SEMANTICS IN ACTION

EVELYNE VIEGAS, KAVI MAHESH, SERGEI NIRENBURG
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003, USA
viegas,mahesh,sergei@crl.nmsu.edu

Abstract. In this paper, we describe our general approach to computational lexical semantics, very briefly, focusing on the dynamics of semantics, and in particular of events. We investigate what semantic information should be encoded in a lexicon entry so that it contributes best to the construction of a Text Meaning Representation (TMR) and to the generation of a text from the TMR. Within this approach, lexical items constitute the building blocks of the sentences of a text. We are interested in showing that static knowledge sources can be changed dynamically, to fit the linguistic context, at processing time if we have the mechanisms to enable such changes. Moreover, our experience shows that determining the meaning of lexical items is not a trivial task in lexicon acquisition, and as humans still play a crucial role in lexicon acquisition, this problem must be addressed head on. We show that it is possible to reconcile the subjectivity of acquirers with the recognition of the "core" meaning of a word, if we develop a theory of computational (lexical) semantics which can account for some of the combinatory and productive principles of natural languages.

1. Introduction

The ultimate goal of studying semantics is to investigate the mechanisms of obtaining a meaning representation given a text and vice versa. A complete semantic theory includes both knowledge and processors. In this paper, we concentrate on the knowledges in action: we investigate the type of information that should be encoded in the knowledge sources, along with mechanisms to enhance them, to have them best used by the processors in order to treat all complex phenomena found in natural languages. Our atti-

tude to the static knowledge bases (the lexicon, the ontology, the grammar, the special "microtheory" rules) is almost strictly functional: the knowledge which is acquired is, from the outset, intended for use to help extract and represent meaning given a text or realizing a text given its meaning.

Mikrokosmos is a Knowledge Based Machine Translation (KBMT) system under development at New Mexico State University jointly with the US Department of Defense (Beale et al., 1995; Onyshkevych and Nirenburg, 1994). The Mikrokosmos team[1] has been working on many topics, such as the relation between underlying meaning and surface realisations, multilinguality aspects, syntactic and semantic dependency knowledge, acquisition of (lexical) semantic knowledge, and application to (multilingual) generation and machine translation. As its experimentation domain, Mikrokosmos has focused so far on translating Spanish news articles to English. By the end of 1995, a core lexicon of approximately 7,000 Spanish word senses supported by an ontology of about 5,000 concepts was in place. After application of lexical rules,[2] the expanded Spanish lexicon contained about 40,000 word senses. High quality semantic analyses of article-length Spanish texts in the domain of company mergers and acquisitions have already been produced, as has been shown through the results we obtained in the Word Sense Disambiguation task, reported in section 2.1. We are now in the process of generating these analyses into English, as reported in (Beale, Viegas and Nirenburg, 1997). A set of graphical tools has been developed to support knowledge acquisition, system development, testing, and demonstration of each phase of language processing and machine translation. In addition, automated techniques and detailed methodologies have also been developed for testing lexicons and ontologies as well as assessing the performance of language processing components.

First, a terminological note on the central notion of the predicate, as this is the topic of the present book. From a logical or syntactic perspective (the ones habitually assumed in NLP),[3] a predicate can be said to be a named n-ary relation between arguments, which concerns itself more with capturing

[1]The Mikrokosmos team includes several other researchers, in particular Stephen Beale, Boyan Onyshkevych and Victor Raskin. The Mikrokosmos project can be seen at the URL http://crl.nmsu.edu/Research/Projects/mikro

[2]We can briefly illustrate the application of morpho-semantic lexical rules on the word *comprar*. Applying the rules on the entry for the Spanish verb *comprar* produced automatically 26 new entries (*comprador-N1*, *comprable-Adj*,...). This includes creating a new mapping with the correct subcategorisations and also the right semantics; for instance, the lexical entry for *comprable* will have the subcategorisation of an adjective and the semantics adds the attribute "feasibility-attribute", as reported in (Viegas et al, 1996).

[3]The oldest notion of predicate from Antiquity lies on a binary semantic opposition between what is called "theme/rheme" or "subject/predicate" and is still active in theoretical linguistics and philosophy (Grimes, 1968; Searle, 1969).

the semantics of events than with capturing the semantics of objects. In other words, this notion of predicate has been usually applied to events, not objects; or rather, since the distinction between natural language and the language of representation has not yet been well appreciated by lexical semanticists, to verbs, not nouns. From this formal viewpoint, this notion seems to be merely equivalent to verbs or relational nouns or prepositions or relational adjectives viewed together with their syntactic dependency (subcategorisation) frames.

For us, semantic *events* are defined by the ontological (that is, non-syntactic) features (and their values) associated with them. Moreover, these features are not equivalent to subcategorisation arguments (which constitute the arguments of a predicate), as they can have no surface realisation at all.

In this paper, we describe our general approach to computational lexical semantics, very briefly, focusing on the dynamics of semantics, and of events in particular. We investigate what semantic information should be encoded in a lexicon entry so that it contributes best to the construction of a Text Meaning Representation (TMR) and to the generation of a text from the TMR. In Mikrokosmos, lexical items constitute the building blocks of the sentences of a text. We use the same frame-based representational formalism (Minsky, 1968; Luger and Stubblefield, 1992; Fillmore, 1985, 1993) to encode information in lexicon entries and to build TMRs; in fact, a lexicon entry constitutes an unsaturated piece of TMR.

We are interested in showing that static knowledge sources can be changed dynamically, to fit the linguistic context, at processing time if we have the mechanisms to enable such changes. Our experience shows that determining the meaning of lexical items is not a trivial task in lexicon acquisition.[4] However, as humans still play a crucial role in lexicon acquisition (Gross, 1984; Lenat et al., 1986; Nirenburg and Raskin, 1987; Normier and Nossin, 1990; McNaught, 1990), this problem must be addressed head on. We show that it is possible to reconcile the subjectivity of acquirers with the recognition of the "core" meaning of a word, if we have developed a theory of computational (lexical) semantics which can account for some of the combinatory and productive principles of natural languages.

In section 2, we give an example of how we use semantics to produce the TMR of a text. In section 3, we address representational issues with respect to the lexicon, showing the (lexical) semantic information that should be in-

[4]One could object that we do not need to determine the meaning of a lexical item but just the relations between lexical items, which is commonly done is lexical semantics (Cruse, 1986). However, for Natural Language Processing (NLP) applications, it is not enough to have the relations between lexical items, especially from a multilingual perspective, and researchers working in computational lexical semantics have added a conceptual layer in their lexicons (e.g. Sanfilippo, 1992).

cluded in the lexical items, paying particular attention to EVENTs, in order to support multilingual NLP applications, in our case Machine Translation (MT). In section 4, we introduce the methodology of acquisition within a multilingual environment, paying particular attention on the one hand to the trade-offs between language dependent and language independent information, and on the other hand between acquirers' different points of view. In particular, we show that if the acquisition is done in a "situated environment", i.e. with particular tasks in mind, and if we rely on a theory of computational (lexical) semantics which allows the static knowledge sources to be changed dynamically, then lexical items can be made to fit the linguistic context in which they appear, at processing time. Finally, in section 5, we compare our approach to other main trends in computational (lexical) semantics; in particular, we address the issue of sense enumeration, arguing that this is a lexicographic concern, not a computational semantic concern.

2. Semantics in Action

As we mentioned earlier, our approach to semantics is almost strictly functional. This is why we start showing what we can do with the static knowledge sources, which will be discussed more thouroughly in later sections. We illustrate our approach to processing text meaning by tracing one of the most difficult tasks in NLP, namely, word sense disambiguation. We begin by presenting the results from teh Mikrokosmos semantic analyser, and then illustrate how they were obtained.

2.1. RESULTS

Experiment 1: Mikrokosmos semantic analyser applied on 4 Spanish texts from real-world texts (news articles on company mergers and acquisitions from the EFE newswire), in which there were no unknown words to Mikrokosmos.

Table 1 shows sample disambiguation results from Mikrokosmos. These are results from analysing four real-world texts. The average text was 17 sentences long, with over 21 words per sentence. For evaluation purposes, correct senses for all the open class words in the texts were determined by a native speaker. Mikrokosmos selects the right sense of open-class words about 97% of the time. Syntactic analysis contributed to about 38% of word sense disambiguation, e.g. when different word senses of a lexeme had different subcategorisations. The performance on the first and third texts, Text 1 and Text 3 respectively, was worse than the performance on the other two texts. These texts had longer sentences (see line 2 in Table 1), and/or many more ambiguous words (see line 4 in Table 1), and construc-

Text	1	2	3	4	Mean
1 - words	347	385	370	353	**364**
2 - words per sentence	16.5	24.0	26.4	20.8	**21.4**
3 - open-class words	183	167	177	177	**176**
4 - ambiguous open-class words	57	42	57	35	**48**
5 - ambiguous words resolved by syntax	21	19	20	12	**18**
6 - ambiguous words correctly resolved	89%	98%	79%	97%	**91%**
7 - total words correctly resolved	51	41	45	34	**43**
8 - % of total words correctly resolved	97%	99%	93%	99%	**97%**

TABLE 1. Mikrokosmos Results in Disambiguating Open Class Words in Spanish Texts.

tions that make disambiguation hard (e.g., ambiguous words embedded in appositions).

