# Lexical Acquisition

Sergei Nirenburg, Victor Raskin[1] and Svetlana Sheremetyeva

Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
{sergei,raskin,lana}@crl.nmsu.edu

## Abstract

This paper describes several aspects of the process of lexical acquisition for one of the most comprehensive applications in natural language processing knowledge-based machine translation. Specifically, the paper concentrates on those components of the lexicon that centrally relate to the task of processing meaning. The usability of the comprehensive ontological semantic lexicon exceeds NLP and extends to theoretical descriptive linguistics and its other applications.

Acquisition of lexical knowledge is a crucial component of building natural language processing applications. The requirements for lexical knowledge differ across different applications. Some of them require only a small amount of information. For example, a lexicon supporting a spelling checker must, at a minimum, only list all the possible word forms in a language. Some other applications require vast quantities of diverse kinds of data. For example, a comprehensive text analysis system may require information about word boundary determination (useful for compounding languages, such as Swedish, where the lexical entries would often match not complete words but parts of compound words); information about inflectional and derivational morphology, syntax, semantics and pragmatics of a lexical unit as well as possible connections among knowledge elements at these various levels.

In this document, we describe lexical acquisition for a semantic analyzer viewed as a component of a knowledge-based machine translation system. The lexicon whose construction we illustrate has been and still is under development at NMSU CRL. Work on this lexicon has been an integral part of the development of *ontological semantics*, a semantic theory for computational applications and a practical approach to automatic treatment of meaning in texts. Over the years, our work involved, on the one hand, the development of general ideas about lexicon structure and content and, on the other, development of actual semantic lexicons for English, Spanish and Chinese. The utility of these lexicons extends beyond natural language processing to the needs of language description proper. The lexicons are flexible as to the grain size of description, thus absorbing and actually implementing the idea of variable-depth semantics (Nirenburg and Raskin 1986, 2001).

## 1. The Lexicon in the Scheme of KBMT Things

In any natural language processing system, the lexicon supports the processes of analysis and generation of text or spoken language at all levels tokenization (that is, roughly, lexical segmentation), part-of-speech tagging and morphological analysis, proper-name recognition, syntactic,

1. Also of the Interdepartmental Program in Linguistics and its Natural Language Processing Laboratory, Purdue University, West Lafayette, IN 47907-1356.

semantic and discourse/pragmatic analysis; lexical selection, syntactic structure generation and morphological form generation.

The basic structure of ontological semantic lexicon is as follows (for a more complete structure see Onyshkevych and Nirenburg 1995, Zajac et al. 1997, or Nirenburg and Raskin 2001). The lexicon for a given language is a collection of *superentries* which are indexed by the citation form of the word (represented in the ORTH-FORM field in the lexicon formalism). A superentry includes all the lexemes which have the same base written form, regardless of syntactic category, pronunciation, or sense. In the examples below, lexemes ( entries ) inside a superentry have names which are formatted using the character + , followed by the citation form, followed by - and an indication of the syntactic category (e.g., **v**, **n**, **adj**) of the entry and its sense number. For example, **+eat-v2** in**troduces the entry for** the second verbal sense of *eat*.

Each lexicon entry is comprised of a number of zones corresponding to the various levels of lexical information. The zones containing information for use by analysis or generation components of an MT system are: CAT (syntactic category), ORTH (orthography abbreviations and variants), PHON (phonology), MORPH (morphological irregular forms, class or paradigm, and stem variants or principal parts ), SYN (syntactic features such as `attributive` for adjectives), SYN-STRUC (indication of sentence- or phrase-level syntactic dependency, centrally including subcategorization) and SEM-STRUC (lexical semantics, meaning representation). The following scheme, in a BNF-like notation, summarizes the basic lexicon structure.

<superentry> :=
ORTHOGRAPHIC-FORM: form
({syn-cat}: <lexeme> * ) *

```
<lexeme> :=
            CATEGORY: { syn-cat}
            ORTHOGRAPHY:
                        VARIANTS: "variants"*
                        ABBREVIATIONS: "abbs"*
            PHONOLOGY: "phonology"*
            MORPHOLOGY:
                        IRREGULAR-FORMS: ("form"
                                                { irreg-form-name} )*
                        PARADIGM: { paradigm-name}
                        STEM-VARIANTS: ("form" { variant-name} )*
            ANNOTATIONS:
                        DEFINITION: "definition in NL" *
                        EXAMPLES: "example"*
                        COMMENTS: "lexicographer comment"*
                        TIME-STAMP: date-of-entry lexicog-id
                        DATE-LAST-MODIFIED: date lexicog-id
                        CROSS-REFERENCES: lexeme *
            SYNTACTIC-FEATURES:
                        SYNTACTIC-CLASS: class
                        IDIOSYNCRATIC-FEATURES: (feature value)*
```

```
SYNTACTIC-STRUCTURE:
            SYNTACTIC-STRUCTURE-CLASS: class
            SYNTACTIC-STRUCTURE-LOCAL: fs-pattern
SEMANTIC-STRUCTURE:
            LEXICAL-MAPPING: lex-sem-specification
```

The following example illustrates the structure and content of the lexicon. The example shows not a complete superentry but just the first verbal sense of the English lexeme *eat*:

```
+eat-v1 :=
        CAT: V
        MORPH:
                    STEM-V: (ate v+past)
                         (eaten v+past-part)
        ANNO:
                    DEF:  ingest solid food through mouth
        SYN:
                    SYN-CLASS: (trans +)
        SYN-STRUC
                                ((root $var0)
                                 (subj ((root $var1)
                                      (cat n)))
                                 (obj  ((root $var2)
                                      (opt +)
                                      (cat n))))
        SEM-STRUC :
                    LEX-MAP:
                                (%ingest
                                  (agent (value ^$var1)
                                       (sem *animal))
                                  (theme (value ^$var2)
                                       (sem *ingestible)
                                       (relax-to *physical-object)))
```

The above states that the verb *eat* takes a subject and a direct object, that its meaning is represented as an instance of the ontological concept INGEST; that the agent of ingesting, which constitutes the meaning of the verb s subject, must be an animal while the entity that is ingested is most often edible, though sometimes it can be something else, as in The baby ate the whole front page of today s paper . The *ontology* is a model of the real world that provides the semantic metalanguage that can be used for specifying the meaning of lexical units in natural languages. The Mikrokosmos project indeed developed such an ontology (see, e.g., Mahesh and Nirenburg 1995; Mahesh 1996; Nirenburg and Raskin 2001) and has used it in lexical acquisition.

The basic stages in lexical acquisition are: a) determining what entities (superentries and entries, that is, lexemes and word senses) must be included in the lexicon; b) grouping such entities by similarity, to attain economies of size when an entire list of lexical units must be treated in the same manner; c) acquiring the lexical unit s syntactic properties; d) describing its meaning; e) describing how syntactic properties of the unit relate to its semantic properties (this process is called linking by theoretical linguists); and f) acquiring ancillary information about the lexi-

cal units definitions, annotations, morphological information, etc.

In what follows, we will describe some of the lexical acquisition procedures used over the years in the Mikrokosmos project at NMSU CRL and related applications.

## 2. General Principles of Semantic Acquisition

The general scenario for lexical acquisition may be presented as follows:

- Decide how many senses for every word must be included into a lexicon entry. Read the definitions of every word sense in a dictionary and try to merge as many senses as possible, so that a minimum number of senses remains.
- Describe the syntax of every sense of the word.
- Describe the semantics of every word sense by mapping it into an ontological concept and constraining if necessary the properties of this concept.
- Link syntactic and semantic properties of a word sense.

Syntactic and Semantic zones of the lexicon can be acquired in any order. Leaving temporarily aside the syntax and the more detailed instructions on acquisition of meaning, we will focus in this section on the general principles for the latter.

### 2.1 The Pre-Acquisition Effort

Before the start of the massive lexical acquisition stage, a master acquirer carries out the pre-acquisition phase. Three major operations are executed in this phase:

- the theoretical foundation of the effort, including the principled architecture of the lexicon, is consolidated and/or developed to make acquisition possible;
- the methodology for the effort is developed;
- the pilot acquisition is carried through.

This is done not necessarily in the order listed above. In fact, the methodology emerges only later, after quite a few pilot descriptions are created, evaluated, and revised. The theory, informing, of course, the initial descriptive efforts, may be revised in the process as well.

The master acquirer needs to possess and utilize:

- the ability to understand the most appropriate meaning of a lexical entity;
- the ability to formalize that meaning in the most economical way;
- the ability to maintain reasonable ontological parsimony, that is, the ability to express lexical meaning using a minimum of ontological (metalanguage) resources.

The determination of the appropriate meaning of a lexical entry or, for that matter, any language unit that has meaning is something that the native speaker is supposed to possess subconsciously and automatically. Explicating it, however, is something neither the native speaker nor a linguist

can do naturally and easily. It is often hard for an untrained researcher to separate the meaning proper from presuppositions, entailments, and other inferences, often of a probabilistic nature. Thus, for a lexical entry such as *marry* it is easy to let into the lexicon all kinds of information about love, sex, fidelity, common abodes, common property, children, etc. The meaning proper of the entry, however, includes only a legal procedure, recognized by the society in question, making, typically but not exclusively, one adult man and one adult woman into a family unit. The links between this proper meaning of an entry and the various inferences should be available separately from the lexicon, in the ontological model, though nothing, of course, formally prevents one from linking it directly to the lexical entry. We are discussing these matters in more detail in the section on semantic heuristics.

The second ability, to develop and to utilize optimal formalisms for representing meaning, comes primarily from experience, often of a negative nature. A master acquirer s effort should be deeply rooted in computational work so that every formal step be associated with a computational price tag. A major consideration here is that the emphasis in the formalism is not on simplicity or elegance but rather on procedural transparence. This is a concern that formal semantics, for instance, considers to be beyond its boundaries.

The third ability is related to the previous one in that it is based on descriptive parsimony. i.e., the strong preference in keeping with Hayes (1979) admonition to keep down the ratio of vocabulary sizes in a metalanguage and its object language for keeping the ontological metalanguage as constrained as possible. Numerous difficult decisions must be made whether to go with a potentially cumbersome representation of a sense within the existing ontology, on the one hand, or to revise the ontology by adding concepts to it, to make the representation easier and, often, more intuitively clear. The additions to ontology and the balance and tradeoffs between an ontology and a lexicon are discussed elsewhere (see Mahesh 1996 or Nirenburg and Raskin 2001), but any such decision should make sure that:

- the addition of a new concept or any other revision of the ontology conforms entirely to the theoretical and methodological principles of ontological acquisition underlying the ontology;
- the change does not cause lexicological problems elsewhere in the lexicon a powerful paradigmatic mechanism for a quick survey of all the related entries should be put in place, if only at the intuitive level (see the section on the mentalist approach to lexical acquisition).

What the master acquirer accomplishes in the pre-acquisition stage is twofold:

- the creation of the major types of representations of the SEM-STRUC zones for the lexical entries;
- the identification of the necessary resources, aids, and tools for the process of lexical acquisition, namely:
    - — lining up the available online resources;
    - — creating desiderata and requirements for new online resources.

Ideally, the results of the pre-acquisition stage are as follows:

- all the necessary templates for the representation of all the existing types of SEM-STRUC zones of the lexicon entry are made available for massive acquisition;
- all the tools and documentation are available for the rapid detection, selection, and modification of the appropriate template for each entry.

