT Press, 1985.

[Mar86] Martin W. and J. Tops (1986) Groot woordenboek Engels-Nederlands. Van Dale Lexicografie. Utrecht.

[Mau93] Maurel Denis, Passage d'un automate avec tables d'acceptablilité à un automate lexical, *A ctes du colloque Informatique et langue naturelle*, Nantes, pp 269-279, 1993.

[May93] Maybury M. (1993) Automated Event Summarization Techniques. In B. Endres-Niggemeyer, J. Hobbs, and K. Sparck Jones editions, Workshop on Summarising Text for Intelligent Communication - Dagstuhl Seminar Report (9350). Dagstuhl, Germany. 1993.

[McA81] McArthur, T. (1981) *Longman Lexicon of Contemporary English.* Longman, London.

[McC68] McCawley, J. D. 'Lexical Insertion in a Transformational Grammar without Deep Structure' in Darden, B., Bailey and Davison (eds.) *Papers from t he Fourth Regional Meeting of the CLS*, Chicago, The University of Chicago Press, 1968.

[McC97] McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

[McD82] McDonald, David B. 1982. *Understanding Noun Compounds.* CMU Technical Report CS-82-102.

[McD83] D. McDonald. Natural Language Generation as a Computational Problem: an Introduction. In M. Brady and R. Berwick, editors, *Computational Models of Discourse*, Cambridge, MA, 1983. MIT Press.

[McD91] McDonald, D. (1991). "On the Place of Words in the Generation Process," in C. Paris, W.R. Swartout, and W.C. Mann, (eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, Kluver Academic Publishers.

[McK85] K. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press, Cambridge, England, 1985.

[McK90] K. McKeown, M. Elhadad, Y. Fukumuto, J. Lim, C. Lombardi, J. Robin, and F. Smadja. Natural Language Generation in COMET. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation.* Academic Press, 1990.

[McK95] McKeown Kathleen and Dragomir R. Radev (1995) Generating Summaries of Multiple News Articles. SIGIR '95, pages 74-82, Seattle, Washington.

[McR92] S. McRoy. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30, 1992.

[Mel89] Melcuk, I. A. 'Semantic primitives from the Viewpoint of the Meaning-Text Linguistic Theory', *Quaderni di Semantica*, 10, 1, Bologna, Il Mulino, 1989.

[Mes91] Mesli, N. (1991). Analyse et traduction automatique de constructions a verbe support dans le formalaisme CAT2. Eurotra-D Working Papers. No19b, IAI, Sarbrücken.

[Met92] Meteer, M. (1992). *Expressibility and the Problem of Efficient Text Planning*, Pinter Publishers, London.

[Mil76] Miller, G., and P. Johnson-Laird (1976). *Language and Perception*, Harvard University Press.

[Mil85a] Miller, G. (1985). "Noun in Wordnet: A Lexical Inheritance System," in *Internation Journal of Lexicography*, vol.3.

[Mil85b] Miller, G. (1985). "WORDNET: a dictionary browser," in *Proceedings of the First International Conference on Information in Data*, University of Waterloo Centre for the New OED, Waterloo, Ontario.

[Mil90a] Miller George (Ed.) "WordNet: An On-line Lexical Database' *International Journal of Lexicography*, 1990.

[Mil90b] Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. (1990) Five Papers on WordNet. CSL Report 43. Cognitive Science Laboratory. Princeton University.

[Mil94] Miller, G., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. (1994). Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*.

[Mok97a] Mokhtar Salah Ait, JP Chanod *Incremental finite-state parsing* In Proceedings of Applied Natural Language Processing 1997, Washington, DC. April 97.

[Mok97b] Mokhtar Salah Ait, JP Chanod *Subject and Object Dependency Extraction Using Finite-State Transducers* ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. 1997, Madrid.

[Mon94] Monachini M., A. Roventini, A. Alonge, N. Calzolari, O. Corazzari, *Linguistic Analysis of Italian Perception and Speech Act Verbs*, Delis Working Paper, Pisa, 1994.

[Mon95] Montemagni, S. (1995). *Subject and Object in Italian Sentence Processing*, PhD Dissertation, Umist Manchester, UK.