Note that we address all types of ambiguity in Mikrokosmos, homography, homomorphy and polysemy.[5] Briefly, homography refers to words which have the same spelling, such as [bank-N1, financial-institution] and [bank-N2, river-bank], we treat such cases at the lexical level in Mikrokosmos, with different entries ([bank-N1] and [bank-N2]); homomorphy refers to words which have the same form and different Parts-of-Speech, such as [fast-Adj] and [fast-Adv], we treat such cases at the syntactic level in Mikrokosmos, with two lexicon entries [fast-Adj1] and [fast-Adv1]; finally, polysemy refers to words which have the same form and different semes, such as [record-N1, music - artifact], we treat such cases in Mikrokomos at the semantic (or ontological) level, we thus have one lexicon entry for [record-N1]. Note that polysemy is a lexicographic or computational concern and as such it is a virtual (or non-existent) ambiguity for people, who can easily select one seme over the other(s) in context. All these types of ambiguity are expressed by concepts along with constraints on the features attached to these concepts.

Note also that most of the researchers using statistic techniques when doing word sense disambiguation also get impressive results; however, they are concerned with disambiguating homographs only (getting about 92% of accuracy on a mean 3-way sense distinction, counting all open class words in

[5]See Weinreich (1964) on contrastive and complementary ambiguities; Cruse (1993) on facets; Pustejovsky (1995) on logical polysemies, for various interesting accounts on homography/polysemy.

(Yarowsky, 1992)), and homomorphs only (getting about 92% counting all open class words in (Wilks, 1996)). As such, lines 6 in Table 1 and Table 2 are absent from the results given by these researchers. In other words, in our task of word sense disambiguation, we address, within a KBMT framework, more issues than have been formerly addressed. However, to be completely accurrate, the numbers shown here should be refined by indicating the percentage mean per n-way sense distinction. We have not done such an experiment on a large scale, just because in the case of Yarowsky, only a dozen words were considered in his experiment whereas we dealt with an average of 48% ambiguous open-class words per text. Although the average of number of senses in our lexicon is about 1.2 sense per lexeme, we still have lexemes which have more than 2 meanings or senses, especially in the case of homographs. The reason why the number of senses in the lexicon is kept lower than in some Machine Readable Dictionaries (MRDs) such as LDOCE, is because polysemy is captured at the semantic level, via the ontology, thus leaving the entries in the lexicon as vague or underspecified, leaving the task to the semantic analyser to produce the right sense in context.

Experiment 2: Mikrokosmos semantic analyser applied on a Spanish text not used in knowledge acquisition.

The above four texts were among about 400 Spanish texts used in the general lexicon and ontology acquisition process in Mikrokosmos. Table 2 shows the results on a previously unseen text. The results were essentially similar to those for the training texts in Table 1.

1 - number of words	390
2 - number of words per sentence	26
3 - number of open-class words	104
4 - number of ambiguous open-class words	26
5 - number of words resolved by syntax	9
6 - % of ambiguous words correctly resolved	88.9%
7 - total of words correctly resolved	23
8 - % of all open-class words correctly resolved	97.1%

TABLE 2. Mikrokosmos Results on an Unseen Text.

The unseen text used in the experiment contained 19 words missing from the Mikrokosmos Spanish lexicon. In such cases, the Mikrokosmos analyser

produces dummy entries, marked as nouns and semantically mapped to ALL, the root concept in the ontology (this has the effect of essentially not including any semantic constraints in the definition). No changes were made in the lexicon, ontology, or the programs. There were many syntactic binding problems with this text; for instance, a lexicon entry would not allow some of the subcategorisations found in the text. We did not fix any of them. We could get even better results if we assumed perfect syntactic output and fixed all the binding problems. Unknown words (which were mapped to ALL) were treated as unambiguous. Twelve of the 19 unknown words appeared to be proper names and only 3 unknown words were in fact ambiguous. Among the 85 words present in the lexicon, only 13 words had been seen in previous analysed texts.

Overall, one can say that the Spanish core lexicon is of a good quality, as it provided enough constraints to help in the task of disambiguation even in the presence of unknown words to our lexicon, and for a text which has not been considered during the acquisition. This experiment does validate our methodology, where the "training" of the static resources only concerns correcting errors in the lexicon or the ontology. This includes such revisions as correcting wrong syntactic class assignments or wrong mappings to a concept or constraints on a concept, and is part of the lexicographic loop in acquisition; but it is not aimed at "hard-coding" lexico-semantic information in the entries so that they give the best results for a particular corpus. Our methodology of "training" is thus quite different from the "training" phase within a statistic approach, where coefficients are tweaked to give best results for a given corpus.

We now address a sentence which is deliberately simplified: it shows a "hitch-free" analysis of a simple example.[6] In reality, the average length of the sentences in the texts on which the analyser was tested was over 21, and the processing often led to outcomes which required additional processing, not a simple process of instantiation and combination, as described in the example. See (Beale et al., 1995), for a more detailed description of the Mikrokosmos analyser.

Consider the Spanish sentence:

[6]This "hitch-free" example is extracted from the sentence of the following text extract. The actual text has 17 sentences: *El grupo Roche, a través de su compañía en España, adquirió el laboratorio farmacéutico Doctor Andreu, se informó hoy aquí. La comunicación oficial no precisó el monto de la operación realizada entre Productos Roche SA y Unión Explosivos Río Tinto SA, hasta ahora mayoritaria en el accionariado.* (The Roche Group acquired the pharmaceutical laboratory Doctor Andreu through its company in Spain, it was announced here today. The official announcement did not specify the exact amount of the transaction which took place between Productos Roche SA and Unión Explosivos Río Tinto SA, which until now had held the majority of the stock.)

(1) El grupo Roche, a través de su compañía en España, adquirió Doctor Andreu.

(The Roche group, through its company in Spain, acquired Doctor Andreu.)

Many words in this sentence are ambiguous: *a través de* can have the semantic roles of Path or Instrument; *compañía* can be CORPORATION or SOCIAL-EVENT; *en* is many ways ambiguous; and *adquirir* can be ACQUIRE or LEARN. In this illustration, we focus on disambiguating *adquirir* although our system resolves all of the above ambiguities. The desired meaning of the sentence is the following, extremely simplified, TMR:

```
ACQUIRE-1
        Agent: ORGANIZATION-1
        Theme: ORGANIZATION-2
        Instrument: ORGANIZATION-3
ORGANIZATION-1
        Object-Name: Grupo Roche
        Agent-Of: ACQUIRE-1
ORGANIZATION-2
        Object-Name: Doctor Andreu
        Theme-Of: ACQUIRE-1
ORGANIZATION-3
        Location: NATION-1
        Instrument-Of: ACQUIRE-1
NATION-1
        Object-Name: Espana
        Location-Of: ORGANIZATION-3
```

It must be noted that all of the labels used in the above TMR are well-defined concepts in the Mikrokosmos ontology that we will develop in the next section. This TMR, in absence of microtheories such as focus, theme/rheme oppositions, can be generated in English by any of the following:
The Roche group, through its company in Spain, acquired Doctor Andreu.
The Roche group acquired Doctor Andreu, through its company in Spain.
The acquisition of Doctor Andreu by the Roche Group was made through its company in Spain.

2.2. GENERATING CONSTRAINTS

The first step for the Mikrokosmos analyser is to gather up all of the possible lexicon entries for each of the words. For instance, for *adquirir*, the two lexicon entries *adquirir-V1* and *adquirir-V2* are retrieved, with mappings into the concepts ACQUIRE and LEARN. For each word sense, the syntactic mapping, done via a variable binding process, must be examined to see if

it fits the current sentence. For *adquirir*, both word senses have identical syntactic mapping, so the variable binding process applies to the two entries *adquirir-V1* and *adquirir-V2*. After variable binding, the semantic analyser examines the semantics of each word sense in order to construct a list of constraints that must be satisfied for that word sense.

2.3. APPLYING CONSTRAINTS

Mikrokosmos employs an ontological graph search mechanism, Onto-Search, (Onyshkevych, 1997) to check constraints. Onto-Search, determines relevant paths between any two concepts and returns a score based on their degree of closeness. For example, the command **check-onto-con**(ACQUIRE EVENT)[7] returns a score of 1.0 (out of 1.0) since ACQUIRE is a type of EVENT. However, **check-onto-con**(ORGANIZATION HUMAN) returns a score of 0.9 along with the path (ORGANIZATION HAS-MEMBER HUMAN). This indicates that ORGANIZATION can stand in the place of HUMAN because it has HUMAN members. This and other types of metonymy are frequent in natural language and are detected automatically by Mikrokosmos.

2.4. DETERMINING THE BEST COMBINATION OF WORD SENSES

Each combination of word senses activates the applicable constraints, which are combined into a total score for the combination. The combination with the best total score is chosen as the basic Semantic Dependency Analysis, the core TMRs to which other microtheories (such as aspect and coreference) can be applied. In the example sentence, the following choices were made:

1. *a-través-de* is INSTRUMENT, since its LOCATION meaning would require *adquirir* to be a PHYSICAL-OBJECT.
2. *en* is LOCATION, since its TEMPORAL meaning would require *españa* to be a TEMPORAL-OBJECT.
3. *adquirir* maps into ACQUIRE, since its LEARN sense would require *Dr-Andrew* to be INFORMATION.
4. *Dr-Andrew* is an ORGANIZATION, since its HUMAN meaning cannot be the theme of an ACQUIRE concept.
5. Mikrokosmos currently has trouble choosing between the CORPORATION and SOCIAL-EVENT meaning of *compañia*, the object of the *á-traves-de* PP-adjunct. Both can have locations in Spain, and both can be INSTRUMENTS of EVENTs. At this point, Mikrokosmos needs to add information into the ontology that ORGANIZATIONs

[7] Which asks "Is ACQUIRE an EVENT?"

can typically fill the INSTRUMENT slot of ACQUIRE acts, but SOCIAL-EVENTs cannot. Another alternative could consist in tuning our lexicons towards the domains using statistical techniques.