Practically, of course, the pre-acquistion stage goes only part of the way, hopefully, most of the way towards these goals. In our practice, it also made sense for a medium-level acquirer to be responsible for assigning senses to their correct types, thus leaving a rank-and-file acquirer only the task of using a pre-identified template for a particular entry a job that would be limited, roughly, to assigning the appropriate fillers to certain properties. It should be noted, however, that this simplified job still requires human intelligence and can be only semi-automated.

The division of labor in the ontological semantic system of lexical acquisition can, thus, be represented as follows:

| Stage of Acquisition | Personnel Level | Results |
|---|---|---|
| Pre-Acquisition | Master Acquirer | Entry |
| Acquisition | Medium-Level | Type assignment |
| Acquisition | Rank-and-File | Entries |
| Post-Acquisition | Master Acquirer | Checked and corrected entries |

In our experience, while it was not unusual for the master acquirer to spend up to 2 hours on a particularly difficult entry type, rank-and-file acquirers can produce up to 30 entries an hour after just a week or two of training.

### 2.2 The Paradigmatic Approach to Semantic Acquisition

The ontological semantic approach to lexicon acquisition is committed to complete coverage of material (see Nirenburg and Raskin 1996). In lexical semantics, this means that every sense of every lexical item should receive a lexical entry, i.e., should be acquired. For practical purposes, every often means every word in a corpus on which the system is based. The corpus-based approach to complete coverage has been our dominant approach.

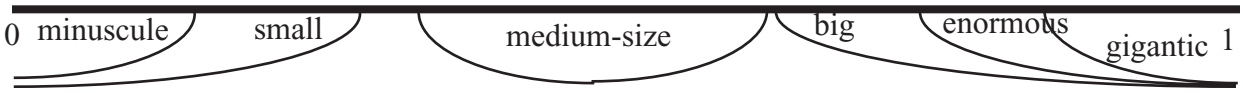
There is, however, an alternative interpretation of every as in every word in the language. This approach, worded like this, does not seem very practical or implementable. There is, however, a way to move towards this possibly unattainable and unnecessary goal quite rapidly and efficiently. We referred to this approach as rapid propagation elsewhere (see, for instance, Raskin and Nirenburg 1995). The linguistic principle on which it is based can be referred to as paradigmatic, or thesaurus-based.

As we have shown in the section on pre-acquisition, the master acquirer produces a set, as exhaustive as possible, of types of semantic representations to be used in large-scale lexicon acquisition. The moment a type is created, it is applied to all the lexical items for which it can be used. In the actual pre-acquisition pilot for Mikrokosmos, we concentrated on English adjectival semantics, so

our examples emerge from that effort (Raskin and Nirenburg 1995, 1998).

Let us start with a simple, typical case of the English adjective for the purposes of illustration. Such an adjective would be a scalar retaining its meaning in both the attributive and predicative use. Our microtheory of adjectival meaning (*op. cit.*) associates its meaning with a region on a numerical scale which is defined as the range of an ontological property. Ontologically, SIZE is a SCALAR-PHYSICAL-OBJECT-ATTRIBUTE, with the term scalar used here, as it is customarily, in the sense of gradable.

Each numerical scale can be measured in actual measuring units, such as LINEAR-SIZE in feet, yards, or millimeters, or TIME in seconds. But often natural language expressions do not refer to absolute magnitudes but rather to abstract relative ones, as in the case of *big*. We assume a 0 to 1 numerical range for such abstract scales. For abstract references to size, the fillers in English can be as follows:




*Big* will, then, get something like a > 0.75 value on the SIZE scale. Such values are a crucial part of the lexical mapping (LEX-MAP) included in the semantics (SEM-STRUC) zone of their lexical entries.

```
(big
        (big-Adj1
                (CAT  adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
                        (2  ((root $var0)
                                (cat adj)
                                (subj ((root $var1)
                                        (cat n))))))
                (SEM-STRUC
                        (LEX-MAP
                                ((1 2) (size-attribute
                                        (domain (value ^$var1)
                                                (sem physical-object))
                                        (range  (value (> 0.75))
                                                (relaxable-to (value (> 0.6)))))))))))
```

In the above example, there are two subcategorization patterns, marked 1 and 2, listed in SYN-STRUC. The former pattern corresponds to the attributive use of the two adjectives: the noun they modify is assigned the variable $var1, and the adjective itself the variable $var0 in the modifier position. The latter pattern presents the noun, bound to $var1, in the subject position and the adjective in the predicative position. In the SEM-STRUC zone, instead of variables which are bound to syntactic elements, the meanings of the elements referred to by these variables (and marked by a caret, ^ ) are used. Thus, ^$var1 reads as the meaning of the element to which the variable

$var1 is bound. Among the constraints listed in the SEM-STRUC zone of an entry, are selectional restrictions (the noun must be a physical object) and relaxation information, which is used for treatment of unexpected ( ill-formed ) input during processing.

Thus, an entry like the above should be read as follows:

- the first line is the head of the superentry for the adjective *big* (in our terminology, an entry is a specification of a single sense, while the superentry is the set of such entries);
- the second line assigns a sense number to the entry within its superentry;
- next, the adjective is assigned to its lexical category;
- the first subcategorization pattern in the SYN-STRUC zone describes the Adj-N construction; the second subcategorization pattern describes the N-Copula-Adj construction;
- the LEX-MAP part of the SEM-STRUC zone defines the lexical semantics of the adjective by assigning it to the class of SIZE adjectives; stating that it is applicable to physical objects and that its meaning is a high-value range on the SIZE scale/property.

In the case of continuous scales, like SIZE, the acquisition of adjectives for the lexicon is greatly facilitated, as all the adjectives of a class served by this scale need only an appropriate range assigned to them, all the rest of the information in the semantic part of the entry being the same. Thus, the entries for *enormous* and *tiny* differ from that for *big* in one value (and the absence of the RELAXABLE-TO facet see Onyshkevych and Nirenburg 1995):

```
((enormous
        (enormous-Adj1
                (CAT  adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
                        (2  ((root $var0)
                                (cat adj)
                                (subj ((root $var1)
                                        (cat n))))))
                (SEM-STRUC
                        (LEX-MAP
                                ((1 2) (size-attribute
                                        (domain (value ^$var1)
                                                (sem physical-object))
                                        (range  (value (> 0.9)))))))))))
((tiny
        (tiny-Adj1
                (CAT  adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
```

```
                (2   ((root $var0)
                        (cat adj)
                        (subj ((root $var1)
                                (cat n))))))
        (SEM-STRUC
                (LEX-MAP
                        ((1 2) (size-attribute
                                (domain (value ^$var1)
                                        (sem physical-object))
                                (range  (value (< 0.2)))))))))))
```

Thus, all the size related adjectives, a few dozen of them in our pilot corpus (*The Wall Street Journal* 1987-89) as well as a few more perhaps outside of the corpus, are easily covered by this general physical-size template. Note, however, that a slight variation of the template can be also used to account for many more adjectives. Thus, one sense of *fat* (see below), as in *fat man*, utilizes, essentially, the same template with a different scale, MASS, substituted for SIZE, and an appropriate SEM facet specified for ^$var1:

```
(fat
        (fat-Adj1
                (CAT   adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
                        (2   ((root $var0)
                                (cat adj)
                                (subj ((root $var1)
                                        (cat n))))))
                (SEM-STRUC
                        (LEX-MAP
                                ((1 2) (mass-attribute
                                        (domain (value ^$var1)
                                                (sem human))
                                        (range  (value (> 0.75))
                                                (relaxable-to (value (> 0.6)))))))))))
```

By varying the scales and the classes of modified nouns in the appropriate slots of the LEX-MAP, as illustrated in above, we have easily obtained the semantic represenations of many other types of adjectival senses based on numerical scales: quantity-related (e.g., *abundant, scarce, plentiful*), price-related (e.g., *affordable, cheap, expensive*), human-height-related (e.g., *tall, short, average-height*), human-mass-related (e.g., *fat, thin, emaciated, buxom, chubby*), container-volume-related (e.g., *capacious, tight, spacious*), and others to the total of 318 adjective senses, all acquired, basically, with one effort at the average rate of 18 entries per hour, including the several hours spent on the discovery and refinement of the template.

Polysemous or homonymous adjectives were handled one sense, not one word, at a time. Obviously, it is not really possible to delimit a single sense without a general view of what the other senses are, but this is part of a more general issue of how we capture lexical meaning (see the section on

semantic heuristics). Thus, for the notoriously difficult adjective *good* (see, for instance, Ziff 1960, Vendler 1963, Katz 1972 see also Raskin and Nirenburg 1995, Section 4), we start with its general, unspecified, and unrestricted evaluative meaning:

```
(good
        (good-Adj1
                (CAT   adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
                        (2   ((root $var0)
                                (cat adj)
                                (subj ((root $var1)
                                        (cat n))))))
                (SEM-STRUC
                        (LEX-MAP
                                (attitude
                                        (type              evaluative)
                                        (attitude-value (value (> 0.75))
                                                (relaxable-to (value (> 0.6))))
                                        (scope             ^$var1)
                                        (attributed-to   *speaker*))))))
```

This gives us several advantages. First, by taking care of *good*, we facilitate the acquisition of all adjectives whose meanings relate to the same evaluative scale, such as *bad*, *excellent*, *terrible*, *mediocre*, etc. Practically, it means the creation of yet another versatile template, which is copied for each new adjective of the same class (116 adjective senses in the pilot study). The entries for all the adjectives of this class will differ from each other again only in range values (see the entry for *excellent*) and constraints on the category of nouns which these adjectives modify (see the entry for *comfortable*).

```
(excellent
                (excellent-Adj1
                        (CAT   adj)
                        (SYN-STRUC
                                (1 ((root $var1)
                                        (cat n)
                                        (mods ((root $var0)))))
                                (2 ((root $var0)
                                         (cat adj)
                                         (subj ((root $var1)
                                                (cat n))))))
                        (SEM-STRUC
                                (LEX-MAP
                                        (attitude
                                                (type              evaluative)
                                                (attitude-value (value (> 0.9))
                                                        (relaxable-to (value (> 0.8))))
```

```
                                        (scope          ^$var1)
                                        (attributed-to  *speaker*))))))


(comfortable
        (comfortable-Adj1
                (CAT  adj)
                (SYN-STRUC
                        (1 ((root $var1)
                                (cat n)
                                (mods ((root $var0)))))
                        (2 ((root $var0)
                                (cat adj)
                                (subj ((root $var1)
                                        (cat n))))))
                (SEM-STRUC
                        (LEX-MAP
                                (^$var1
                                        (instance-of (sem (OR (furniture clothing))))
                                (attitude
                                        (type           evaluative)
                                        (attitude-value (value (> 0.75)
                                                (relaxable-to (value (> 0.6))))
                                        (scope          ^$var1)
                                        (attributed-to  *speaker*)))))))
```

Second, this approach maximizes the use of each type of lexical entry and, by the same token, of the ontological material it is based upon (ontological concepts, properties, facets, etc.) and thus contributes significantly to the parsimony of the ontology. Third, it makes uses of synonymy, antonymy, and other paradigmatic relations among words to generate lists of adjectives that can be acquired using a given lexical entry template. Availability of thesauri and similar online resources facilitates this method of acquisition.