[Moo90] J. Moore and W. Swartout. A reactive approach to explanation: Taking the user's feedback into account. In C. Paris, W. Swartout, and W. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics*. Kluwer Academic Publishers, July 1991. Presented at the Fourth International Workshop on Natural Language Generation. Santa Catalina Island, California, July, 1988.

[Mor91] Morris & Hirst (1991) Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17:21-48.

[Nag92] Nagao M. (1992). 'Some Rationales and Methodologies for Example-Based Approach, in *Proceedings of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing*, 30-31 July 1992, Manchester, UK, pp. 82-94.

[Nat95] Nationalencyklopedins ordbok (1995-96). Nationalencyklopedins ordbok. 1995-6. Språkdata, Göteborg, och Bokförlaget Bra Böcker AB, Höganäs.

[Nel95] Nelson, S.J. et al. (1995) Identifying concepts in medical knowledge, pp.33-36, Medinfo 1995.

[Ng-96] Ng, H. and Lee, H. (1996). Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of ACL'96*.

[Nik95] N. Nikolov, C. Mellish, and G. Ritchie. Sentence Generation from Conceptual Graphs. In *Proceedings of 3rd Int. Conf. on Conceptual Structures (ICCS'95)*, number 954 in LNAI, Santa Cruz, 1995. Springer-Verlag.

[Nir93] Nirenburg, S., C. Defrise, Lexical and Conceptual Structure for Knowledge Based Translation. In J. Pustejovsky (ed.) *Semantics and the Lexicon*. Kluwer, Dordrecht. 1993.

[Nir93b] Nirenburg S., Domashnev C., Grannes D.I. (1993). 'Two Approaches to Matching in Example-Based Machine Translation, in *Proceedings of TMI-93*, pp. 47-57.

[Noy97] Noy N.F.and C.D. Hafner. The state of the art in ontology design. *AI Magazine*, 18(3):53–74, 1997.

[Oxf94] Oxford University Press-Hachette, "The Oxford Hachette French Dictionary", 1994.

[Pai90] Paice C. D. (1990) Constructing Literature Abstracts by Computer: Techniques and Prospects. In Information Processing & Management, 26(1), pages 171–186.

[Pal94] Palmer, F.R. (1994). *Grammatical Roles and Relations*, Cambridge University Press, Cambridge.

[Par90] Parsons T. *Event in the semantics of English: A study in subatomic semantics*, Cambridge:MIT Press, 1990

[Par91] C. Paris. Generation and explanation: Building an explanation facility for the explainable expert systems framework. In C. Paris, W. Swartout, and W. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics.* Kluwer Academic Publishers, July 1991. Presented at the Fourth International Workshop on Natural Language Generation. Santa Catalina Island, California, July, 1988.

[Ped97] T. Pedersen and R. Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, August 1997.

[Pel86] Pelletier F.J. - L.K. Schubert (1986) "Mass expressions," in D. Gabbay and F. Guenthner (eds.) *Handbook of philosophical logic*, 4. Dordrecht: Reidel.

[Per91] Perron, 1991

[Per92] Pereira F., Tishby N. (1992). 'Distributional Similarity, Phase Transitions and Hierarchical Clustering, Working Notes, Fall Symposium Series, AAAI, pp. 108-112.

[Pes82] Pesetsky, D., *Paths and Categories*, MIT doctoral dissertation, 1982.

[Pia96a] Pianesi, F. & A. Varzi, Events, Topology and Temporal Relations, in *The Monist*, 79, pp. 89-116, 1996.

[Pia96b] Pianesi, F. & A. Varzi, Refining Temporal Reference in Event Structures, in *Notre Dame Journal of Formal Logic*, 37, pp. 71-83, 1996.

[Pic78] Picabia L. *Les constructions adjectivales en français. Systématique transformationnelle*, Genève, Paris: Droz, 1978.

[Pin89] Pinker, S., *Learnability and Cognition: The acquisition of argument structure*, MIT Press, 1989.

[Pol87] Pollard C. and Sag I. (1987) "Information-Based Syntax and Semantics, Vol 1: Fundamentals" CSLI Lecture Notes no. 13. Stanford.