To summarise, disambiguation decisions for a word can rarely be done independently of the decisions on other words. The Mikrokosmos analyser therefore operates as follows:

- Derive selectional constraints from the lexicon and the ontology for each pair of syntactically dependent words, in both directions.
- Check each constraint by finding the "distance" between the pair of concepts in the ontology.
- Combine the results in an efficient constraint satisfaction algorithm (Beale et al., 1996) to select the best combination of senses for all the words in a sentence.

The above description gives a general view of our approach to the application of selectional constraints during processing, and some insight into our approach to computational semantics. We now describe the Mikrokosmos overall process including the static knowledge sources in Figure 1.

Input texts are passed through a multi-stage morpho-syntactic analyser for Spanish called Panglyzer (Farwell et al., 1994). The resulting syntactic trees are transformed to an LFG-like syntactic structure (Bresnan, 1982). Lexical entries from the Mikrokosmos lexicon are instantiated for each of the root forms in the syntactic structure. Syntactic variables in the lexical instantiations are bound to one another using the syntactic patterns in the lexical entries to establish syntactic dependencies and map them to semantic dependencies. In addition, ontological concepts referred to the semantic zones of the lexical entries are instantiated. In the next step, selectional constraints are retrieved from the ontology and added to those encoded in the lexicon. Individual selectional constraints are checked by an ontological search program called Onto-Search (Onyshkevych, forthcoming). The resulting preference values for each constraint are combined in an efficient control and search algorithm called Hunter-Gatherer that combines constraint satisfaction (Tsang, 1991), branch and bound, and solution synthesis techniques (Freuder, 1978) to pick the best combination of word senses of the entire sentence in near linear time, as described in (Beale, 1997), (Beale et al., 1996). Chosen word senses are assembled into TMR frames using the lexical semantic representations from the lexicon. Finally, a variety of microtheories are applied to further analyse elements of text meaning such as time, aspect, propositions, sets, coreference, and so on, to produce the final TMR.[8]

[8] A few microtheories are already in place and several are currently being developed by the Mikrokosmos team.

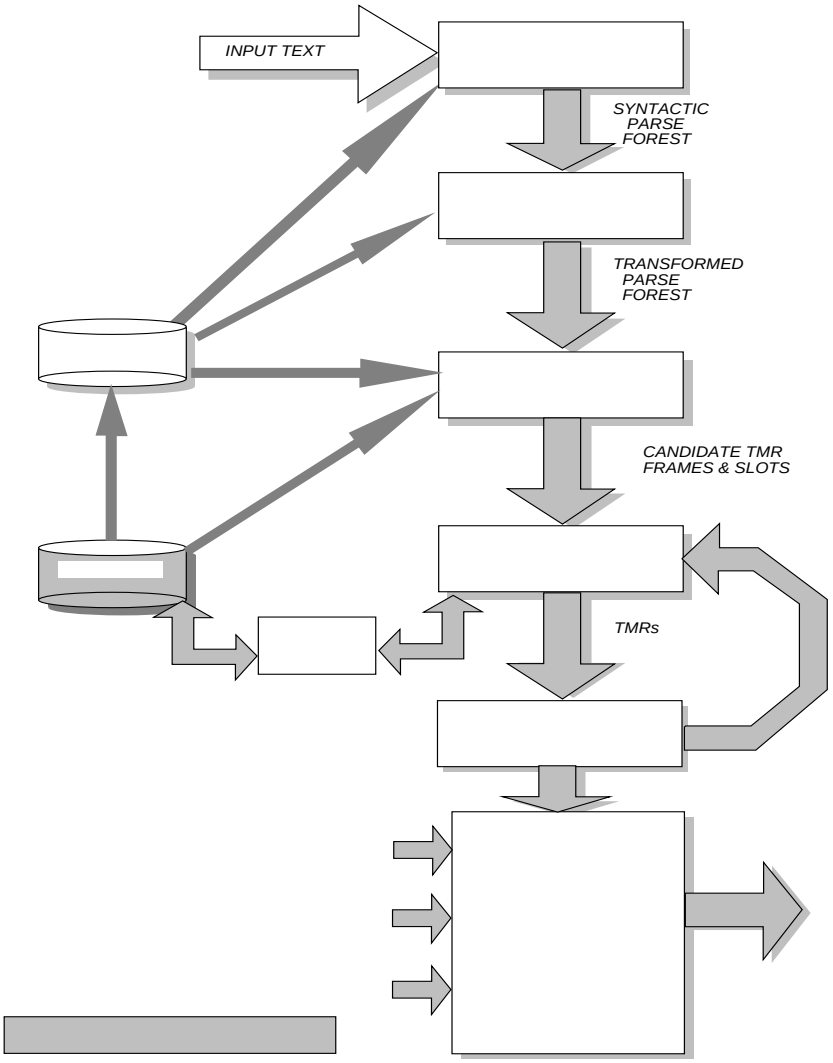


Figure 1. The Mikrokosmos Analysis Architecture.

In this section, we have shown the general architecture of the Mikrokosmos MT project, in terms of knowledge resources that are required to complete an NLP task, such as analysis. We have provided an evaluation of the Mikrokosmos analysis process through the task of word sense disam-

biguation, showing that we obtain even better results than statistic-based approaches which constrain their word sense disambiguation algorithms to deal with a sub-set of the types of ambiguity. One could object that our approach has obvious limitations in that it is expensive to evaluate comprehensively and on a large corpus of inputs, due to high start-up costs of knowledge acquisition and on the inevitable complexity of any realistic application-oriented evaluation scheme. However, our experience shows that it is possible to reduce the acquisition costs. In Mikrokosmos, we acquired an ontology of about 5,000 concepts (Mahesh, 1996; Mahesh and Nirenburg, 1995) and have acquired semi-automatically about 7,000 word senses for the Spanish core lexicon, with an average of 1.2 meaning per word-form, as described in (Viegas and Raskin, 1997). This semi-automatic acquisition of the core lexicon, has been extended to the automatic acquisition of about 32,000 new entries, using morpho-semantic lexical rules (derivational morphology), as reported in (Viegas et al., 1996).

3. Representational Issues

3.1. THE LEXICON

The lexicon primarily connects with the ontology and the onomasticon (a special-purpose lexicon of named entities such as cities, corporations, or products names), thus becoming the locus of links between lexical units in texts and the TMR.[9] Each lexical entry contains a representation of its semantics, represented by using terms from the ontology, in addition to other non-ontological primitives, e.g., to reflect speaker attitudes and modality. The advantages of connecting the lexicon to an ontology are threefold: 1) it allows one to capture the conceptual properties of particular words of a natural language in a "language-impartial" way, thus favouring cross-linguistic communication; 2) it allows for better processing in a multilingual environment; 3) it is cost-effective for multilingual NLP applications, as only the "language-dependent" properties have to be acquired when adding new natural languages to the system.

It is important to note that we adopt a transcategorial approach, where syntactic categories and semantic or ontological categories are not automatically related. For example, although many verbs are EVENTs and a number of nouns are represented by concepts from the OBJECT subtree (such as the class of artifacts), frequently this is not the case. This is particularly true with words derived via Lexical Rules, (LRs).

[9]For a review on computational lexical semantics, the reader can consult (Nirenburg and Raskin, 1996), where the authors describe the different approaches to lexical semantics.

3.1.1. *SYN Zone*

The content of the SYN zone of a lexicon entry provides the basis of the syntax-semantics interface. The information contained in this zone essentially amounts to an underspecified piece of a syntactic parse of a sentence using the lexeme. For instance, in our example sentence, while processing the *adquirir* lexicon entry, *Grupo Roche* will be bound to [1] as the SUBJ, while *Dr. Andreu* will be bound to [2] as the OBJ.

3.1.2. *SEM Zone*

The SEM zone provides the mapping to the output semantics. Each SEM zone provides the Lexical Semantic Representation (LSR) of the lexical item, and constitutes an unsaturated TMR fragment which includes as much meaning as can be extracted from the word being processed. The interaction of SEM zones from all the words in the sentence, (along with information added by other microtheories), result in the final TMR outputs. The formalism for the lexical semantic specification in this zone in our lexicon has been discussed in detail in other sources, such as (Onyshkevych and Nirenburg, 1994), (Onyshkevych, 1995).