Finally and perhaps most importantly, the judicious selection of an entry template and an ontological scale facilitates the acquisition of adjective entries across languages. The single word senses acquired in this fashion are, essentially, language-independent, as our work on Spanish and other languages has confirmed. In other words, each of the English senses will have an equivalent sense expressed in another language; what will vary from language to language is, essentially, how these single senses will be grouped in a superentry.

Obviously, this one-sense-at-a-time, maximum concept utilization approach has already yielded quite a few pairs, triples, etc., of senses of the same English adjective, but we assembling them into superentries comes much later in the overall acquisition process. Thus, *great-Adj1* is an evaluative sense of *great* and *great-Adj2* is its importance sense (note that the latter is treated as an attitude just like evaluative adjectives such as *good* and pretty much for the same reasons).

```
(great
        (great-Adj1
                (CAT  adj)
```

```
        (SYN-STRUC
                (1 ((root $var1)
                        (cat n)
                        (mods ((root $var0)))))
                (2   ((root $var0)
                        (cat adj)
                        (subj ((root $var1)
                                (cat n))))))
(SEM-STRUC
        (LEX-MAP
                (attitude
                        (type             evaluative)
                        (attitude-value (value (> 0.9))
                                (relaxable-to (value (> 0.8))))
                        (scope            ^$var1)
                        (attributed-to   *speaker*))))))
(great-Adj2
        (CAT   adj)
        (SYN-STRUC
                (1 ((root $var1)
                        (cat n)
                        (mods ((root $var0)))))
                (2   ((root $var0)
                        (cat adj)
                        (subj ((root $var1)
                                (cat n))))))
(SEM-STRUC
        (LEX-MAP
                (attitude
                        (type             salience)
                        (attitude-value (value (> 0.75)))
                        (scope            ^$var1)
                        (attributed-to   *speaker*))))))
```

**2.3 Semantic Heuristics**

In this section, we tackle the central issues of computational semantics and lexicography, namely, how to capture the meaning of an entry. This question includes at least these three aspects:

- how to determine what exactly the entry means;
- how to determine how detailed ( fine-grained ) the representation should be; and
- how best to represent the meaning given the ontological and lexical means we have.

We have basically two resources for capturing meaning, and their status is quite different: one of them, the speaker s intuition, works very well for humans but not for machines (it is difficult to represent it explicitly); the other, human-oriented published dictionaries, represents meaning

explicitly but are known to be faulty and unreliable and, moreover, do not contain sufficient amounts of information to allow for automatic capturing of word meaning from them. Dictionaries also typically list too many different senses. If a computational lexicon recognizes the same number of senses, it would be very difficult formally to specify how each of them differs from the others, and the human-oriented dictionaries do not always provide this information. Thus, in a computational application it becomes important to reduce the number of senses to a manageable set.

### 2.3.1 Reducing Polysemy

In his critique of Katz and Fodor (1963), Weinreich (1966) accused them of having no criteria for limiting polysemy, i.e., for determining when a sense should no longer be subdivided. Thus, for example, having determined that one of the senses of *eat* is ingest by mouth, should we subdivide this sense of *eat* into eating with a spoon and eating with a fork, which are rather different operations? Existing human-oriented dictionaries still do not have theoretically sound criteria for limiting polysemy of the sort Weinreich talked about. They definitely have their own implicit rules of thumb, but the results come out quite uneven. More importantly, the number of senses listed for each entry is usually quite high.

It is often difficult to reduce the number of senses for a word even in a computationally-informed lexical resource, as can be illustrated by an example from WordNet, a most useful and consistent online lexical resource (Miller et al. 1988; Fellbaum 1998). In WordNet, each sense in an entry is determined by a synset, a set of synonyms, rather than by a verbal definition. The list below contains the 12 synsets WordNet lists for the adjective *good*:

Sense 1
good (vs. evil) -- (morally admirable)
=> angelic, angelical, saintly, sainted -- (resembling an angel or saint in goodness)
=> beneficent, benevolent, gracious -- (doing or producing good)
=> white -- ( white magic )
Also See-> good, moral, right, righteous, virtuous, worthy

Sense 2
good (vs. bad) -- (having positive qualities, asp. those desirable in a thing specified: good news ; a good report card ; a good joke ; a good exterior paint ; a good secretary )
=> bang-up, bully, cool, corking, cracking, dandy, great, keen, neat, nifty, not bad(predicate), peachy, swell, smashing -- ((informal) very good)
=> fine -- (very good of its kind or for its purpose: a fine gentleman ; a fine mind ; a fine speech ; a fine day )
=> redeeming(prenominal), saving(prenominal) -- (offsetting some fault or defect: redeeming feature ; saving grace )
=> safe, sound -- ( a good investment )
=> satisfactory -- (meeting requirements: good qualifications for the job )
=> suitable -- (serving the desired purpose: Is this a good dress for the office? )
=> unspoiled -- ( the meat is still good )
=> well-behaved -- ( when she was good she was very good )
Also See-> best, better, favorable, genuine, good, obedient, respectable, sound, well(predicate)

Sense 3
benevolent (vs. malevolent), good -- (having, showing, or arising from a desire to promote the welfare or happiness of others)
=> beneficent, charitable, generous, kind -- ( a benevolent contributor )
=> good-hearted, kindly, openhearted -- ( a benevolent smile ; take a kindly interest )
Also See-> beneficent, benefic, charitable, kind

Sense 4
good, upright, virtuous -- (of moral excellence: a genuinely good person ; an upright and respectable man ; the life of the nation is secure only while the nation is honest, truthful, and virtuous - Frederick Douglass; the...prayer of a righteous man availeth much - James 5:16)
=> righteous (vs. unrighteous)

Sense 5
estimable, good, honorable, respectable -- ( all reputable companies give guarantees ; ruined the family's good name )
=> reputable (vs. disreputable)

Sense 6
good, right, seasonable, timely, well-timed -- (occurring at a fitting time: opportune moment ; a good time to plant tomatoes ; the right time to act ; seasonable summer storms ; timely warning ; the book's publication was well-timed )
=> opportune (vs. inopportune)

Sense 7
good, pleasing -- (agreeable or pleasant: we had a nice time ; a nice day ; nice manners )
=> nice (vs. nasty)

Sense 8
good, intact -- (not impaired in any way: I still have one good leg )
=> unimpaired (vs. impaired) -- (not damaged or diminished)

Sense 9
good -- (not forged: a good dollar bill )
=> genuine (vs. counterfeit)

Sense 10
good -- ( good taste )
=> discriminating (vs. undiscriminating)

Sense 11
good, Sunday, Sunday-go-to-meeting(prenominal) -- (used of clothing: my good clothes ; his best suit ; her Sunday-go-to-meeting clothes )
=> best (vs. worst) -- (superlative of good : the best film of the year )

Sense 12
full, good -- ( gives full (good) measure ; a good mile from here )
=> ample (vs. meager) -- (more than enough in size or scope or capacity)

The first thing one notices about the 12 senses is that noun classes which they modify vary a great deal in size. Sense 2 dwarfs all the other senses in this respect. Senses 1 and 3-5 all pertain to humans and their acts and are very similar to each other, at least in the sense that the association of one of these senses with a noun strongly entails or presupposes the association of the others with the same noun. Thus, in the examples below, *good* will probably be understood in a somewhat vague combination of all these four senses, and the native speaker will not feel the need for further specification:

Fred is a good man.
Fred s behavior in that difficult situation was very good.
Mom & Pop, Inc. is a good company

This feeling, if captured reliably by a well-defined procedure, is the basis Weinreich sought for determining if further polysemy is required. A group of individuals, if defined as *good*, is indeed more likely to be understood in WordNet Sense 5, but none of the other three is excluded. In fact, other than in a context of at least several sentences, if not paragraphs, it is very hard to use *good* specifically in one of these similar senses and not simultaneously in the others. This observation can serve as an operational criterion for limiting polysemy: if it is hard to pinpoint a sense within a one-sentence example, the status of the meaning as a separate sense in the lexical entry should be questioned.

It is crucially important that if there is a shift in meaning at all from across the above examples, it is caused by the shift from one person to a group of individuals. The determining role of the nominal meaning on the meaning of *good* is even more obvious in the other WordNet senses for the adjective. Starting with Sense 6, the noun classes to which these senses apply shrink in size, and with Senses 8-12 come dangerously close to phrasals consisting of *good* and the corresponding nouns, phrasals in which the meaning of *good* varies significantly.

WordNet itself recognizes some of the observations above by reducing the 12 senses of *good* in to the following three senses in response to a different set of parameters:

Sense 1
good (vs. evil) -- (morally admirable)
    => good, virtue, goodness -- (the quality of being morally excellent or admirable)

Sense 2
good (vs. bad) -- (having positive qualities, esp. those desirable in a thing specified: good news ; a good report card ; a good joke ; a good exterior paint ; a good secretary )
    => goodness -- (being of positive value)

Sense 3
benevolent (vs. malevolent), good -- (having, showing, or arising from a desire to promote the welfare or happiness of others)
    => benevolence -- (an inclination to do kind or charitable acts)

This short list of the main senses of *good* is still rather unbalanced with respect to the size of noun classes they modify, and the distinction between Senses 1 and 3 remains perhaps only slightly less problematic than the distinction among Senses 1 and 3-5. It is, however, the long WordNet list rather than the short one that is very similar to typical dictionary fare: compare the entries from the

online Webster s (1963) and the American Heritage Dictionary (1992) we list only meaning-related information in both entries.