[Pot74] Pottier, B. *Linguistique générale, Théorie et Description*, Paris, Klinsksieck, 1974.

[Poz96] Poznanski V. & A. Sanfilippo (1996) Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora. In B. Boguraev and J. Pustejovsky (eds) *Corpus Processing for lexical Acquisition*, MIT Press.

[Pri88] Pritchett, B. (1988) "Garden path phenomena and the grammatical basis of language processing." *Language*, **64**, 539-576.

[Pro78] Procter, P. (Eds) (1978) Longman Dictionary of Contemporary English. Longman, Harlow and London.

[Pus88] Pustejovsky, J., The Geometry of Events, in: *Studies in Generative Approaches to Aspect*, C. Tenny (ed.), MIT Press, 1988.

[Pus91a] Pustejovsky, J. (1991) The Generative Lexicon. *Computational Linguistics*, 17(4).

[Pus91b] Pustejovsky, J. (1991) The syntax of event structure. Cognition, 41, 47-81. 1991.

[Pus94a] Pustejovsky, J. (1994) Linguistic Constraints on Type Coercion. In P. St.Ditzier and E. Viegas (eds) *Computational Lexical Semantics*, CUP, in press.

[Pus94b] James Pustejovsky. Semantic Typing and Degrees of Polymorphism. In Martin-Vide (ed.), *Current Issues in Mathematical linguistics.* Elsevier, 1994.

[Pus95a] James Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, MA, 1995.

[Pus95b] James Pustejovsky, Bran Boguraev and Michael Johnston. A Core Lexical Engine: The Contextual Determination of Word Sense. Technical Report, Department of Computer Science, Brandeis University, 1995.

[Pus96] James Pustejovsky. The Semantics of Complex Types. *Lingua*, 1996.

[Qui94] Quirk R., S. Greenbaum, G. Leech, J. Svartvik *A Comprehensive Grammar of the English Language*, London et New York:Longman, 1994, (Twelfth impression).

[Rad89] Rada R., Hafedh M., Bicknell E. and Blettner M. (1989) Development and application of a metric on semantic nets. *IEEE Transactions on System, Man, and Cybernetics*, 19(1):17-30.

[Rad96] Radford et al. (1996)

[Rap88] Rappaport, M., Levin, B., What to do with θ-roles ?, in *Syntax and Semantics 21: Thematic Relations*, W. Wilkins (ed.), Academic Press, 1988.

[Ras95] Raskin V. et S. Nirenburg *Lexical Semantics of Adjectives. A microtheory of Adjectival meaning*, in: MCCS-95-288, 1995.

[Rau94] Rau L. F., R. Brandow, and K. Mitze (1994) Domain-Independent Summarization of News. In Dagstuhl Seminar Report 79: Summarising Text for Intelligent Communication, B. Endres-Niggemeyer, J. Hobbs, and K. Sparck-Jones, editors, Dagstuhl, Germany.

[Rav90] Ravin, Y., *Lexical Semantics without Thematic Roles*, Oxford Univ. Press, 1990.

[Rei91] N. Reithinger. POPEL—a Parallel and Incremental Natural Language Generation System. In C. Paris, W. Swartout, and W. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics.* Kluwer, 1991.

[Res95] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI.*

[Res97] Resnik, P. (1997) Selectional Preference and Sense Disambiguation, In proce. of ACL SIGLEX, workshop on tagging text with lexical semantics, April 1997 Washington.

[Reu93] Reuland, E., Abraham, W., (eds) *Knowledge and Language, Vol II*, Kluwer Academic, 1993.

[Rib94] Ribas Framis F. (1994). 'An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus', in *Proceedings of COLING-94*, Kyoto, Japan, pp. 769-774.

[Rib95] [Ribas, 1995]Ribas-95 Ribas, F. (1995). On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–118.

[Ric83] Rickheit G. et M. Bock, *Psycholinguistic Studies in Language Processing*, Walter de Gruyter:Berlin, 1983.