We illustrate the SYN and SEM notions through relevant fragments of the Spanish lexicon entry for *adquirir*, shown in (Figure 2).[10]

Figure 2 shows the SEM zone of adquirir-V1 calls for an ACQUIRE concept (Figure 3) with AGENT and THEME slots that will be filled by the TMR names that are produced by *Grupo Roche* ([1]) and *Dr. Andreu* ([2]), respectively. Other words in the sentence can fill in additional information in the ACQUIRE TMR. One of the meanings of "a través de," treated as a phrasal entry, will add an INSTRUMENT slot. The adquirir-v2 SEM zone calls for a LEARN concept (Figure 4) with a theme of type INFORMATION. The information shown in the SYN zone here is partial. In fact, adquirir-V1 has optional pp-adjuncts, lexicalising the semantics of ACQUIRE, such as the SOURCE, e.g. *from ....*

3.1.3. *Other Lexicon Zones*

The SYN zone and SEM zone are the main zones used to produce the TMRs. The lexicon includes many other zones, some which participate in the production of the TMR, some which provide information mostly targeted at the task of generation, and finally some which serve knowledge management purposes. Briefly, a lexeme is minimally described via **10 zones** corresponding to various levels of lexical information, relevant to phonology, orthography, morphology, syntax-semantic linking, stylistics,

[10] We use the typed feature structures (tfs) as described in (Pollard and Sag, 1987).

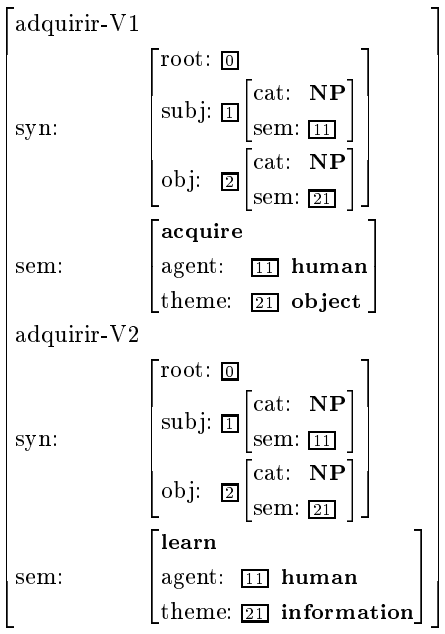


Figure 2. Sense (Partial) Entries for the Spanish lexical item *adquirir*.

and paradigmatic and syntagmatic information, along with sub-zones containing triggers for analysis and generation.[11]

3.2. THE ONTOLOGY

The ontology is a large collection of information about EVENTs, OBJECTs and PROPERTYs in the world (Mahesh and Nirenburg, 1995; Mahesh, 1996).[12] In addition to the taxonomic multi-hierarchical organisation, each concept has a number (currently averaging 14) of other local or inherited links to other concepts in the ontology, via relations (themselves defined in the PROPERTY sublattice). Figure 3) and Figure 4 shows the information found in the events ACQUIRE and LEARN respectively.

This section dealt with the way we represented knowledge in the lexicon, mainly by mapping the meanings of words onto a "language-impartial" conceptual world or ontology. In next section, after briefly discussing the methodology for acquisition, we turn to the decisions which have to be

[11] See (Viegas and Raskin, 1997), (Meyer et al., 1990) for explanations on the necessity of these zones, and on how to acquire them.

[12] The Mikrokosmos ontology is available on-line for browsing or downloading for research purposes at the URL http://crl.nmsu.edu/Research/Projects/mikro/ontology/

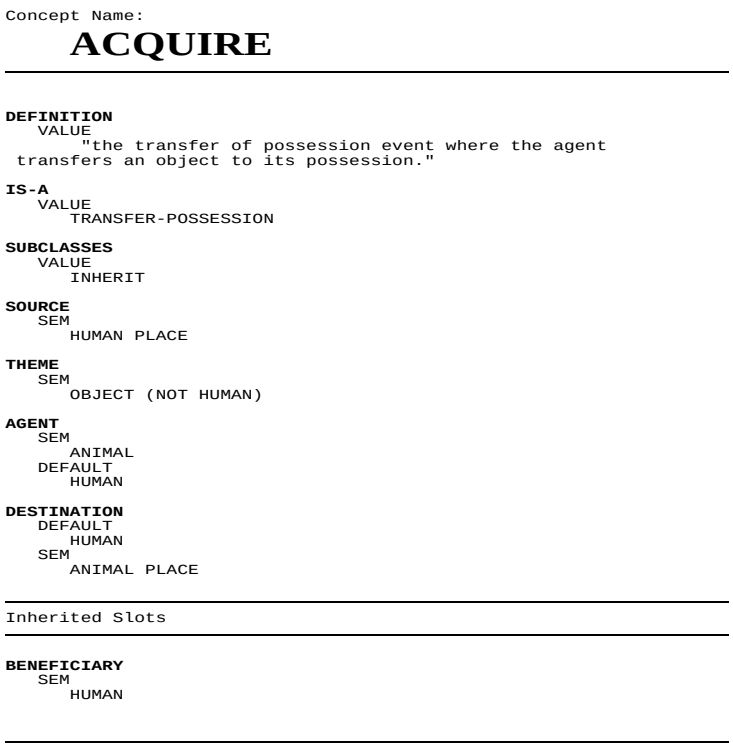


Figure 3. Conceptual frame for ACQUIRE.

made by the acquirers of ontologies and lexicons to build lexical entries which will contribute best to 1) capturing the meanings of words as well as 2) constituting the building blocks of the meaning of a text. In particular, we will address the issues of "language-impartial" versus "language-dependent" trade-offs within a multilingual environment, and of semantic ambiguity in terms of sense enumeration. We argue that these are mainly lexicographic concerns, and not computational semantic concerns: the static knowledge sources should be allowed to be changed dynamically at processing time, to 1) fit the linguistic context in which they appear, 2) reconcile different acquirers' points of views on the data.

4. Acquisition and Multilinguality

4.1. ONTOLOGY ACQUISITION

The ontology is being acquired incrementally, relying on continuous interactions with lexicon and semantic analyser teams. A series of negotia-

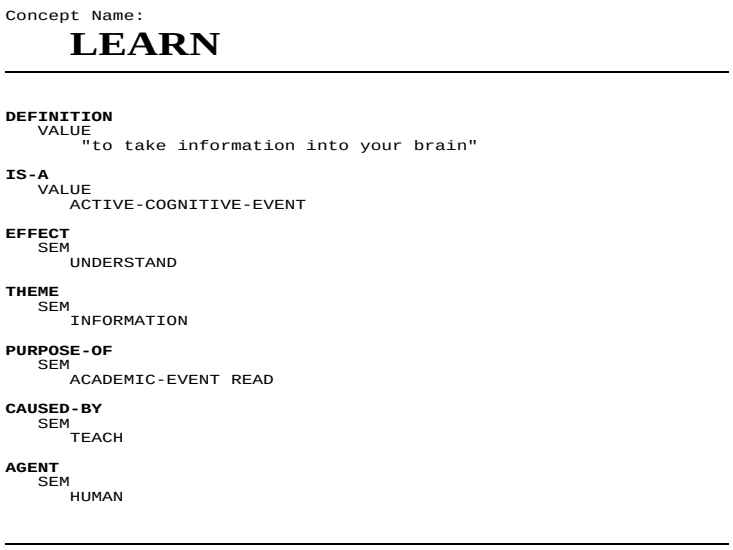


Figure 4. Conceptual frame for LEARN.

tions between lexicographers, ontologists, and developers of Onto-Search (see Section 2.3.) leads to the best choice of meaning representation in each case. It also ensures that every entry in each knowledge base is consistent, compatible with its counterparts, and has a purpose towards the ultimate objective of producing quality TMRs. We refer to this method as *situated development* of ontologies and lexical resources. This method is ideal for a multilingual situation such as in Mikrokosmos where it ensures that representational needs of more than one language are taken into account.

Ontology acquisition is a very expensive empirical task. Situated development is a good way to constrain the process and make it attainable. For example, the acquirer must focus on concepts in the domain of the input texts and thereby increase the ratio of the number of concepts (or their slots) that are actually used in processing a set of texts to the total number of concepts present in the ontology. The best example of a large ontological database acquired with enormous efforts but entirely out of any situation is CyC (Lenat and Guha, 1990). While the utility of CyC in a particular situation such as large scale NLP is yet to be demonstrated (Mahesh

et al., 1996), it is also true that most projects cannot afford to spend as many resources as it has taken to develop CyC and must strive to constrain acquisition significantly or share existing ontologies.

Methodology and Guidelines The basic methodology for concept acquisition employed in the Mikrokosmos project involves a fine-grained cycle of requests for concepts from the lexicon acquisition team and the resulting responses which may involve pointing out an existing concept, adding a new concept, enhancing the internal structure of one or more concepts, or suggesting a different lexical mapping for the word in question. If it is determined that a word sense requires a new concept in the ontology, the "algorithm" applied for adding the new concept hinges on viewing the ontology as a discrimination tree. The acquirer discriminates from the top down until at some point there is no child that subsumes the meaning in question. A new concept is added as a child at that point. In the Mikrokosmos project, sets of guidelines have emerged for making various kinds of decisions in ontology acquisition. These guidelines, some of which are shown in Figures 5, collectively define the methodology for ontology building (Mahesh, 1996).