(Webster s) 1. good... 1a1: of a favorable character or tendency {~ news} 1a2: BOUNTIFUL, FERTILE {~ land} 1a3: COMELY, ATTRACTIVE {~ looks} 1b1: SUITABLE, FIT {~ to eat} 1b2: SOUND, WHOLE {one ~ arm} 1b3: not depreciated {bad money drives out ~} 1b4: commercially reliable {~ risk} 1b5: certain to last or live {~ for another year} 1b6: certain to pay or contribute {~ for a hundred dollars} 1b7: certain to elicit a specified result {always ~ for a laugh} 1c1: AGREEABLE, PLEASANT 1c2: SALUTARY, WHOLESOME {~ for a cold} 1d1: CONSIDERABLE, AMPLE {~ margin} 1d2: FULL {~ measure} 1e1: WELL-FOUNDED, COGENT {~ reasons} 1e2: TRUE {holds ~ for society at large} 1e3: ACTUALIZED, REAL {made ~ his promises} 1e4: RECOGNIZED, HONORED {in ~ standing} 1e5: legally valid or effectual {~ title} 1f1: ADEQUATE, SATISFACTORY {~ care} 1f2: conforming to a standard {~ English} 1f3: DISCRIMINATING, CHOICE {~ taste} 1f4: containing less fat and being less tender than higher grades - used of meat and esp. of beef 2a1: COMMENDIBLE, VIRTUOUS, JUST {~ man} 2a2: RIGHT {~ conduct} 2a3: KIND, BENEVOLENT {~ intentions} 2b: UPPER-CLASS {~ family} 2c: COMPETENT, SKILLFUL {~ doctor} 2d: LOYAL {~ party man} {~ Catholic}: in effect: VIRTUALLY {as good as dead}: VERY, ENTIRELY {was good and mad}

(American Heritage good
1. Being positive or desirable in nature; not bad or poor: a good experience; good news from the hospital.
2.a. Having the qualities that are desirable or distinguishing in a particular thing: a good exterior paint; a good joke. b. Serving the desired purpose or end; suitable: Is this a good dress for the party?
3.a. Not spoiled or ruined: The milk is still good. b. In excellent condition; sound: a good tooth.
4.a. Superior to the average; satisfactory: a good student. b. Used formerly to refer to the U.S. Government grade of meat higher than standard and lower than choice.
5.a. Of high quality: good books. b. Discriminating: good taste.
6. Worthy of respect; honorable: ruined the family's good name.
7. Attractive; handsome: good looks.
8. Beneficial to health; salutary: a good night's rest.
9. Competent; skilled: a good machinist.
10. Complete; thorough: a good workout.
11.a. Reliable; sure: a good investment. b. Valid or true: a good reason. c. Genuine; real: a good dollar bill.
12.a. In effect; operative: a warranty good for two years; a driver's license that is still good. b. Able to continue in a specified activity: I'm good for another round of golf.
13.a. Able to pay or contribute: Is she good for the money that you lent her? b. Able to elicit a specified reaction: He is always good for a laugh.
14.a. Ample; substantial: a good income. b. Bountiful: a good table.
15. Full: It is a good mile from here.
16.a. Pleasant; enjoyable: had a good time at the party. b. Propitious; favorable: good weather; a good omen.
17.a. Of moral excellence; upright: a good person. b. Benevolent; kind: a good soul; a good heart. c. Loyal; staunch: a good Republican.
18.a. Well-behaved; obedient: a good child. b. Socially correct; proper: good manners.
19. Sports. Having landed within bounds or within a particular area of a court: The first serve was wide, but the second was good.
20. Used to form exclamatory phrases expressing surprise or dismay: Good heavens! Good grief!

The high quality of WordNet as an online resource and of its (largely unspecified) method of distinguishing senses is confirmed by the fact that the other two dictionaries do not really add any significant new senses to the WordNet long list--all they do is further specify the senses and add

new phrasal and near-phrasal senses.

Our two main objections to adding a sense to an entry are as follows:

- the lack of clear distinction between the candidate sense and those already in the entry, and
- the small size of the set of nouns to which this sense applies.

The rules of thumb to be used by lexicon acquirers for reducing polysemy are based on the above observations:

- Try to make the word carry the candidate sense in a one-sentence example. If you need to provide additional context for this to happen, this sense should be rejected and subsumed by one of the existing senses in the entry.
- Check if the candidate sense applies only to a small number of semantically similar nouns. If it does, reject this sense; its meaning will either be subsumed by one of the existing senses in the entry or will become a part of the meaning of an idiom.

Both these rules are manifestations of a general linguistic principle of complementary distribution, widely used for establishing variance and invariance in phonology and morphology: if two different senses of the same adjective can only be realized when used with two different nouns or sets of nouns, they should be seen as variants of the same sense. In a way, some dictionaries try to capture this in their entries by grouping all senses into a small number of main ones which are further divided, often iteratively. Thus, Webster s has only two main senses and two levels of specification under them, but American Heritage prefers putting 20 senses on the top level, with minimum further subdivision. Both from the point of view of theoretical linguistics (the essential complementary distribution principle) and of natural language processing, entries like that in Webster s are the least helpful.

The objections to the entry in American Heritage push us in an obvious direction: we see *good* as having one sense, which takes different shades, depending on the meaning of the modified nouns. This sense of *good* is something like assigning a high positive value range to a selected property of the noun. Our entry for *good* (see above) captures this meaning but refuses to specify the noun property, and we have a good reason for doing that. *Good* is, of course, an adjective with a very broadly applicable meaning, but the same objections to excessive polysemy hold for other adjectives as well. The same principle of polysemy reduction pertains to other lexical categories: thus, in Nirenburg *et al* (1995), we reduced 52 listed senses for the Spanish verb *dejar* to a manageable set of just 7.

### 2.3.2 Grain Size

Reducing the number of senses in a polysemous lexical item affects the grain size of its semantic representation: the fewer the number the larger the grain size. We are interested in formulating principles of granularity which will allow us to keep the number of senses in an entry low.

#### 2.3.2.1 Grain Size on a Principled Basis: Practical Effability

The grain size of lexical representation in an NLP system must be principle-based and application-oriented

The overt principles that guided the acquisition of the Mikrokosmos lexicon entries for adjectives have been presented above. It is certain that more principles were used in this process, without the acquirers realizing this. However, they still await their formulation due to the usual difficulties with explicating intuitions. In fact, we have imposed yet another principle which we expect to remove in later work. For reasons of simplicity of acquisition, we have strived to maintain the grain size of description at a uniform level of detail throughout the system. In fact, however, variable-depth meaning descriptions (see Nirenburg and Raskin 1986) are necessary when one wants to balance economy of effort and ability to disambiguate.

It has been demonstrated in various schools of semantics that natural languages dissect reality differently and fill their words with different chunks of that reality (see, for instance, Hjelmslev 1959, Whorf 1956, Hayakawa 1975). It has also been experienced by translators who know that word-for-word rendition of a text from one language to another never makes sense. This is the reason for crucial semantic difficulties in MT, necessitating the kind of meaning analysis we intend to support by ontological-semantic lexicons.

The interlingual text meaning representation resulting from such meaning analysis presupposes lexicons for the source and target languages which represent enough different word and phrase senses to give serious credence to a hope that a meaning expressed in one language will be expressible in another language at the same grain size without much difficulty in most cases. There are, however, cases when this will not happen, and it is those cases that require a finer grain size of semantic analysis than the others. As a result, the most general approach will expect variable grain-size meaning descriptions.

One such case would be a situation when one word in a source language can be translated into a target language as either one of two words, and the decision as to which word to use requires additional information that the source text may not contain at all or at least not in an easily extractable way. For example, the English *corner* can be rendered in Spanish as either *rinc n* (inside) corner, nook or as *esquina* (outside) corner, street corner ; the English *blue* can be rendered in Russian as either *siniy* dark blue, navy blue or *goluboy* light blue, baby (sky) blue . As a result, it is difficult to translate the sentences: *He could see the corner clearly* and *She wore a blue dress* into Spanish and Russian, respectively.

Refining the grain size for *corner* and *blue* in their lexical entries by adding to their lexicon definitions appropriate distinguishing properties in order to accommodate Spanish and Russian is possible though often practice useless. This is because the data on which lexical constraints can be checked may not be present in either the text or extralinguistic context. Such situations are difficult for human translators as well. The reason for this state of affairs can be that language always underdetermines reality (cf. Barwise and Perry 1983: 30): any sentence leaves out numerous details of the situation described in it, and in the case of the above examples, English underdetermines it more than Spanish or Russian.

Katz (1978: 209) has formulated the principle of effability, or mutual intertranslatability of natural languages: [e]ach proposition can be expressed by some sentence in any natural language (see

also Katz 1972/1974: 18-24, Frege 1963: 1, Tarski 1956: 19-21, and Searle 1969: 19-21 a view which is opposite to that formulated by Quine 1960: 26-30). In our work, we directly rely on a stronger form of this principle. Its generic formulation, expressed in the terms of the philosophical debate on effability, is as follows:

*Hypothesis of Practical Effability*: Each sentence can be translated into another natural language on the basis of a lexicon with equally limited polysemy.

A version more attuned to the environment of computational microtheories can be formulated as:

*Hypothesis of Practical Effability for Computational Microtheories*: Any text in the source language can be translated into the target language in an acceptable way on the basis of a lexicon for the source language and a lexicon for the target language, such that their respective entries are limited in exactly the same fashion with regard to polysemy.

The equally limited polysemy we practice recommends fewer than 10, preferably fewer than 5 senses per lexical entry. The limitation does not, of course, affect the scope of the word meaning: all the possible senses of a lexical item are captured in the superentry. The small number of these senses simply means a larger grain size. In a limited domain, however, some senses of the same word can be ignored because they denote concepts which are not used in the domain, are not part of the sublanguage that serves the domain, and thus are unlikely to occur in the corresponding corpora (see Nirenburg and Raskin 1987; Raskin 1971, 1987).

### 2.3.2.2 . Grain Size in Entries

The practical effability hypothesis was successfully applied to a corpus of English with 1,506 adjective senses. Let us see how exactly it is reflected in the choices forming the lexical entries. The adjective *good* is, again, a good place to start. We will show how, for this adjective, we settled on a grain size of description larger than the most detailed semantic analysis possible. We will then see how the same principle of not specifying in detail the specific noun property modified by an adjective applies to all the other adjectives as well. And we will briefly discuss the conceptual and computational status of those properties which are introduced by the scales we need to postulate for our adjective entries.

We interpret *good* in a sentence like *This is a good book* as, essentially, *The speaker evaluates this book highly*. We realize that in this sentence *good* may have a large variety of senses, some of which are illustrated in the possible continuations of the sentence:

- ...because it is very informative.
- ...because it is very entertaining.
- ...because the style is great.
- ...because it looks great on the coffee table.
- ...because it is made very sturdy and will last for centuries.

Obviously, *good* may have additional senses when used to modify other nouns, as in *This is a good breadmaker*, *He is a good teacher*, *She is a good baby* or *Rice is good food*.

In each case, *good* selects a property of a noun and assigns it a high value on the evaluation scale associated with that property. The property changes not only from noun to noun but also within the same noun, depending on the context. The finest grain-size analysis requires that a certain property of the modified noun is contextually selected as the one on which the meaning of the noun and that of the adjective is connected. This is what many psychologists call a salient property. In our approach, the representation solution for *good* would be to introduce an evaluation attitude, with a high value and scoped over this property.

Now, it is difficult to identify salient properties formally, as is well known, for instance, in the scholarship on metaphor, where salience is the determining factor for the similarity dimension on which metaphors (and similes) are based (see, for instance, Black 1954-55, 1979; Davidson 1978; Lakoff and Johnson 1980, Lakoff 1987; Searle 1979; on salience, specifically, see Tversky and Kahnemann 1983). It is, therefore, wise to avoid having to search for the salient property, and the hypothesis of practical effability offers a justification for this. What this means, in plainer terms, is that if we treat the meaning of *good* unspecified with regard to the nominal property it modifies, there is a solid chance that there will be an adjective with a matching generalized, unspecified meaning like that in the target language as well.

In fact, however, we go one step further with the lexical entry of *good* and other adjectives from the same scale and remove their meaning from the nouns they modify, making them contribute instead to an evaluative attitude pertaining to the whole sentence. It can be argued, of course, that since the scope of the attitude remains the modified noun, all that changes is the formalism and not the essence of the matter. We do not wish to insist, therefore, that this additional step constitutes a step towards an even larger grain size.

Non-attitude-based scalars are treated in a standard fashion: their lexicon entries effectively execute the following, informally defined, procedure: insert the scale name and scale value for an adjective as a property-value pair in the frame describing the meaning of the noun the adjective modifies.