[Ric95] Richardson, R. and Smeaton, A. (1995). Using wordnet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe.*

[Rij] C.J. van Rijsbergen, *Towards an Information Logic*, at http://www.dcs.gla.ac.uk/Keith/

[Ril93] Riloff E., "A Corpus-Based Approach to Domain-Specific Text Summarisation: A Proposal", in B. Endres-Niggemeyer, J. Hobbs, and K. Sparck Jones editions, Workshop on Summarising Text for Intelligent Communication - Dagstuhl Seminar Report (9350). Dagstuhl, Germany. 1993.

[Riv87] R. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[Rob93] Robert P., "Le Nouveau Petit Robert: Dictionnaire de la langue Francaise", 1993.

[Roc92] Roca, I.M. (ed.), *Thematic Structure: its Role in Grammar*, Mouton de Gruyter, Berlin, 1992.

[Roc94] Rocha, R.A. and Rocha, B. and Huff, S.M., Automated Translation between Medical Vocabularies using a frame-based Interlingua, *Proc. of SCAMC '94*, 1994, 690-694

[Roz89] Rozwadowska B. (1988) Thematic Restrictions on Derived Nominals. In Wilkins, W. (ed.) *Syntax and Semantics, Volume 21*, Academic Press, New York.

[Sad93] Sadler L. and D. Arnold, *Prenominal Adjectives and the Phrasal/Lexical Distinction*, mss., 1993.

[Sag90] Sager, J.C., 1990. A Practical Course in Terminology Processing. John Benjamins Publishing Company.

[Sai98] Saint-Dizier, P., Verb Semantic Classes in French, to be published in *Predicative forms in language and lexical knowledge bases*, P. Saint-Dizier (ed.), Kluwer Academic, 1998.

[Sal97] Salton G., A. Singhal, M. Mitra, and C. Buckley (1997) Automatic Text Structuring and Summarization, Information Processing and Management, 33(2), 193–208.

[San94b] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings, ACM Special Interest Group on Information Retrieval*, pages 142–151, 1994.

[San91] Sanfilippo, A., Thematic and Aspectual Information in Verb Semantics in *Belgian Journal of Linguistics 6*, 1991.

[San92a] Sanfilippo A. & V. Poznański (1992) The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento.

[San92b] Sanfilippo A., T. Briscoe, A. Copestake, M. A. Martì, M. Taulé e A. Alonge 1992 'Translation Equivalence and Lexicalization in the ACQUILEX LKB', in Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal.

[San93a] Sanfilippo A. (1993) Grammatical Relations in Unification Categorial Grammar. *Lingua e Stile, XXVII/2.*

[San93b] Sanfilippo, A., LKB encoding of lexical knowledge. In *Default inheritance in unification-based approaches to the lexicon*, T. Briscoe and A. Copestake and V. de Paiva (eds.), Blackwell, Cambridge, 1993.

[San94] Sanfilippo, A., K. Benkerimi, D. Dwehus 'Virtual Polysemy', in *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, 1994.

[San95] Sanfilippo A. (1995) Lexical Polymorphism and Word Disambiguation. In *Working Notes of the AAAI Spring Symposium on The Representation and Acquisition of Lexical Knowledge*, Stanford University, Ca.

[San97] Sanfilippo, A. (1997). Using semantic similarity to acquire co-occurrence restrictions from corpora. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

[San98] Sanfilippo, A. Ranking Text Units According to Textual Saliency, Connectivity and topic Aptness. COLING-ACL'98, August 10-14, 1998, University of Montreal, Montreal, Quebec, Canada.

[SanFC] Sanfilippo, A. (forthcoming) Lexical Undespecification and Word Disambiguation. In Viegas E. (eds) *Breadth and Depth of Semantic Lexicons*, Kluwer.

[Sat] Sato, S. (1992) CTM: An example based translation aid system, in Coling 1992, 1259-1263.

[Sch73] Schank R.C. (1973) Identification of Conceptualizations Underlying natural Language, in RC. Schank and K.M. Colby *Computer Models of Thought and Language*, Freeman and Company, San Fransisco, pp. 187-247.

[Sch75] Schank R.C. (1975) *Conceptual Information Processing*, Fundamental Studies in Computer Science, 3, Elsevier North-Holland/American.