4.2. LEXICON ACQUISITION

Acquiring a large-scale computational semantic lexicon is a very expensive enterprise, this is why it is advantageous to build lexicons which are reusable for other domains or applications. We need lexicons that are multi-purpose, supporting the three following paradigms:

a **multi-lingual**: French, English, Japanese, Russian, Spanish, etc..., *(encoding of characters)*

b **multi-media**: containing linguistic information for natural language processing, phonological information, essentially for speech recognition and production, and graphics, motion for visual processing, ... *(structure of the lexicons)*

c **multi-process**: applicable for analysis, generation (both mono- and multilingual), MT, summarization, information extraction, or speech processing,... *(reversibility of the lexicons)*

In other words, we develop one lexicon per language, and this lexicon can be used for different applications by automatically reindexing the lexicon as we did for instance in (Viegas and Beale, 1996) to create an English generation lexicon from an existing Spanish analysis lexicon. Having as a goal a multi-purpose lexicon saves a lot of time in acquisition.

The process of acquisition itself has been reported in many documents by the members of our team. People interested in the detailed process of

1. Do not add instances as concepts in the ontology. Rules of thumb for distinguishing an instance from a concept are:
   - **Instance-Rule1:** See if the entity can have its own instance. Instances do not have their own instances; concepts do.
   - **Instance-Rule2:** See if the thing has a fixed position in time and/or space in the world. If yes, it is an instance. If not, it is a concept. For example, SUNDAY is a concept, not an instance, because it is not a fixed position in time ("last Sunday," "the first Sunday of the month," etc.).
2. Do not decompose concepts further into other concepts merely because you can. It is important to focus on building those parts that are needed immediately for the Mikrokosmos task. For example, though EVENTs like BUY or MARKETING can be decomposed to a great extent, unless there is an indication that detailed decompositions are needed for the task, do not decompose such EVENTs.
3. Do not add a concept if there is already one "close" to it or slightly more general than the one being considered. Consider the expressiveness of the representation provided by gradations (i.e., attribute values) before adding separate concepts. For example, we do not need separate concepts for *suggest*, *urge*, and *order*. They are all gradations of the same concept, a DIRECTIVE-ACT, with various degrees of force which can be captured in an appropriate attribute.
4. Do not add specialized EVENTs with particular arguments as new concepts. For example, we do not need separate concepts for "walk to airport terminal" and "walk to parking lot."
5. Certain elements of text meaning such as aspect, temporal relations, attitudes, and so on, that are instance-specific belong only in the TMRs. For example, BREAKFAST is probably a concept in the ontology (and a subclass of MEAL, say) but a meal that happened at 3 O'clock on a particular day is not a separate concept in the ontology.
6. If any part of a meaning representation is specific to a particular language that part does not belong in the ontology.
7. Mikrokosmos representations have a very expressive `set` and `subset` notation. Hence, there is no need to create ontological concepts for collections of different types of things in the world.

Figure 5. Guidelines for Deciding What Concepts to Add.

acquisition can consult (Viegas and Raskin, 1997), (Viegas and Nirenburg, 1996). Briefly, lexicon acquirers (whether lexicographers or terminologists) have access to various on-line resources, such as **corpus search**, **look-up dictionary**, **ontology browser** tools. The acquisition is done semi-automatically; in other words we advocate human intervention to develop

the lexicons.

Apart from the tools to help acquiring the data, we have also developed programs to check the semi-automatically acquired data. Using this approach, we have acquired about one-fifth of our lexicon, and have developed a morpho-semantic acquisition program, which has allowed us to expand our core Spanish lexicon (containing 7,000 word senses) to 40,000 word senses, entirely automatically, through the application of lexical rules using a morpho-semantic generator (Viegas et al, 1996).

The acquisition of large-scale computational semantic lexicons is a very time consuming task, and the trade-offs between interlingua and lexical knowledge (or "language-impartial" versus "language-dependent" knowledge) on one hand and, between different senses within a lexicon, is not always easy to determine as we will see in next two sections.

4.3. ONTOLOGY-LEXICON TRADE OFFS

Word meanings in Mikrokosmos are represented in the lexicon, with meanings partly anchored in the ontology. The ontology has been built for NLP purposes and as such its acquisition involved continual trade-offs between the ontology and the lexicon.[13] (Mahesh and Nirenburg, 1996) argued for an intermediary position between a highly minimalist and a highly excessive number of primitives. The question of primitives is often left unspecified in NLP systems. Semantic lexicons are sometimes built by introducing a number of primitives as needed for representing word meanings. Neither the set of primitives, nor the taxonomic or other relationships between the primitives is specified in such systems. The best known example of an extremely minimalist position can be seen in Schank's (1973) Conceptual Dependency theory (CDs), an ontology of 11 events.[14] Such a small number of primitives is not practical for building large scale systems that attempt to capture the full richness of meaning that is necessary for MT or other NLP tasks in a domain. When we attempt to decompose complex events such as "a takeover bid for a company" in CDs, the resulting meaning representations will be lengthy, convoluted, and hard to acquire on a large scale. Moreover, they are unsuitable for MT since it is very hard to generate the equivalent word(s) in a target language from such highly decomposed meaning representations. The other extreme position, a popular one in NLP, makes almost every word sense in a natural language a primitive by itself. Examples of such "ontologies" are WordNet (Miller and Fellbaum, 1991) and

[13] Note that it could be possible to develop the ontology to support inferencing if needed.

[14] Other well-known minimalist approaches include Jackendoff's (1990) lexical-conceptual structures (LCS). See also (Dorr 1993), (Wilks, 1992) or (Onyshkevych and Nirenburg, 1994) for criticisms of the minimalist approach to meaning representation.

Sensus/Pangloss (Knight and Luk, 1994). In this approach, the large set of primitives is necessarily tied to a particular language (English in the above systems), which may be desirable for MT if the target language is always the same (say English).

We take an intermediate approach and propose a set of primitives that is much bigger than CDs or LCSs but significantly smaller than the typical size (of the order of 50,000) of a "word sense taxonomy" such as WordNet (Miller, 1990). Our experience in Mikrokosmos and its predecessor projects shows that fewer than 10,000 primitives are sufficient for building practical MT systems in a nontrivial domain, such as company mergers and acquisitions.[15] For successful, multilingual MT, such a system must be provided with a rich compositional structure in its meaning representations. The 6000-8000 primitives must also be organised in a highly interconnected ontological network. Using such a scheme, we have built a Spanish core lexicon with over 7000 words that use less than 2500 primitives in their meaning representations.[16]

A smaller ontology is not only cheaper to acquire, but we can also ensure better quality of concepts and inter-conceptual relations when the size is small. However, a smaller number of concepts necessitates a greater degree of decomposition in meanings in representing word senses in the lexicon. This not only explodes the cost of training and lexical acquisition, it also creates problems in analysis and generation. In Mikrokosmos, we have strived to achieve an intermediate grain size of meaning representation in both the lexicon and the ontology. Many word senses have direct mappings to concepts in the ontology; many others must be decomposed and mapped indirectly through composition and modification of ontological concepts. In a multilingual situation, the set of primitives must not be anchored in any natural language. A more compositional meaning representation with a smaller number of primitives is much more practical for constructing large-scale semantic lexicons for multiple languages.

For example, the English word *acquire* can be considered to be just the beginning phase of the word *own* so that the two words can be mapped to the same event in the ontology. However, it is often desirable to add separate concepts for the two word senses even from a strictly ontological perspective. For example, in the domain of company mergers and acquisitions, there may be a need to further classify the concept ACQUIRE into types of acquisitions. This cannot be done if teh word *acquire* was mapped

[15] It must be noted that we do not propose to encode only those meanings of words that are in the chosen domain. We in fact encode all meanings of words in a corpus using much less than 10,000 primitives. This number excludes any onomasticon entries needed.

[16] Construction of a Japanese lexicon using the same set of primitives has also begun recently.

to OWN in the ontology with no separate concept for acquisitions. Similarly, certain attributes or relations in the ontology may be applicable only to events involving transfers of possession and hence cannot be applied to the word *own*. There is no algorithm to determine when to decompose a word meaning and when to make it a new concept in the ontology. We have developed a set of guidelines and a training methodology that results in acceptable quality and uniformity in lexical and ontological representations.

Heuristics: In principle, the separation between ontology and lexicon is as follows: "language-neutral" meanings are stored in the former; language-specific information in the latter.

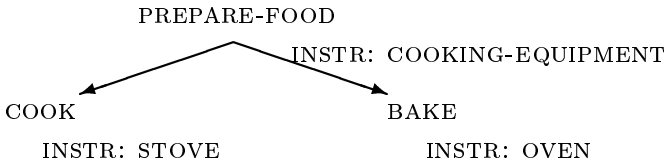
In a multilingual situation, as we saw earlier, it is not always easy to determine the boundary between ontology and lexicon. As a result, ontology and lexicon acquisition involves a process of daily negotiations between the two teams of acquirers.

For instance, if we consider the English verbs *cook* and *bake*, which translate "approximately" into the French equivalents *cuire [+/- sur le feu]* (cook on the fire) and *cuire [+/- au four]* (cook in the oven) respectively, then there are three different ways of doing the mappings, depending on the information available in the ontology:

a One-to-one Mapping between Ontology and Lexicon
b Lexicon Unspecification
c Lexicon Ontology Balance

One-to-one Mapping between Ontology and Lexicon For instance, one could consider that we have two concepts BAKE and COOK, subtypes of PREPARE-FOOD as in Table 3 to which correspond the English verbs *bake* and *cook* and the French expressions *cuire [+/- sur le feu]* and *cuire [+/- au four]* respectively. This solution seems more artificial for French than for English, because the verb *cuire* by itself is not ambiguous in French, it is just underspecified with respect to English.[17]

[17] The issue of actually translating *bake* and *cook* into French involves more than representational issues as discussed in (Viegas, 1997) and (Beale and Viegas, 1996).