If *house,* in one of its senses, has the following lexicon entry

```
(house-N1
        (CAT   n)
        (SYN-STRUC
                (1        ((root $var0)
                           (cat n))))
        (SEM-STRUC
                (LEX-MAP
                        (2        (private-home))))
```

then the meanings of the phrases *big house* and *red house* will be represented in TMRs as follows:

```
(private-home
        (size-attribute (value > 0.75)))

(private-home
```

(color-attribute (value red)))

In the former example, the linking attribute is selected rather high in the hierarchy of attributes in the ontology SIZE-ATTRIBUTE is the parent of such properties as LENGTH-ATTRIBUTE, WIDTH-ATTRIBUTE, AREA-ATTRIBUTE, WEIGHT-ATTRIBUTE, etc. If the context does not allow the analyzer to select one of those, a coarser-grain solution is preferred. In other words, we represent the meaning of *big house* without specifying whether *big* pertains to the length, width, height or area of a house.

The advantage of a larger grain size is, of course, increased feasibility and manageability of lexical description and analysis, thanks, for instance, to much more limited polysemy. But there is a price to pay for this position: if a translation depends on the availability of a finer-grain polysemy, the translation will fail. We are obviously gambling on the infrequency of such situations and on the applicability of the practical effability principle in many cases.

### 2.3.3 What Does This Word Mean?

In this subsection, we discuss two related but distinct issues, namely, how a lexicon acquirer can discover what an adjective means and how to decide how to represent this meaning. Much more effort has been spent in the field on the latter question, though the former is a much more difficult issue. We intend to demonstrate that the difficulties of determining what the meaning actually is are often not appreciated by researchers.

#### 2.3.3.1 Representation of Meaning

Our commitment to the ontological representation in the lexicon helps us to determine the actual representation of a lexical entry but it does not make it a deterministic process: there are still choices to make and, accordingly, principled bases for making these choices.

One good example of such a choice and a theoretical basis for making it is our treatment of the adjective *abhorrent*. In general, this adjective is morphologically related to the verb *abhor*, and its lexical entry is derived from that of the verb. There are, however, at least two very distinct ways of representing *abhor*, one as an event and the other as an attitude. If we had an event concept LIKE, for instance, we would easily present *abhor* as an intensified negation of LIKE. Alternatively, we can represent *abhor* and *like* as an attitude: *like* is represented pretty much along the lines of *good* , and *abhor* simply replaces the $> 0.75$ value of *like* on the evaluation scale with something like $< 0.1$. Accordingly, either an event concept or an evaluative attitude appear in the LEX-MAP for *abhorrent*. Which should it be?

The answer is based on our independently motivated position with regard to the representation of verbs: we represent actions, but not states, as instances of ontological events. This disqualifies *abhor* and *like*, along with many other evaluative states from an event-based representation, and the adjective *abhorrent* gets an attitude-based treatment:

```
(abhorrent
    (abhorrent-Adj1
        (CAT    adj)
        (SYN-STRUC
            (1 ((root $var1)
                (cat n)
```

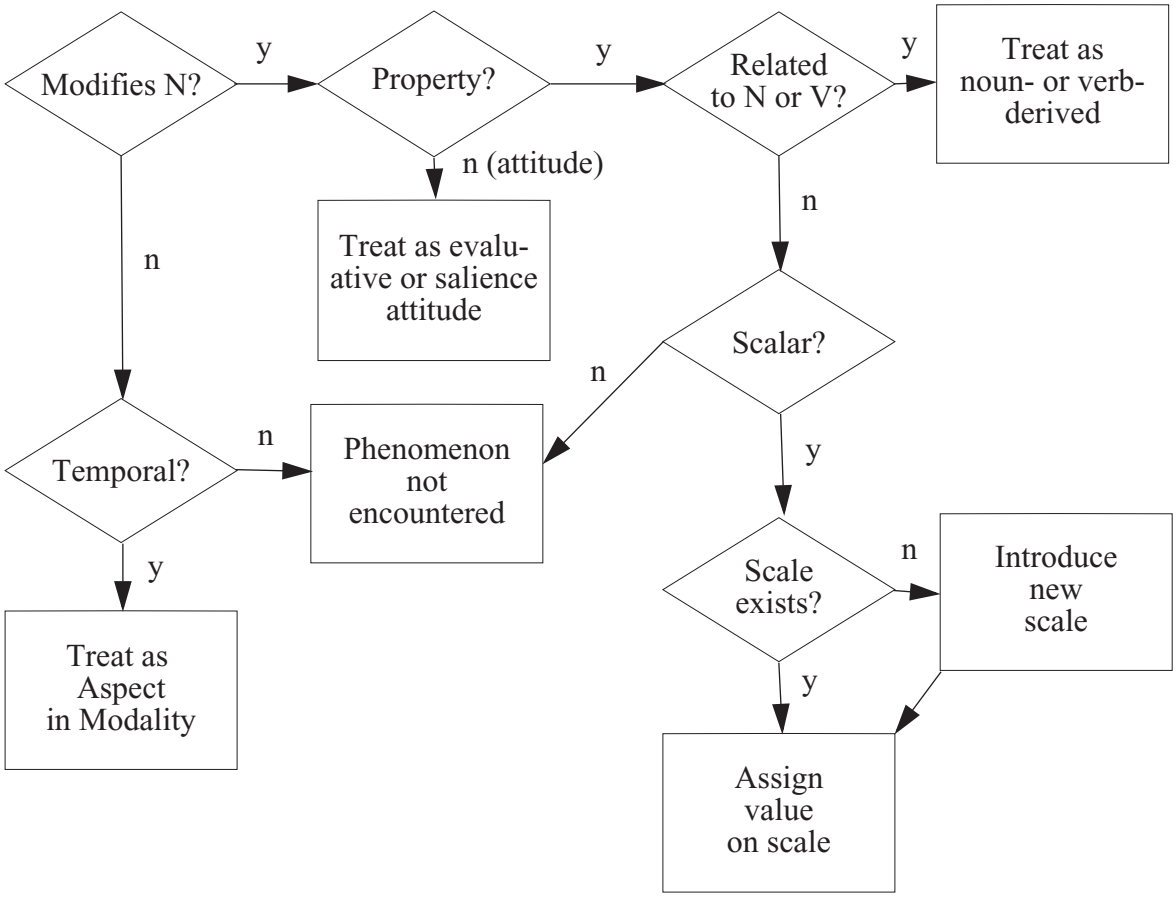
```
                (mods ((root $var0)))))
        (2   ((root $var0)
                (cat adj)
                (subj ((root $var1)
                        (cat n))))))
(SEM-STRUC
        (LEX-MAP
                (attitude
                        (type    evaluative)
                        (attitude-value (value (< 0.1)))
                        (scope           ^$var1)
                        (attributed-to   *speaker*))))))
```

Even if we decide on this form of representation for *abhorrent*, there are still choices to make. The same lexical entry can be derived from the entry for *abhor*, as we mentioned before, or directly acquired as an evaluative adjective. In the former case, *abhorrent* is treated as a member of the adjective set including such words as *lovable*, *likable*, *repulsive*, *repellent*, etc. In the latter case, it belongs to a set of adjectives, such as *awful*, *terrible*, *dismal*, *abysmal*, etc. The difference between the two sets is the fact that the former is associated with the morphologically related verbs and the latter is not.

Other choices may not be related to any other principle of representation or notation. In fact, they are addressed in traditional lexicography, *albeit* in its typically intuitive fashion, not directly related to any linguistic semantic insights. Thus, the twelve basic principles of lexicographic definition by Benson *et al*. (1986: 203-226) favor 1) referential definition by the same part of speech, with 2) a minimum of, if any, metalanguage, 3) the use of synonyms, and 4) illustrative phrases, 5) definitions... [that] give only those relevant features that are necessary to differentiate the referent from other members of its general class (op. cit.: 211), 6) the inclusion of lexical patterns, 7) the use of field, temporal, regional, etc. labelling, 8) a certain order of defined meanings, 9) a clear and separate treatment of connotation, 10) clarity of definitions, 11) consistency of definitions, and 12) the need for objectivity.

This is pretty standard fare in traditional lexicographic literature (cf., for instance, Svens n 1990: 112-139, or Landau 1984: 120-174; see also Benson 1985, Congleton *et al*. 1979, Hartmann 1983, Ilson 1986, Kipfer 1984, McDavid and Duckert 1973, Zgusta 1971 and 1980). Those of the twelve steps above that are readily interpretable describe a desirable state of lexicographic affairs but contain very few clues as to how they can be realized. The assumption is that an expert knows that intuitively. There is no criterion of evaluation and little methodology: whatever is there is of a negative nature, such as do not use a different part of speech in the definition, use as little metalanguage in the definition as possible, preferably none at all, or do not use any features other than the distinctive ones (cf. McCawley 1986, ridiculing this very principle on the example of a standard dictionary definition of *horse*). As to the positive methods of lexicographic description, for instance, how to identify relevant features, how to achieve clarity or objectivity, this literature does not provide answers, except by pointing out at some positive examples and sometimes comparing them with the less successful ones. Lexicography is treated as an art or a craft (Landau 1984), and this is, indeed, a self-fulfilling prophecy: without a reliable methodology and an explicit theoretical foundation, it is guaranteed not to become a science.

The following flow chart summarizes our methodology for creating lexical entries for adjectives:




Obviously, we need reasons for all the decisions represented by the diamonds in the flow chart, and for that, we need to determine what an adjective means as well as to have guidance for the actual choices. A good methodology offers such guidance in the form of tests, and this is precisely what we will attempt to formulate next.

#### 2.3.3.2 Determining Meaning and Making Choices

Our lexicographic resources include a variety of online dictionaries, tools for showing the actual usages of the word in context on a wide selection of corpora, and an arsenal of acquisition tools which can display useful templates, bring up the entries of similar words, and traverse our lexicon and ontology in other ways. Nevertheless, no tool has the capability of determining the meaning of the word automatically: our acquisition of the lexicon (and of the ontology) is semi-automatic in that it requires a human participant, even though human intuition is guided and checked through these tools.

For nouns and verbs, the simplest outcome is the discovery of an ontological concept which directly corresponds to their meanings. Failing that, the next step is an attempt to discover the concept for the hyperonym of the word: a positive outcome will result in either adding a new daughter concept in the ontology into which the word would be mapped directly, or adding meaning constraints

to the entry for this word in the lexicon, while linking it to the concept for its hyperonym.

With adjectives, however, we will not find the corresponding concept in the ontology unless the adjective pertains to a noun or a verb and the LEX-MAP of the adjective is to be derived from that of the corresponding noun or verb. In this case, the adjectival meaning can be determined only after the meaning of the noun or verb is specified.

Before pursuing the particular path of the determination of the adjectival meaning, let us pause to consider what resources we can use in the determination of meaning in general. The dictionary entries and/or the speaker s intuition give us the general idea. The contexts in which the word occurs allows us to remove the word, replace it with an antonym, synonym, or other cognate word and see what happens to the meaning of the sentence. Obviously, if the entry being acquired is synonymous to a previously acquired entry, the problem is solved even though there are no identical synonyms in languages, and the distinctions between two synonyms, represented by their diagnostic constructions, i.e., contexts in which one of them fits and the other does not (see, for instance, Raskin and Weiser 1987: 115), should be checked out; we may, in fact, have identical synonyms in the selected grain-size: in that case the set of diagnostic constructions will be conveniently empty.