[Sch81] Scha R. (1981) "Distributive, collective and cumulative quantification," in Groenendijk, Janssen, Stokhof (eds.), *Formal methods in the study of language*, Amsterdam: Mathematical Center.

[Sch96] Schiller, Anne. *Multilingual finite-state noun phrase extraction*, Workshop on *Extended finite state models of language*, European Conference on Artificial Intelligence (ECAI), Budapest, 1996.

[Sea69] Searle J. (1969) *Speech Acts.* CUP.

[Seg95] Segond Frédérique, Tapanainen Pasi, Using a finite-state based formalism to identify and g enerate multiword expressions. *Technical Report MLTT-019* Rank Xerox Research Centre, Grenoble, July 1995.

[Seg97] Segond, F., Schiller, A., Grefenstette, G., and Chanod, J. (1997). An experiment in semantic tagging using hidden markov model tagging. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

[Sek92] Sekine S., Carroll J.J., Ananiadou S., Tsujii J. (1992). 'Linguistic Knowledge Generator, in *Proceedings of COLING-92.*

[Sha48] Shannon (1948)

[Sha82] S. Shapiro. Generalized Augmented Transition Network Grammars for Generation from Semantic Networks. *American Journal of Computational Linguistics*, 8(2):12 – 25, 1982.

[Sig91] B. Sigurd. Referent Grammar in Text Generation. In C. Paris, W. Swartout, and W. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics.* Kluwer, 1991.

[Sil77] Silva G. et S.A. Thompson *On the syntax of adjectives with 'it' subject and infinitival complements in English*, in: Studies in Language, 1:1, 1977, pp. 109-126.

[Sim87] Simons, P. *Parts: A study in Ontology*, Clarendon Press, Oxford, 1987.

[Sin94] Sinclair J., M. Hoelter, C. Peters (Eds.) (1994) *The Languages of Definition: The Formalisation of Dictionary Definitions for Natural Language Processing*, Studies in Machine Translation and Natural Language Processing, Office for Official Publications of the European Communities, Luxembourg.

[Sma93] Smadja F. (1993) 'Retrieving Collocations from Text: Xtract, *Computational Linguistics*, vol. 19, n.1, March 1993, pp. 143-177.

[Sme95] Smeaton, A., Kelledy, F., and O'Donnell, R. (1995). Trec-4 experiments at Dublin city university: Thresolding posting lists, query expansion with wordnet and pos tagging of Spanish. In *Proceedings of TREC-4.*

[Sme96] Smeaton, A. and Quigley, A. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the* 19^{th} *International Conference on Research and Development in IR.*

[Sod95] Soderland, S. and Fisher, D. and Aseltine, J. and Lehnert, W., CRYSTAL: Inducing a Conceptual Dictionary, *Proc. of the 14th International Joint Conference on Artificial Intelligence*, 1995

[Som87] Somers H., 1987, Valency and Case in Computational Linguistics, Edinburgh University Press, Edinburgh.

[Sow84] Sowa, J.F. (1984) Conceptual Structures: Information processing in mind and machine. Reading, MA: Addison-Wesley publishing company.

[Sow92] Sowa, J. (1992) *Principles of Semantic Networks. Explorations in the representation of knowledge.*

[Sta96] Stairmand, Mark A. & William J. Black (1996). "Conceptual and Contextual Indexing of Documents using WordNet-derived Lexical Chains." In Proc. of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research.

[Ste88] Steiner, E., EckertU., Roth, B., Winter J. (1988). The Development of the EUROTRA-D System of Semantic Relations, In: Steiner, E., Schmidt, P., & Zelinsky-Wibbelt (eds), rom Syntax to Semantics: Insights from Machine Translation, London: Frances Printer.

[Sto81] Stowell, T., *Origins of Phrase Structure*, MIT doctoral dissertation, 1981.

[Sto89] Stowe, L. (1989) "Thematic Structures and Sentence Comprehension." In Carlson, G. and Tanenhaus, M. (eds) *Linguistic Structure in Language Processing*, Kluwer Academic Publishers, pp. 319-357.

[Str95] T. Strzalkowski. Information retrieval using robust language processing. In *AAAI Spring Symposium on Representation and Aquisition of Lexical Information*, pages 104–111, Stanford, 1995.