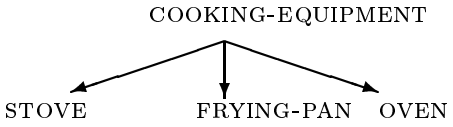
Concepts	COOK	BAKE
English words	*cook*	*bake*
French words	*cuire [+/- sur le feu]*	*cuire [+/- au four]*

TABLE 3. Lexicon Ontology Trade-offs in a Multilingual Environment: one-to-one mapping

This solution is the easiest, so it seems, for lexicographers who do not have to worry about changing the ontological constraints inside the lexicon. Here, the distinctions will be made in the ontology, with COOK and BAKE having STOVE and OVEN as their respective instruments. There are two problems with this solution though, lexical and ontological. First, we do not necessarily want a one-to-one mapping between concepts and lexemes, or in other words, we do not consider that every lexeme in every language constitutes a primitive in the ontology, as we discussed earlier. Second, such a representation introduces an unnecessary computational ambiguity in the French lexicon, as *cuire* by itself is not ambiguous, it is just underspecified with respect to English for the type of cooking involved such as (*cook, bake*).

Lexicon Unspecification Another solution would be to map all the entries to just one concept, labeled PREPARE-FOOD as in Table 4 for instance with different constraints on the INSTRUMENT slot of the concept, as shown below:

PREPARE-FOOD
INSTR: COOKING-EQUIPMENT



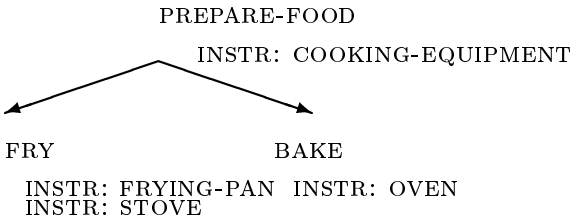
Concepts	PREPARE-FOOD	PREPARE-FOOD
+ instrument		OVEN
English words	*cook*	*bake*
French words	*cuire [+/- sur le feu]*	*cuire [+/- au four]*

TABLE 4. Lexicon Ontology Trade-offs in a Multilingual Environment: Lexicon Unspecification

Note that in this case we do not constrain the instrument of the English *cook* to be of type STOVE, as one can also *cook* with a barbecue for instance. Also note that STOVE and OVEN are subtypes of COOKING-EQUIPMENT.

Although it is possible to have underspecified entries such as in Table 4, it may be desirable to have BAKE as a separate event in the ontology. Without such an event, it is not possible in the ontology to add constraints on instruments of baking, for instance, due to the limited expressiveness of ontological representation in Mikrokosmos, a highly desirable feature from the point of view of ontology acquisition (Mahesh, 1996). As noted earlier, the BAKE node in the ontology is necessary also if there is a need to further classify baking events.

Lexicon-Ontology Balance



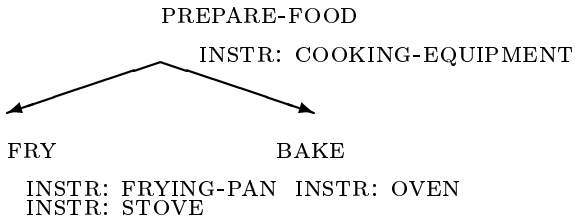
Concepts	PREPARE-FOOD	BAKE
+ instrument	COOKING-EQUIPMENT	
English words	*cook*	*bake*
French words	*cuire*	

TABLE 5. Lexicon Ontology Trade-offs in a Multilingual Environment: Lexicon Ontology Balance

Table 5 presents the best solution from the lexicon-ontology trade-off perspective as it limits unnecessary decomposition in the lexicon, as would have been the case otherwise in Table 4. It also allows the treatment of language gaps, by lessening the gap between the two types of mismatch and divergence for *cuire* [+/- au four], (Viegas, 1997), which was not the case in Table 3. Finally, it acknowledges the fact that the lexical items *fry, bake* are more specific than the lexical item *cook*, and that *cuire* by itself is only underspecified with respect to other languages.

4.4. RECONCILING ACQUIRERS' VIEWPOINTS

We have discussed so far the lexicon-ontology trade-offs, we now turn to the choices a lexicon acquirer has to face when creating a new lexicon entry. In the case of the lexical items discussed here, *cook, bake*, and the French *cuire*, and assuming the ontological knowledge illustrated below,[18]



the lexicon acquirer is faced with the two distinct choices:

[18] In reality and as pointed out earlier, there is a continuous interaction between the developers of ontologies and the developers of lexicons; for expository purposes, we will assume however the ontological information as shown in the last conceptual diagram.

1 the type of ontological mapping, which can be either direct or constrained
2 the number of entries to create per lexical item

Our experience shows that different acquirers, who have been trained, will arrive at the same number of entries and/or mapping. What might vary is how vague or underspecified an entry can be. This is due mostly because of the presence of a rich well structured and organised ontology,[19] which supports the lexicon.

In the following paragraphs, we discuss the two types of mappings between ontology and lexicon and then we address the issue of semantic ambiguity within a lexicon.

Direct Mapping In a direct mapping (Table 6), the lexicon acquirer, just maps the lexical item to a concept in the ontology, checking that all the slots and constraints on these slots fits the lexical item. In our case, this is the solution adopted for English (as described formerly in Table 4), where we have the following mappings:

Concepts	PREPARE-FOOD	BAKE
English words	*cook*	*bake*

TABLE 6. Direct Mapping

Constrained Mapping In a constrained mapping (Table 7), the lexicon acquirer maps the lexical item to a concept where the values of the slots have to be constrained. For instance, in our example, we could constrain the INSTRUMENT of PREPARE-FOOD to be STOVE in the lexicon entry for COOK if *cook* in English was always used with such a restriction, which is actually not the case as you can also cook on a barbecue or grill.

Number of Entries Another concern in lexical acquisition is the number of entries to create. If we look at data from a corpus, without relying on some underlying semantics, then we could come up with as many as the entries in Table 8.

Table 9 presents our solution where we wrote one sense per mapping and other meanings (such as the BAKE meaning for *cook in the oven*) are

[19] On theoretical discussions about ontologies, see (Nirenburg et al., 1994).

Concepts	PREPARE-FOOD
Instruments	STOVE
English words	*cook*

TABLE 7. Constrained Mapping

Words	**Concepts**	**Examples**
cook-V1	PREPARE-FOOD	*Cook the meat rare*
cook-V2	PREPARE-FOOD INSTRUMENT: OVEN	*Cook the pasta au gratin for 35 min. in the oven*
bake-V1	BAKE	*Bake the pasta au gratin 35 min.*
cuire-V1	PREPARE-FOOD	*Cuis les pates al'dente* (cook the pasta al'dente)
cuire-V2	BAKE	*Cuis le gâteau 45 min.* (bake the cake 45 mns)
cuire-V3	PREPARE-FOOD INSTRUMENT: OVEN	*Cuis les pâtes à four moyen.* (bake the cake at medium heat)

TABLE 8. Entries for lexical items.

created at processing time, in the context of other constraints. Moreover, we use mechanisms such as generalisation and specialisation, as described in (Mahesh et al., 1997) to dynamically go from PREPARE-FOOD (along with contextual constraints on the slot fillers) to BAKE and vice versa, to account for all the examples given in Table 8 for *cuire*.

cook-V1	PREPARE-FOOD
bake-V1	BAKE
cuire-V1	PREPARE-FOOD

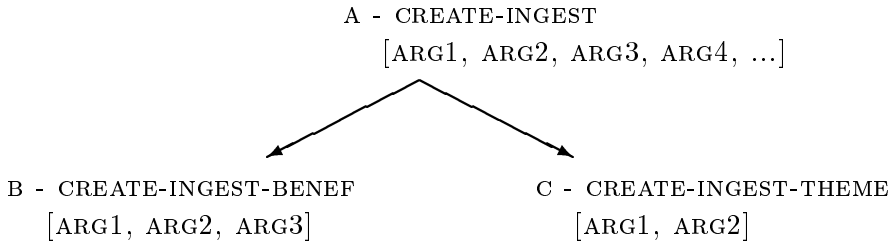
TABLE 9. Trade-off Output

Another issue we started addressing in Mikrokosmos is to take into

account the lexicon acquirers' points of view.[20] If we look at the following data and their subcategorisations:

(2) I fixed the meal - [NP1, NP2]
(3) I fixed a sandwich for you - [NP1, NP2, PP1]
(4) I fixed you a sandwich - [NP1, NP3, NP2]

where *fix* means PREPARE-FOOD, then our system must allow the analysis and generation of any of sentences (2), (3) and (4), whether we have a mapping of *fix* onto the concept A or on the frames B or C:[21]