Let us focus, however, on the most difficult and interesting case, when we do not get any easy escapes. If we find no previously acquired close meaning cognates and we cannot quite figure out how to bend a dictionary entry into our framework, we should focus on the contexts in which the word occurs. Bloomfield s unfriendly advice would be to look for the features brought by the word into each situation in which it is uttered, focusing on the distinctive features (see Bloomfield 1935: 139; cf. Alston 1964: 26-28). Now, Bloomfield introduced this definition only to declare meaning unknowable. Alston is not too hopeful about this approach either.

If, while working on the word *shirt*, the very one that baffled Alston, in its garment sense, we attempt an IS-A hypothesis, linguistics offers a standard test to confirm or falsify such a hypothesis: we do it by negating it. If the resulting negative sentence makes no sense, the semantic element hypothesized about is indeed part of the meaning of the word (cf. Raskin 1986: 53-54). For example, the following two sentences form the test pair:

¥ A shirt is a garment.

¥ *I have a shirt but I have no garment.

If, on the other hand, the negation of a hypothesis is meaningful, the semantic element hypothesized about is not part of the meaning of the word:

¥ Shirts have collars.

¥ I have a shirt which has no collar.

This test, which we call the deniability filter, weeds out presuppositions, entailments, and inferences from the meaning of the word proper. An adjective like *round* can be (successfully) described using the property of shape or (unsuccessfully) the property of ability to roll:

¥ Roundness is a shape

¥ *I saw a round object, which had no particular shape.

¥ Roundness means the ability to roll.

¥ I saw a round object, but it could not roll.

At this point, we are ready to see how the algorithm presented in the flow chart above helps the process of acquisition. The first decision to make is whether the adjective semantically modifies the noun it modifies syntactically. The decision is made much easier for us by the finding that all those adjectives which definitely do not are of a temporal nature (see, for instance, example (40) above). Our framework assigns temporal information to events, and it properly belongs together with the aspect-related information in modality.

The next question to answer is whether the basic meaning of the adjective is an attitude or a property. We have two attitudes, evaluative and saliency, so all the evaluation-attribute adjectives, such as *good*, *bad*, *superb*, *awful*, etc., and the saliency-attribute adjectives, such as *important*, *unimportant*, *significant*, *prominent*, etc. belong here. All other adjectives are treated as properties.

We are approaching the most critical part of the procedure, but there is one easy question left: is the adjective a morphological derivative of a noun or a verb, such that the meaning of the adjective follows from the meaning of the underlying noun or verb. This is particularly easy to establish when the morphological derivation follows the standard N--->Adj or V--->Adj route, as in the examples below:

| | | | |
|---|---|---|---|
| ¥ | abusive | <--- | abuse |
| ¥ | national | <--- | nation |

If a noun and a verb of the same root may both claim an adjective, we give preference to the latter:

| | | | |
|---|---|---|---|
| ¥ | *successful | <--- | success |
| ¥ | successful | <--- | succeed |
| ¥ | success | <--- | succeed |

It is a little less trivial to relate an adjective to an underlying verb, for instance, suppletively, which can, of course be done only semantically:

audible <--- hear

The noun-derived adjectives are clearly demarcated from cases like

red ---> redness,

where the noun is obviously derived from the adjective. Such nouns are, in fact, treated as attributes semantically as well.

The LEX-MAPs of the adjectives truly derived from verbs or nouns (deverbals and denominals, respectively) can be created from the LEX-MAPs of the corresponding verbs or nouns with the help of

lexical rules.

The remaining adjectives are either scalars or members of property sets that do not form scales. The task we face is to assign each of them to an appropriate scale or value set. Obviously, with each new adjective, we must decide if it fits into an existing scale or set or it requires the addition of a new one.

### 2.4 Sample Instructions in the Mikrokosmos Lexicon Acquisition Tool.

While in the above sections we described the various pre-acquisition steps and considerations, in what follows we concentrate on the steps that a rank-and-file acquirer must go through in order to acquire the core lexical information for a lexical unit. While the pre-acquisition steps were illustrated largely using adjectives, the illustration of the actual acquisition process concentrates on nouns and verbs. Also, while describing adjectival entries, we concentrated on their semantic properties. In the description of the nominal and verbal lexical entries, we will illustrate the operations devoted to acquisition of syntactic knowledge as well as linking, the information about co-reference of syntactic dependency roles for clause heads and other governing elements and semantic case roles (participants) for event types that express their meanings.

The information below is excerpted from the online *tutorial* supplied with the Mikrokosmos lexicon acquisition tools. We do not describe ontology acquisition in this article, even though acquisition of ontology and the lexicon is carried out in lockstep. For the purposes of this narrative, we assume that ontological support for lexical acquisition is already available.

Note that the presentation of parameters and definitions in what follows in many respects differs from the familiar pedagogical or theoretical approaches. It was carefully developed specifically for the purpose of facilitating ontological-semantic lexicon acquisition by people who may have no formal linguistic training. This special pedagogical angle requires the introduction of some redundancy in the descriptions, as the acquirers are expected to be somewhat slow in digesting the definitions and concepts with which they operate.

#### 2.4.1 Acquiring Syntactic Information

The first step in syntactic description of nouns is to understand with what kind of a noun sense you are dealing. We divide all nouns (or noun senses) into simple nouns and complex nouns. The rules of thumb (which may not work in 100% of cases) to help determine whether a noun is simple or complex are as follows:

Simple nouns are those nouns that

- are not derived from verbs (non-deverbal nouns) and represent objects or events such as *table* (object), *weather* (event), *athletics* (event), but NOT properties. For example, *age*, *size*, *weight* are not simple nouns, as they are properties of things.
- can be derived from verbs but in this case they should represent objects, not events or properties. For example, the word sense *management*-N2 ( those in charge of running a business ) is a simple noun because, though it is derived from the verb *manage* it represents an object. Compare with *management*-N1 ( the act of managing something ) which is a

deverbal noun; semantically, an event, not an object; and thus not a simple noun.

Complex nouns are all the other nouns, i.e., nouns derived from verbs (deverbal nouns) that represent events, such as *movement* (from *move*) or non-deverbal nouns which represent properties of things such as height (of smth), etc.

For simple nouns you skip the syntactic stage of description (the necessary markers are added automatically) and, after correcting the definition of the sense (if necessary), proceed directly to the semantic description of the word sense. Complex nouns require syntactic description.

### 2.4.2 Syntactic description of complex nouns

Complex nouns are nouns derived from verbs (deverbal nouns) such as *movement* (from *move*) or non-deverbal nouns which represent properties of things such as *height* (of smth.), *weight* (of smth), etc.

Syntactic description of complex nouns is done in terms of their *subcategorization patterns*. A subcategorization pattern for a word is a set of phrases which may be required to co-occur with this word in a sentence. For example, the subcategorization pattern of the word *movement* may optionally include three prepositional phrases as in: *movement of the ball from the window to the door*. In the formalism we currently use, this subcategorization pattern can be represented as

N [optional]Adjunct1(PP preposition:of)
[optional]Adjunct2(PP preposition:from)
[optional]Adjunct3(PP preposition:to).

The interface will create this record automatically. Your task is only to select appropriate Adjuncts (components of a subcategorazation pattern) from the interface menu.

A word can have more than one subcategorizarion pattern. For example, *announcement* subcategorizes for *of*, *about* or *that*.

The interface will allow you to choose several subcategorization patterns for a word sense if necessary. For *announcement* the two patterns are as follows:
N [optional]Adjunct1(PP(preposition:of, about) and N [optional]Complement1(S) .

### 2.4.3 Syntactic Description of Verbs

Syntactic description of verbs centrally involves specifying their *subcategorization patterns*. A subcategorization pattern of a verb is a set of phrases that are required by this verb to form a grammatically correct sentence. For example, the verb *place*-V1 ( to put smth in a particular location ) requires *he* and *the* chair for the sentence *He placed the chair by the window very quickly* to be correct and complete. Indeed, omitting these phrases leads to unacceptable results:

**Placed the chair by the window very quickly*.
*He placed by the window very quickly.
The phrase *by the window* is an optional member of the subcategorization pattern of the verb place-V1. Indeed, while the sentence *He placed the chair* can be accepted as appropriate, it is still

understood by the reader that the chair must be put somewhere. The phrase *by the window* is licensed by this sense of *place*. By contrast, the phrase *very quickly* can be skipped without affecting the sense of the verb *place*.

In general, as a rule of thumb, adjuncts denoting time and location of events are not included in the subcategorization pattern, though when the event directly refers to time or location, this rule does not hold. For example, in *He waited for four hours* the time adjunct is a member of the subcategorization pattern.

Thus the subcategorization pattern of the verb *place*-V1 includes the following components:

NP1 V NP2 (optional)PP1, where, for the example sentence, NP1 = *he*, NP2 = *the chair*, PP1 = *by the window.*

Subcategorization patterns often include the main syntactic functions of the pattern components, such as subject, direct object and indirect object. They may also include some morphological markers, for example, case. In the formalism we currently use, the subcategorization pattern of the above example is represented as

```
V Subject(NP case:Nominative)
   DirectObject(NP case:Accusative)
   [optional]Adjunct1(PP preposition:by)
```

where the syntactic function Adjunct is assinged to a component other than subject, direct or indirect object or a clause. (the latter can also be a component of a subcategorization pattern. It is called a complement in our formalism). The editor will create this record automatically. Your task is only to select appropriate components of a subcategorazation pattern from the menu.
A verb can have more than one subcategorization pattern. The interface will allow you to choose as many subcategorization patterns as necessary.

#### 2.4.3.1 Subject

The subject typically refers to the agent of an action, its perpetrator, as it were, or to general topics of conversation. In English, the subject can be a noun phrase, pronoun, finite or nonfinite clause (a clause headed by an infinitive or gerund):

*The dog* ate my homework.
*He* is tough.
*A large tip* is out of the question.
*How you got there* doesn't concern me.
*To read* is useful.
*Reading* is interesting.

In English, all sentences require subjects, even if those subjects do not have a lexical meaning, as in *It is raining*.
However, some languages permit subjectless sentences. There are two types of subjectless sentences:

¥ those in which the subject can be left out (or ellided) if it is implied by the context. For

example, in Russian you can say both *Ponimaju* ( I understand) and *Ja ponimaju* ( I understand ), and

- ¥ those that can never take a subject For example, Russian *Temneet* ( It is getting dark ).

A technical point: In languages which use grammatical cases, the subject is usually put either in the nominative case or in the ergative or absolutive case.

#### 2.4.3.2 Objects

Objects are nouns, pronouns, or noun phrases that are either recipients of a verbal action, or complements of (i.e., the thing that goes with) a preposition or postposition. There are two types of objects:

- ¥ Direct Objects, which are the direct recipients of verbal action (and are generally in the objective or accusative case, in languages that have case), e.g.,: I hit *him*; He loves *me*; The dog ate *the bone*; Grandma baked *a cake*, and
- ¥ Indirect Objects, which are indirect recipients of the action: I gave the gift *to her*. He handed the book *to me*.