[Sun91] Sundheim, B. M. (1991). "Overview of the Third Message Understanding Evaluation and Conference," in *Proceedings of MUC-3.*

[Sus93] Sussna, M. (1993) Word sense disambiguation for free-test indexing using a massive semantic network. *Proceedings of the 2nd International Conference on Information and Knowledge Management.* Arlington, Virginia, USA.

[Sve86] Svensk ordbok (1986) Språkdata och Esselte Studium AB.

[Swa91] Swarts, H., *Adverbs of Quantification: A Generalised Quantifier Approach*, PhD Dissertation, Groningen University, 1991.

[Tal76] Talmy, L. 'Semantic Causative Types', *Syntax and Semantics*, 6, 1976.

[Tal85] Talmy, L. 1985 'Lexicalization Patterns: Semantic Structure in Lexical Form', in Shopen, T. (ed.), Language Typology and Syntactic Description: Grammatical Categories and the Lexicon, Cambridge, Cambridge University Press.

[Tap94] Tapanainen Pasi, RXRC Finite-State rule Compiler, *Technical Report MLTT-020*, Rank Xerox Research Centre, Grenoble 1994.

[Tau94] Taulé, Delor M., 'Towards a VRQS Representation', Esprit-BRA 7315, Acquilex-II WP 32, 1994.

[Tra94] Trask R.L. (1994). *A Dictionary of Grammatical Terms in Linguistics*, Routledge, London and New York.

[Tru92] Trujillo, A., Locations in the Machine Translation of Prepositional Phrases, in proc. TMI, Montreal, 1992.

[Tsu92] Tsujii J., Ananiadou S., Arad I., Sekine S. (1992). 'Linguistic Knowledge Acquisition from Corpora', in *Proceedings of 'International Workshop on Fundamental Research for the Future Generation of Natural Language Processing*, 30-31 July 1992, Manchester, UK, pp. 61-81.

[Tsu93] Tsujii, J.I. and Ananiadou, S. (1993) Knowledge based processing in MT. In Proceedings of the 1st international conference on Knowledge Bases and Knowledge Sharing, Tokyo.

[Van93] Van Valin, R. D., A Synopsis of Role and Reference Grammar, in *Advances in Role and Reference Grammar*, R. D. Van Valin (ed.), John Benjamins Publishing Company, Amsterdam, 1993.

[Van94] Vandeloise, C., Spatial Prepositions, Chicago University Press,1994.

[VanE94] Van den Eijk (1994) Van der Eijk (1993) Automating the acquisition of bilingual terminology In proc. of 6th Conf. of EACL, pp.113-119.

[Ven67] Vendler, Z., Verbs and Times, *Philosophical Review 56*, 1967.

[Ven68] Vendler Z. *Adjectives and Nominalization*, La Haye:Mouton, 1968.

[Ver72] Verkuyl, H. (1972) On the compositional nature of the aspects. Dordrecht: Reidel. 1972.

[Ver89] Verkuyl , H. (1989) Aspectual classes and aspectual distinctions Linguistics and Philosiphy, 12, 39-94. 1989

[Ver93] Verkuyl, H., *A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure*, CUP, Cambidge, 1993.

[Vic97] Vickery B.C. Ontologies. *Journal of Information Sciences*, 23(4):277–286, 1997.

[Voo93] Voorhees, E. (1993) Using WordNet to Disambiguate Word Senses for Text Retrieval. SIGIR-93.

[Vos93] Vossen P. and A. Copestake (1993) Untangling definition structure into knowledge representation In E.J. Briscoe, A. Copestake and V. de Paiva (eds.) Default inheritance in unification based approaches to the lexicon. Cambridge: Cambridge University Press. 1993.

[Vos95a] Vossen, P. (1995) Grammatical and Conceptual Individuation in the Lexicon, PhD. Thesis, University of Amsterdam, IFOTT, Amsterdam. 1995.

[Vos95b] Vossen, P., P. Boersma, A. Bon, T. Donker 1995 A flexible semantic database for information retrieval tasks. In: Proceedings of the AI'95, June26-30, Montpellier.