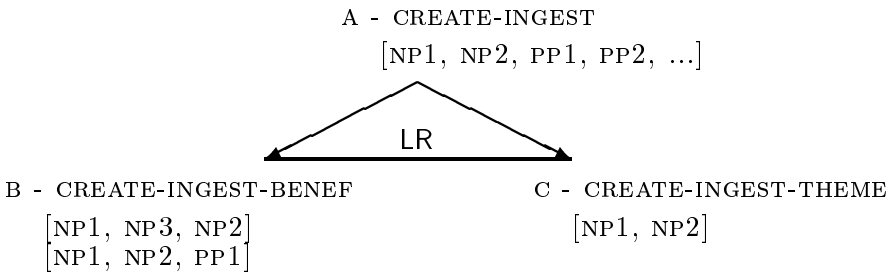
Looking at example 1), and in absence of examples 2) and 3) in the corpus, *fix* could easily be mapped into A or C, whereas with examples 2) and 3), and in absence of example 1) in the corpus, it could be easily mapped into A or B by the acquirers. We claim that this is of no importance as far as we have the mechanisms to go from one to the other. The diagram above is a computational linguist construct and has no "reality" per se; A, which is a concept in the ontology, also belongs to a truly multilingual hierarchy of semantic classes (Cahill and Gazdar, 1996) subsumed by the ontology.[22] B and C are constructs which provide for every semantic class the different semantic patterns that a particular semantic class accepts, such as the pattern CREATE-INGEST-BENEF requires 3 arguments and CREATE-INGEST-THEME 2. This diagram can be further specified for a particular natural language, where the required arguments are mapped to syntactic arguments and where lexical rules for a particular language provide the link between the different semantic patterns for a semantic class. The diagram

[20] This work is in progress and has not yet been implemented.

[21] We give the general diagram for CREATE-INGEST events, as PREPARE-FOOD is a subtype and will inherit all the properties of the semantic class CREATE-INGEST.

[22] The difference between the ontology and the multilingual hierarchy of semantic classes, is that there are less nodes in the latter; FRY, BAKE, GRILL are equivalent to the semantic class PREPARE-FOOD, and as such do not appear in the multilingual hierarchy.

below is for English where subcategorisations in between square brackets are associated to the lexical items mapped to A, B and C.



The corpus can indeed influence the way a lexicon acquirer will do the mapping, so if a lexicon acquirer creates an underspecified entry (mapping *fix* on (A), as opposed to (B) or (C)), dynamic mechanisms such as specialisation or generalisation (Kameyama et al., 1991) would enable the system to get to (B) and (C) from (A) and vice versa (to (A) from (B) or (C)). Moreover, if a lexicon acquirer decides to map to (B) instead of (C) or vice versa, then a lexical rule (LR) between (B) and (C) will enable the system to go from (B) to (C) and vice versa.

In other words, although there are 3 potentially different ways of writing the lexicon entry for *fix* for example sentences (2), (3) and (4), these different ways of encoding *fix* should remain a virtual difference at processing time. In other words, the system must encode mechanisms and rules to reconcile the different points of view of different acquirers, so that the system can treat sentences (2), (3) and (4) from any of the three potential lexicon entries.

In this section, we have outlined the needs for allowing lexicon entries to be dynamically changed to fit different linguistic contexts and different acquirers' analysis of the data. In next section, we compare our approach to the main trends in computational (lexical) semantics.

5. Comparison with Other Major Approaches

The comparison below is partial and incomplete: partial, because it is difficult to compare and do justice to all approaches as they differ in their goals; incomplete, because we highlight from other theories only some of the points addressed in this paper.

The main difference between Mikrokosmos and other approaches is that in Mikrokosmos the lexicon is only one part of the program of studying

semantics. Other approaches to computational semantics concentrate on issues connected with economy of lexicon acquisition based on exploiting lexical regularities, such as, for instance, derivational morphology (e.g., Viegas et al, 1996) or standard metonymies (such as, e.g., *creator* for *creation*, as in *Rachmaninov played Chopin*), (cf. (Ostler and Atkins, 1992), (Briscoe and Copestake, 1991), (Copestake and Briscoe, 1992), (Briscoe et al., 1993), (Pustejovsky, 1995)). The main research goal of workers in this approach is to eliminate the need for specifying some word senses altogether and instead have rules for deriving the necessary word senses automatically from the main (and, often, only) sense, which most of the time consists of an underspecified meaning.[23] From the point of view of a processing application, the question is rather when to apply the sense expansion rules. For instance, if the derivational morphology-oriented rules are applied at load time, then the resulting system lexicon includes word senses. If it is done "when needed," the only real difference is in the trade-off between space (to store the expanded system lexicon) against time (to apply lexical rules at need time). An important point to be made is that lexical rules constitute a powerful conceptual tool to generate the semantics of a word on the fly as illustrated in (Viegas et al., 1996). Our approach is different in that we discuss the nature and content of the semantic knowledge sources only in the context of the discussion of the mechanisms of extraction, representation and use of meaning.

We claim that a sense enumeration approach is neither too restrictive nor totally inadequate or inflexible to handle sense disambiguation even for new uses of words in novel contexts. It is indeed true that it is impossible to secure the exhaustive list of meanings for every single word, or complex expression (for instance what about the treatment of metonymies or the whole range of metaphors), as it is true that there is no theory of context available to help constructing meanings on the fly. Our position is to take advantage of most of the information listed in lexicons and look at processing methods to dynamically find new meanings which were not listed in a lexicon entry.

There is not yet a single semantics-based operational system which can avoid some form of enumeration. Whether *dejar* in Spanish has 7 senses, as in (Nirenburg et al., 1994), or the 54 meanings listed in the Collins Spanish-English dictionary; whether *break* has 7 meanings in (Palmer and Wu, 1995) or the 18 in LDOCE, or a single underspecified meaning, as claimed in (Pustejovsky, 1995), should be determined by the existing organisation of an ontology or conceptual domain or lexical conceptual paradigms. Meaning underspecification as an alternative to a sense enumeration approach

[23] See (Kees van Deemter and Stanley Peters, 1996) for computational treatments on underspecification.

is a very appealing idea, although until we are able to operationally define *contextual meaning* it will remain but an idea. Saturating some underspecified lexical representation by context is still today a very difficult task.

The advantage of underspecified lexicon entries, is that it obviously limits sense enumeration plus avoids spurious ambiguity. The problems with such an approach are worth mentioning. It becomes very difficult to expand a lexicon through lexical rules (which constitute a powerful conceptual tool for rapid acquisition); it requires the availability of a theory of context for computational semantics (not yet available); finally, the underspecified entries seem very far away from the linguistic data (thus complexifying the task of acquisition). Lexical semantic underspecification is still today difficult to consider for NLP applications, not because it is the worng way to go, on the contrary, but because as of today, we still lack the mechanisms to produce the specified meaning of a word in context. In that sense, we believe that underspecification of lexical items should not constitute a goal in itself for NLP apllications, but rather we can consider it when we have the available mechanisms to "surf" the hierarchy to go from one underspecified entry to a fully specified one in context.

Finally, another main difference in our approach to computational lexical semantics is that we try to reconcile computational linguistic theoretical concerns (in terms of understanding the combinatory and productive principles involved in the analysis and generation of natural languages) with NLP concerns (developing efficient non-toy working systems). From this perspective, one should mention that a computational (lexical)-semantic approach to NLP, although feasible (as demonstrated in Section 2), would be considered an overkill, due to high costs in acquisition, for applications such as Information Retrieval, where a statistics based approach would be more cost-effective. The future of computational (lexical)-semantics for NLP applications will rely on the ability to create knowledge at a much lower cost.

6. Conclusion

In this paper, we presented our approach to computational (lexical) semantics, focusing on the dynamics of semantics. From our viewpoint, we must discuss the nature and content of the semantic knowledge sources only in the context of the discussion of the mechanisms of extraction, representation and use of meaning.

We discussed in this context the lexicon ontology trade-offs, or in other words, the language related versus language neutral knowledge. In the case of EVENTs realised as verbs, we advocated a truly semantic approach, with an event hierarchy which can help predict the semantic behaviour of a verb,

rather than a syntactically-driven classification of verbs, as is usually done.

We stressed that the debate on sense enumeration should remain a legitimate lexicographers' concern, but is not the central concern of theories and applications of semantic processing.

Finally, we want to emphasise that in terms of knowledge acquisition, once we have developed the core lexicon for a natural language, more attention and work should be devoted to the entirely automatic generation of (lexical) semantic data at run time, to overcome some virtual incompleteness due to lexicon acquirer's points of views.

Further research in computational lexical semantics includes investigating how to bypass the closed world assumption under which knowledge-base systems work, i.e. under the simplistic assumption that the knowledge sources are complete.

References

Beale, S., Nirenburg, S. and K., Mahesh (1995). Semantic Analysis in the Mikrokosmos Machine Translation Project. In Proceedings of the *Second Symposium on Natural Language Processing* (SNLP-95), August 2-4. Bangkok, Thailand.

Beale, S. (1997). *HUNTER-GATHERER: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics.* Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University.

Beale, S., S. Nirenburg and K., Mahesh (1996) HUNTER-GATHERER: Three Search Techniques Integrated for Natural Language Semantics. In the *Thirteenth National Conference on Artificial Intelligence (AAAI96)*, Portland, Oregon.

Beale, S. and E., Viegas (1996) Intelligent Planning meets Intelligent Planners. In Proceedings of the Workshop on *Gaps and Bridges: New Directions in Planning and Natural Language Generation, ECAI'96*, Budapest, August 12-13.