#### 2.4.3.3 Examples of subcategorization patterns

| Categories of verbs | Subcategorization patterns | Probable case-roles | Examples |
|---|---|---|---|
| Intransitive<br><br>One obligatory argument | NP-V | theme / agent / experiencer | He suddenly appeared (at the wedding). |
| Transitive<br><br>Two obligatory arguments | NP-V-NP<br>NP-V-(optional)NP<br><br>NP-V-(clause)Compl<br>NP-V-(Inf/-ing)XCompl | agent, theme | I only smelt the roses.<br>She already ate (the meal).<br><br>We learned that people thought.<br>I like to read books.<br>I like reading books. |
| Bi-transitive<br><br>Three obligatory arguments | NP-V-NP-NP<br>NP-V-NP-PP<br><br>NP-V-NP-(clause)Compl<br>NP-V-NP-(Inf)XCompl | agent, experiencer, theme | I gave Mary the roses.<br>I told Mary (that) I would come .<br>I persuaded Mary to go. |

### 2.4.4 Semantic Description of Nouns

Semantic description of the sense of a word is done in terms of ontological concepts and their properties. The main task is to find the ontological concept that best represents the sense of the word. We will say that the sense of a word should be *mapped* into an ontological concept.

Below are some rules of thimb about how to find an appropriate ontological concept for the given sense of a word.

First of all, read the definition of a noun sense to decide into what kind of ontological concept it should be mapped. Nouns can map into any class of concepts: OBJECTS, EVENTS or PROPERTYs. If a noun sense corresponds to an OBJECT concept its meaning implies persistence in time. If a noun sense maps into an EVENT concept its meaning often implies causation. For example, *flood* is an EVENT, not an OBJECT because it is caused by something else and is not persistent in time (compared, for example, with *river*). Some nouns, for example, *cost* (of smth), *weight* (of smth), *decline* (of smth), etc. map into PROPERTY concepts.

#### 2.4.4.1 How to map a word sense into an ontological concept

The first step in semantic description of a word sense is to find an ontological concept which best matches this sense. It is probably best to proceed as follows:

Try to find a concept whose definition coincides with the word sense on which you are working. For example, for the word sense: *table*-N1 (a piece of furniture), there is a concept TABLE (a flat horizontal surface with legs).

If such a concept is difficult to find determine whether your word sense represents an object, an event or a property and search along the corresponding branches of the ontology tree. For example, when describing the word sense of *accord*-N1 ( a written agreement between two or more parties ) you could feel that it is an object; scrolling through the branches of the OBJECT subtreein the ontology will lead you along the path: OBJECT > MENTAL-OBJECT > REPRESENTATIONAL-OBJECT > LANGUAGE-RELATED-OBJECT > DOCUMENT>CONTRACT.

At this point, both the concepts DOCUMENT ( anything printed or hand-written that is relied upon or used as proof of something ) and CONTRACT ( an agreement written up between people, especially one that can be enforced by law ) might seem to be appropriate for representing the semantics of *accord*-N1 . But CONTRACT does not really cover your sense, as *accord*-N1 is a kind of agreement not necessarily between two people. It might be an international treaty between countries, etc. So the right choice is the concept DOCUMENT.

**Important**. When the appropriate concept is found, check whether the definition of the concept covers the definition of your word sense. If not, search for another concept.

Further steps in semantic description depend upon the type of a concept (OBJECT, EVENT or PROPERTY) you found for your word sense. Mapping word senses into OBJECTS and EVENTS differs from mapping them into PROPERTIES.

#### 2.4.4.2 A word sense corresponds to an Object or Event. Direct and Constrained mapping

After you found that a noun sense corresponds to a particular OBJECT or EVENT, read the content of the DEFINITION and IS-A slots of the concept and decide whether you have a direct or a con-

strained mapping. A mapping is direct when the content of the concept presented in the DEFINITION and IS-A slots closely conforms to the word sense with which you are dealing. A mapping is constrained when the content of the concept is close but does not fully coincide with the sense of the word. In this case the procedure is to tighten or relax, as needed, the fillers of some properties (slots) of the concept to adjust it to the sense of the word you are working on.

In the case of direct mapping you should only put the concept in the semantic description of a word sense and proceed to the linking stage. In the case of constrained mapping in addition to the concept, you should also indicate those concept properties whose fillers you are changing to make them reflect your word sense. For example, when mapping the word sense *accord*-N1 ( a written agreement between two or more parties ) into the concept DOCUMENT you should constrain the property OUTCOME-OF by changing its filler from EVENT (in the ontological concept) to SOCIAL-EVENT or DIPLOMATIC-EVENT (in the lexicon), to yield:

accord-N1 (a written agreement between two or more parties)
CONCEPT: DOCUMENT
PROPERTY: OUTCOME-OF
FACET: SEM
FILLER: SOCIAL-EVENT, DIPLOMATIC-EVENT

**Important**. You can only relax or tighten the fillers of the properties which are present in the concept frame. For example, when mapping the word sense chief-N2 (a leader; ruler; person with highest rank) into the concept SOCIAL-ROLE you cannot use the property HEAD-OF to constrain this concept, as there is no such property in the SOCIAL-ROLE concept. You can either request an ontology developer to introduce such a property for this concept or use the property LEADERSHIP, which is a scalar attribute, and change its filler from (<> 0 1) into (> 0.9).

If to constrain the content of a concept you need a property which is not found in the ontological concept frame, you may request an ontology developer to include this property into the concept description.

#### 2.4.4.3 A word sense corresponds to a PROPERTY concept. Non-propositional mapping.

If you have determined that your word sense corresponds to a PROPERTY concept you cannot follow the mapping procedure for EVENTS or OBJECTS, as no word in the lexicon is allowed to be mapped directly to a PROPERTY concept. The reason for this is that a property does not exist on its own but is always attached to an object or event which it describes. This should be reflected in the ontological mapping. Such a mapping is called a non-propositional mapping. Let us consider how to do it on the example.

The word sense *decline*-N2 (a slope) corresponds to the SLOPE-ATTRIBUTE concept in the ontology. It means the slope of a mountain (NATURAL-OBJECT) or the slope of a road (ARTIFACT). Usually objects or events whose properties a word describes are represented by syntactic adjuncts of the property words (the adjuncts are mountain and road in our example). It is the meaning of these adjuncts that is used as the main concept in the ontological mapping.

decline-N2 (a slope)
CONCEPT: ^$ADJUNCT1
PROPERTY: IS-A
FACET: SEM

FILLER: NATURAL-OBJECT, ARTIFACT
PROPERTY: SLOPE-ATTRIBUTE
FACET:VALUE
FILLER: (<>0 1)

In this example, the property IS-A is used to constrain ^$ADJUNCT1 (the caret symbol ^ reads as the meaning of ) to NATURAL-OBJECT or ARTIFACT which are in the DOMAIN of the property SLOPE-ATTRIBUTE and the filler (<> 01) is in the RANGE of this property.

It is sometimes necessary and possible to make the semantic description of a word sense more precise. An elaboration of the above example presented below.

### 2.4.5 Deep semantic description of a word sense

To make the semantic description of a word sense more accurate you should attempt to factorize the word sense into relevant components; find a corresponding ontological concept for every sense component; put the concepts in the correct relative position within the semantic description.

To do this make sure you understand what kinds of concept (PROPERTY, OBJECT or EVENT ) you are dealing with.

Remember:

- a PROPERTY to be used to constrain the meaning of a particular OBJECT (EVENT) should be already defined in the ontological frame of this OBJECT (EVENT). In other words, to have a certain PROPERTY as a slot in its frame, an OBJECT(EVENT) should be in the DOMAIN of this PROPERTY;
- a PROPERTY cannot be used as a filler of a PROPERTY;
- an OBJECT or an EVENT cannot be used as a slot of a concept.

As an example, consider the semantic description of *decline*-N2 (a slope) given above. The meaning of *decline*-N2 includes such meaning components as being a slope expressed by the SLOPE-ATTRIBUTE concept and being a property of some natural or artificial object expressed by the IS-A, ARTIFACT or NATURAL-OBJECT concepts.

To attain a more precise semantic description of *decline*-N2, it is worth indicating that a slope is a property of something which runs downwards. This leads us to the EVENT concept CHANGE-LOCATION and to the property DIRECTION-OF-MOTION. The scalar attribute DIRECTION-OF-MOTION has downward among the literals included into its range, which makes this concept appropriate for our purpose.

It is essential to put the concepts at the appropriate level in the semantic description. Let us consider where to place the property DIRECTION-OF-MOTION. We can only include DIRECTION-OF-MOTION as a property of an OBJECT or EVENT, but none of the OBJECT concepts considered so far (ARTIFACT and NATURAL-OBJECT) has the slot DIRECTION-OF-MOTION in its ontological frame. The EVENT CHANGE-LOCATION does have DIRECTION-OF-MOTION as its slot. This is why we put DIRECTION-OF-MOTION as a property of DIREC-

TION-OF-MOTION. The filler of DIRECTION-OF-MOTION in our description will be constrained to downward.

What remains to be determined is where to place CHANGE-LOCATION. It is an EVENT and cannot be used as a slot in any concept. It can only be a filler. It is a filler of the PATH-OF slot in both the ARTIFACT concept and the NATURAL-OBJECT concept. As a result we have the following semantic description:

decline-N2 (a slope)
CONCEPT: ^$ADJUNCT1
  PROPERTY: IS-A
    FACET: SEM
      FILLER: NATURAL-OBJECT, ARTIFACT
  PROPERTY: PATH-OF
    FACET: SEM
      FILLER: CHANGE-LOCATION
  PROPERTY: DIRECTION-OF-MOTION
    FACET: VALUE
      FILLER: downward
  PROPERTY: SLOPE-ATTRIBUTE
    FACET:VALUE
      FILLER: (<>0 1)

### 2.4.6 Semantic description of verbs

First of all, read the definition of a verb sense to decide into what kind of ontological concept it should be mapped.

- Verbs most often map into EVENTS and never to OBJECTS.
- Verbs can sometimes map into PROPERTY concepts (usually RELATIONs).

For example, the verb *describe*-V1 ( to give an account or representation of in words ) maps into the PROPERTY concept DESCRIBES.

#### 2.4.6.1 How to map a verb sense into an ontological concept

The first step in the semantic description of the sense of a word is to find an ontological concept which covers the sense on which you are working. It is probably best to proceed as follows.

First try to understand whether your verb sense represents an EVENT or a PROPERTY and search along the corresponding branches of the ontology tree.

Important. When the concept is found, check whether the definition of the concept covers the definition of your word sense. If not, search for another concept.

For example, when describing the verb sense of appear-V2 (come into sight) you can feel that it is an event and scrolling through various branches of the EVENT subtree will eventually lead you to the path EVENT > PHYSICAL-EVENT> CHANGE-LOCATION > APPEAR (come into sight).

When the concept is found, further steps in semantic description depend upon the type of a concept (EVENT or PROPERTY) you found for your word sense. Mapping verb senses into EVENTS differs from mapping them into PROPERTIES.

#### 2.4.6.2 A verb sense corresponds to an Event: Direct and Constrained mapping

When your verb sense corresponds to a particular EVENT then if the content of the EVENT completely corresponds to the sense of the verb, you have what is called a direct mapping and your semantic description of the word sense is done.

For example for the word sense: *appear*-V2 ( come into sight ) there is a direct mapping into the concept APPEAR (come into sight).