[Vos96] Vossen P. and A. Bon (1996) Building a semantic hierarchy for the Sift project, Sift LRE 62030, Deliverable D20b, University of Amsterdam. Amsterdam.

[Vos97a] Vossen, P. (1997) EuroWordNet: a multilingual database for information retrieval, In Proceedings of the Delos workshop on Cross-language Information Retrieval, March 5-7, 1997, Zurich.

[Vos97b] Vossen, P., P. Diez-Orzas, W. Peters (1997) The Multilingual Design of EuroWordNet. In P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.

[Vos98] Vossen, P. ???

[Wan94] Wanner, L. (1994). "Building another Bridge over the Generation Gap," in *Proceedings of the 7th International Workshop on Natural Language Generation*, Kennebunkport, Maine.

[War87] Warren, Beatrice, 1987. *Semantic Patterns of Noun-Noun Compounds*. Gothenburg Studies in English 41. Acta Universitatis Gothoburgensis, Gothenburg.

[Wie72] Wierzbicka, A. *Semantic Primitives*, Frankfurt, Athenaum Verlag, 1972.

[Wie80] Wierzbicka, A. *Lingua Mentalis: the Semantics of Natural Language*, Sydney, Academic Press, 1980.

[Wie85] Wierzbicka, A. *Lexicography and Conceptual Analysis*, Ann Arbor, Karoma Publ. Inc., 1985.

[Wie88] Wierzbicka, A. (1988) *The Semantics of Grammar*, Amsterdam, John Benjamins.

[Wie89a] Wierzbicka, A. 'Semantic Primitives and Lexical Universals', in *Quaderni di Semantica*, 10, 1, Bologna, Il Mulino, 1989a.

[Wie89b] Wierzbicka, A. 'Semantic Primitives - the Expanding Set', in *Quaderni di Semantica*, 10, 2, Bologna, Il Mulino, 1989.

[Wil75] Wilks, Y. (1975). "A Preferential Pattern Seeking Semantics for Natural Language Inference," *Artificial Intelligence*, Vol. 6.

[Wil97] Y. Wilks and M. Stevenson. The Grammar of Sense: using part-of-speech tags as a first step in semantic disambiguation. To appear in *Journal of Natural Language Engineering*, 4(3).

[Wil81] Williams, E. (1981a) "Argument Structure and Morphology." *Linguistic Review*, **1**, 81-114.

[Wil94] Williams, E., *Thematic Structure in Syntax*, Linguistic Inquiry monograph no. 23, MIT Press, 1994.

[Wil95] Wilkens Mike and Julian Kupiec, 'Training Hidden Markov Models for Part-of-speech Tagging', Internal document, Xerox Corporation, (1995).

[Win87] Winston M.E. - R. Chaffin - D.J. Hermann (1987) A taxonomy of part-whole relations In Cognitive Science, 11. Norwood NJ: Ablex Publ. Corp.: 417-444, 1987.

[Wis87] Wiston, M.E., Chaffin, R., Hermann, D., A Taxonomy of Part-Whole Relations, *Cognitive Science*, 11, 417-444, 1987.

[Woo] Woods, W.A. Conceptual Indexing in, http://www.sunlabs.com/technical-reports/1995/annualreport95/conceptual.html

[Yan94] Yang, Y. and Chute, C.G., Words or concepts: the features of indexing units and their optimal use in information retrieval, *Proc. of SCAMC '94*, 1994, 685-689

[Yar92] D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France, 1992.

[Yar93] D. Yarowsky. One sense per collocation. In *Proceedings ARPA Human Language Technology Workshop*, pages 266–271, Princeton, NJ, 1993.

[Yar95] D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Lainguistics (ACL '95)*, pages 189–196, Cambridge, MA, 1995.

[Zel88] Zelinsky-Wibbelt, C. (1988). From Cognitive Grammar to the Generation of Semantic Interpretation in Machine Translation. In: Steiner, E., Schmidt, P., & Zelinsky-Wibbelt (eds), rom Syntax to Semantics: Insights from Machine Translation, London: Frances Printer.

[Zel93] Zelinsky-Wibbelt, C., The semantics of Prepositions, Mouton de Gruyter, 1993.