Stephen Beale, Evelyne Viegas and Sergei Nirenburg (1997) Breaking Down Barriers: The Mikrokosmos Generator. In Proceedings of Natural Language Processing Pacific Rim Symposium, Phuket, Thailand.

Bresnan, J. (Ed.) (1982) *The Mental Representation of Grammatical Relations.* Cambridge MA: MIT Press.

Briscoe, T. and A., Copestake (1991) Sense extensions as lexical rules. In *Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language.* Sydney, Australia, pp. 12-20.

Briscoe, T., de Paiva, V., and A., Copestake (eds.) (1993) *Inheritance, Defaults, and the Lexicon.* Cambridge: Cambridge University Press.

Briscoe, T., Copestake, A. and A., Lascarides (1995) Blocking. In P. Saint-Dizier and E. Viegas (Eds.), *Computational Lexical Semantics.* Cambridge University Press.

Cahill, L. and G., Gazdar (1996). " Multilingual Lexicons for Related Lexicons", In Proceedings of *AISB'96 Workshop on Multilinguality in the Lexicon*, Brighton, UK, April 1-2, 1996.

Copestake, A. and T., Briscoe (1992) Lexical operations in a unification-based framework In J. Pustejovsky and S. Bergler (eds), *Lexical Semantics and Knowledge Representation.* Berlin: Springer, pp. 101-119.

Cruse, D.A. (1986) *Lexical Semantics.* Cambridge Texbooks in Linguistics.

Dorr, B.J. (1995) A lexical-semantic solution to the divergence problem in machine translation. In St-Dizier P. and E., Viegas (eds), Computational Lexical Semantics: CUP.

Dorr, B., and J., Klavans (eds.) 1994/1995. Special Issue: Building Lexicons for Machine

Translation I. *Machine Translation 9: 3-4.*
Dorr, B., and J., Klavans (eds.) 1995. Special Issue: Building Lexicons for Machine Translation II. *Machine Translation 10: 1-2.*
Farwell, D., S. Helmreich, W. Jin, M. Casper, J. Hargrave, H. Molina-Salgado and F. Weng. PANGLYZER: Spanish Language Analysis System. In Proceedings of *the 1st Conference of the AMTA*, p.56-64, 1994.
Fillmore, C.J. (1985) Frames and the Semantics of Understanding. *Quaderni de Semantica*, VI.2.
Fillmore, C.J. (1993) Frame Semantics and Perception Verbs, contribution to the Daghstuhl Seminar on Universals in the Lexicon, in: Hans Kamp, James Pustejovsky (eds.), *Universals in the Lexicon: At the Intersection of Lexical Semantic Theories*, 1993, ms, Daghstuhl.
Freuder, E.C. (1978) Synthesizing Constraint Expressions. in *Communications ACM* 21(11): 958-966.
Grimes, J.E. (1968) *The Thread of Discourse.* ERIC Document ED-019669.
Gross, M. (1984) Lexicon-Grammar and the Syntactic Analysis of French. In *Proceedings of the 10th Coling*, Stanford, Ca.
Jackendoff, R. (1990) *Semantic Structures.* Cambridge, MA: MIT Press.
Kameyama M., Ochitani, R. and S., Peters (1991) Resolving Translation Mismatches With Information Flow. In Proceedings of the *Association for Computational Linguists, 1991.*
Knight, K. and S.K., Luk (1994). Building a Large-Scale Knowledge Base for Machine Translation. In Proceedings of the *Twelfth National Conf. on Artificial Intelligence*, (AAAI-94).
Lenat, D., M. Prakash and M. Shepherd (1986) CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine, VI.*
Lenat, D. B. and R.V., Guha (1990). *Building Large Knowledge-Based Systems.* Reading, MA: Addison-Wesley.
Levin, B. (1993) *English Verb Classes and Alternations.* The University of Chicago Press.
Luger, G.F. and W.A., Stubblefield (1992) Rule based expert systems. In *Computer Engineering Handbook*, C.H. Chen, ed. New York, NY: McGraw-Hill.
Mahesh, K. (1996) *Ontology Development for Machine Translation: Ideology and Methodology.* Memoranda in Computer and Cognitive Science, MCCS-96-292. Las Cruces, N.M.: New Mexico State University.
Mahesh, K., and S. Nirenburg (1995). A situated ontology for practical NLP. In the Proceedings of *IJCAI'95* Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, August 19-21.
Mahesh, K., Nirenburg, S., Cowie, J. and D. Farewell (1996). *An Assessment of CyC for Natural Language.* Memoranda in Computer and Cognitive Science, MCCS-96-302. Las Cruces, N.M.: New Mexico State University.
Mahesh, K., Nirenburg, S. and S. Beale (1997) If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation. In Proceedings of *TMI-97.*
McNaught, J. (1990) Reusability of Lexical and Terminological Resources; Steps towards Independence.
Meyer, I., Onyshkevych, B., and L. Carlson (1990). *Lexicographic principles and design for knowledge- based machine translation.* Technical Report CMU-CMT-90-118, Center for Machine Translation, Carnegie Melon University, Pittsburgh, PA.
Miller, G.A., and C. Fellbaum (1991). Semantic Networks of English. *Cognition 41*, pp. 197-229.
Minsky, M. (1968) *Semantic Information Processing.* MIT Press, Cambridge, MA.
Nirenburg, S., Raskin, V. and B., Onyshkevych (1994) Apologiae Ontologia. MT Summit'94.
Nirenburg, S., Mahesh, K. and S. Beale (1996) Measuring semantic coverage, In Proceedings of *Coling-96*, Copenhagen.
Nirenburg, S. and V. Raskin (1987) The Subworld Concept Lexicon and the Lexicon

Management System. In *Computational Linguistics*, Volume 13, Numbers 3-4.

Nirenburg, S. and V. Raskin (1996) *Ten Choices for Lexical Semantics. Memoranda in Computer and Cognitive Science*, MCCS-96-304. Las Cruces, N.M.: New Mexico State University.

Normier, B. and M. Nossin (1990) GENELEX Project: EUREKA for Linguistic Engineering. In *Proceedings of the international Workshop on Electronic Dictionaries*, OISO, Kanagawa, Japan.

Onyshkevych, B. and S. Nirenburg (1994). *The lexicon in the scheme of KBMT things.* Memoranda in Computer and Cognitive Science, MCCS-94-277. Las Cruces, N.M.: New Mexico State University. Reprinted as: A lexicon for knowledge-based machine translation, in: Dorr and Klavans 1995 (eds).

Onyshkevysh, B. (1997) *An Ontological-Semantic Framework for Text Analysis.* Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University.

Ostler, N. and B.T.S. Atkins (1992) Predictable meaning shift: Some linguistic properties of lexical implication rules. In J. Pustejovsky and S. Bergler (eds), *Lexical Semantics and Knowledge Representation.* Berlin: Springer, pp. 87-100.

Palmer, M. and Z., Wu (1995) Verb Semantics for English-Chinese Translation. *Machine Translation*, Volume 10, Nos 1-2.

Pollard, C. and I. Sag (1987) An Information-based Approach to Syntax and Semantics: Volume 1 Fundamentals. CSLI Lecture Notes 13, Stanford CA.

Pustejovsky, J., P., Anick (1988) On the Semantic Interpretation of Nominals. In *Coling 1988*, vol.2: 518-523.

Pustejovsky, J. (1995) The Generative Lexicon. Cambridge, MA: MIT Press.

Sanfilippo, A. (Ed.) (1992) *The (Other) Cambridge Acquilex Papers.* Technical Report No. 253. Computer Laboratory, University of Cambridge, New Museums Site, Pembrocke Street, Cambridge.

Schank, R. (1973) Identification of conceptualizations underlying natural language. In *Computer Models of Thought and Language*, Schank, R. and Colby, K., editors, San Francisco:W. H. Freeman Co.

Searle, J.R. (1969) *Speech Acts. An Essay in the Philosophy of Language.*. Cambridge University Press.

Tsang, E.P.K. (1993) *Foundations of Constraint Satisfaction.* London: Academic Press.

Viegas, E. and S. Beale (1996) Multilinguality and Reversibility in Computational Semantic Lexicons. In Proceedings of *INLG 96*, Sussex, England.

Viegas, E., Onyshkevych, B., Raskin, V., and S. Nirenburg (1996) From *Submit* to *Submitted* via *Submission*: on Lexical Rules in Large-scale Lexicon Acquisition. In Proceedings of *ACL96*, University of California, Santa-Cruz, June 23-28, California.

Viegas, E. and V. Raskin (1997) *Lexical Acquisition: Guidelines and Methodology.* Memoranda in Computer and Cognitive Science. MCCS-97-2xx. Las Cruces, Computing Research Laboratory, New Mexico State University.

Viegas, E. (1997) Mismatches and Divergences: the Continuum Perspective. In Proceedings of *TMI-97*, Santa Fe.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics.* pp. 454-460, Nantes, France.

Weinreich, U. (1964) Webster's Third: A Critique of its Semantics. In *International Journal of American Linguistics* 30: 405-409.

Wilks, Y. (1992) Review of Ray Jackendoff's *Semantic Structures, Computational Linguistics*, vol. 18:1, pp 95-97.

Wilks, Y. (1996). Homography and part-of-speech tagging. Presentation at the MikroKosmos Workshop. Las Cruces, N.M.: Computing Research Laboratory, August 21.