If the content of the EVENT is close but does not completely correspond to the sense of the verb, you have what is called a constrained mapping and your semantic description of the word sense includes relaxing or constraining some of the property values of the concept.

For example, when describing the verb sense of *appear*-V1 ( seem, give a certain impression or have a certain outward aspect ) after scrolling through various branches of the EVENT subtree along the path EVENT > SOCIAL-EVENT> ABSTRACT-SOCIAL-ACTIVITY > IMPRESS (to affect strongly the mind or emotions of ) you will probably feel that the content of the concept IMPRESS and the sense of *appear*-V1 differ in that the verb sense contains a component of uncertainty and is less intensive.

To introduce these semantic components into the semantic description of the verb *appear*-V1, look through the properties of the concept IMPRESS and try to find a property (or properties) that would help you to introduce these meaning components into your description. There are such properties as ADEQUACY and INTENSITY whose value range is between zero and one (<> 0 1). The value 0 means that the event represented by the concept does not have these properties, the value 1 means that the event is absolutely adequate and most intensive. The meaning of *appear*-V1 consists in that what you feel might not be completely adequate or intensive. So, you should include these properties in the semantic description with the value (0.5). The resulting semantic description will be as follows:

```
appear-V1
  CONCEPT: IMPRESS
    PROPERTY: ADEQUACY
      FACET: VALUE
        FILLER: (0.5)
    PROPERTY: INTENSITY
      FACET: VALUE
        FILLER: (0.5)
```

#### 2.4.6.3 A word sense corresponds to a PROPERTY concept. Non-propositional mapping

It may happen that the sense of a verb corresponds to a PROPERTY concept. In such a case you cannot follow the mapping procedure for EVENTS. as no word in the lexicon is allowed to be mapped directly to a PROPERTY concept. The reason for this is that a property does not exist on its own but is always attached to an object or event that it describes. This is reflected with the help of a non-propositional mapping. Let us consider how to do it on an example.

The verb sense *involve*-V1 ( engage as a participant: You involved me in your family affairs ) corresponds to the PROPERTY (RELATION) PARTICIPANT-IN-SCHEME (The relation between an organized event and a human who participates in it). This means that the property of participation refers to some social event (wedding, party or, in our example, somebody's family affairs). This event is referred to by one of the syntactic adjuncts from the subcategorization pattern of the verb *involve*-V1. In the format used in the interface this pattern looks as follows:

(Adjunct1 (NP case:nominative))-V-(Adjunct2(NP case:accusative))-(Adjunct3(PP preposition:in))

It is the meaning of Adjunct3 that refers to the event (described by the property HAS-PARTICIPANT) and consequently is used as the main concept in the ontological mapping.

involve-V1 (engage as a participant)
CONCEPT: ^$ADJUNCT3
  PROPERTY: IS-A
    FACET: SEM
      FILLER: SOCIAL-EVENT
  PROPERTY: HAS-PARTICIPANT
    FACET:SEM
      FILLER: HUMAN

In this example, the property IS-A is used to constrain ^$ADJUNCT3 (the meaning of ADJUNCT3) to SOCIAL-EVENT which is in the DOMAIN of the property HAS-PARTICIPANT while the filler HUMAN is in the RANGE of this property.

### 2.5 Acquiring Syntax-to-Semantics Mapping Information

#### 2.5.1 Linking Nouns

Linking shows how the syntactic description is connected to the semantic description of a word sense. This means that you should skip the linking stage for simple nouns, as they do not have a syntactic description in the lexicon. For complex nouns, at the linking stage you should indicate semantic content for every component of the word sense subcategorization pattern.This is done in terms of linking an Adjunct or a Complement to a corresponding semantic argument called a case role.

For example, in the subcategorization pattern of the word "movement" the three prepositional phrases correspond to the following case roles:

movement
  (THEME: of the ball)
  (SOURCE: from the window)
  (DESTINATION: to the door)

In the format we currently use this linking is represented as

(THEME([opt]Adjunct1(PP(prep:of))))-(SOURCE[opt](Adjunct2(PP(preposition:from))))
 -(DESTINATION([opt]Adjunct3(PP prep:to))))

The interface will create this record automatically. Your task is only to select an appropriate set of case-roles for every Adjunct or Complement. When a word sense has more than one subcategorization pattern, you should link each pattern. For example,

announcement (THEME: of/about smth)
announcement (THEME: that...)

After you select the case-roles for the adjuncts the interface will present the linking as follows:

(THEME([opt]Adjunct1(PP( prep:of,about))))
(THEME([opt]Compl1(S)))

**Important**. In the case of non-propositional mapping, it is impossible to define the case role for an adjunct in the subcategorization pattern of a word sense that denotes a property. The notion of a case role may not be applicable at all if this adjunct is an ontological OBJECT. OBJECTS do not have case roles.

### 2.5.2 Linking verbs

Linking shows how the syntactic description is connected to the semantic description of a word sense. At the linking stage, you should indicate the semantic content corresponding to every syntactic component of the word sense subcategorization pattern. This is done in terms of linking an Adjunct or a Complement to a corresponding semantic argument called a case role.

For example, in the subcategorization pattern of the verb *move*-N1 ( to change the position of smth ), the four components of this pattern correspond to the following case-roles:

He [AGENT] moved the ball [THEME] from the window [SOURCE ] to the door [DESTINATION ]

In the format we currently use this linking is represented as

(AGENT(Subject(NP case:nominative)))-V-(THEME(DirectObject(NP case:Accusative)))-(SOURCE[opt](Adjunct1(PP preposition:from)))-(DESTINATION([opt]Adjunct2(PP prep:to))).

The interface will create this record automatically. Your task is only to select an appropriate set of case roles for every Subject, Direct Object, Indirect Object, Adjunct or Complement.

When a word sense has more than one subcategorization pattern, you should link each pattern.

**Important**. When acquiring verb senses, linking is done for every kind of mapping (direct, constrained or non-propositional mapping). Linking for a non-propositional mapping is illustrated using the example of the verb sense *involve*-V1 ( engage as a participant ):

(AGENT(Subject(NP case:nominative)))-V-(EXPERIENCER(DirectObject(NP case:accusative)))-(THEME(Adjunct1(PP preposition:in)))

## 3. Conclusion

Acquiring lexical entries is a difficult business. We have tried to make it less so by:

- ¥ developing a well-structured environment, in which the task can be formulated precisely and implemented systematically;
- ¥ discussing the general principles of semantic acquisition and illustrating them with appropriate examples;
- ¥ deploying a battery of semi-automatic acquisition tools which enable rank-and-file acquirers to produce high-quality homogeneous results after very limited training;
- ¥ offering a series of simplified instructions for using the acquisition tools in a representative environment.

We believe that both the general principles and practical instructions are applicable beyond not only Mikrokosmos but also natural language processing and spill over naturally into any form of descriptive semantics and its applications.

## References


Alston, William P. 1964. Philosophy of Language. Englewood Cliffs, N.J.: Prentice-Hall.

The American Heritage Electronic Dictionary of the English Language 1992. 3rd Edition. New York: Houghton Mifflin.

Barwise, Jon, and John Perry 1983. Situations and Attitudes. Cambridge, MA: M.I.T. Press.

Benson, Morton 1985. Lexical combinability. Papers in Linguistics 18. Special Issue: William J. Frawley and R. Steiner (eds.), Advances in Lexicography, pp. 3-15.

Benson, Morton, Evelyn Benson, and Robert Ilson 1986. Lexicographic Description of English. Amsterdam: John Benjamins.

Black, Max 1954-55. Metaphor. Proceedings of the Aristotelian Society 55, pp. 273-294. Reprinted in: Max Black, Models and Metaphors. Ithaca, N.Y.: Cornell University Press, 1962.

Black, Max 1979. More about metaphor. In: Andrew Ortony (ed.), Metaphor and Thought. Cambridge: Cambridge University Press.

Bloomfield, Leonard 1935. Language. London: George Allen and Unwin.

Congleton, J. E., J. Edward Gates, and Donald Holar (eds.) 1979. Papers in Lexicography in Honor of Warren N. Cordell. Terre Haute, IN: The Dictionary Society of America-Indiana State University.

Davidson, Donald 1978. What metaphors mean. Critical Inquiry 5:1. Reprinted in: Sheldon Saks (ed.), On Metaphor. Chicago: University of Chicago Press, 1979, pp. 29-45.

Fellbaum, Christiane (ed.) 1998. Wordnet : An Electronic Lexical Database (Language, Speech and Communication). Cambridge, MA: MIT Press.

Frege, Gottlob 1963. Compound thoughts. Mind 72, pp. 1-17.

Hartmann, R. R. K. (ed.) 1983. Lexicography: Principles and Practice. London: Academic Press.

Hayakawa, S. I. 1975. Language in Thought and Action, 3rd ed. New York: Harcourt Brace Jovanovich.

Hayes, Patrick 1979. Naive physics manifesto. In: D. Mitchie (ed.), Expert Systems in the Micro-Electronic Age. Edinburgh, UK: Edinburgh University Press.

Hjelmslev, Louis 1959. Pour une s mantique structurale. In: Louis Hjelmslev, Essais Linguis-

tiques. Copenhague: Cercle Linguistique de Copenhague.

Ilson, Robert (ed.) 1986. Lexicography: An Emerging International Profession. The Fulbright Colloquium on Lexicography. Manchester: Manchester University Press/The Fulbright Commission (London).

Katz, Jerrold 1972. Semantic theory and the meaning of good. Journal of Philosophy 61, pp. 736-760.

Katz, Jerrold J. 1972/1974. Semantic Theory. New York: Harper and Row.

Katz, Jerrold J. 1978. Effability and translation. In: F. Guenthner and M. Guenthner-Reutter (eds.), Meaning and Translation: Philosophical and Linguistic Approaches. London: Duckworth, pp. 191-234.

Katz, Jerrold J., and Jerry A. Fodor 1963. The structure of a semantic theory. Language 39:1, pp. 170-210.

Kipfer, Barbara Ann 1984. Workbook on Lexicography. Exeter: University of Exeter.

Lakoff, George 1987. Women, Fire, and Dangerous Things. Chicago: Chicago University Press.

Lakoff, George, and Mark Johnson 1980. Metaphors We Live by. Chicago: Chicago University Press.

Landau, Sidney I. 1984. Dictionaries: The Art and Craft of Lexicography. New York: Charles Scribner s Sons.

Mahesh, Kavi 1996. Ontology development for machine translation: Ideology and methodology. Memoranda in Computer and Cognitive Science MCCS-96-292. Computing Research Laboratory, New Mexico State University.

Mahesh, Kavi, and Sergei Nirenburg 1995. A situated ontology for practical NLP. A paper presented at the IJCAI 95 Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, August 19-21.

McCawley, James D. 1988. The Syntactic Phenomena of English, Vol. II. Chicago: Chicago University Press.

McDavid, Raven I., and Audrey R. Duckert (eds.) 1973. Lexicography in English. Annals of the New York Academy of Sciences.

Miller, George A., Christiane Fellbaum, J. Kegl, and K. Miller 1988. Wordnet: An electronic lexical reference system based on theories of lexical memory. R vue Qu becoise de Linguistique 17, pp. 181-211.

Nirenburg, Sergei, and Victor Raskin 1986. A metric for computational analysis of meaning: Toward an applied theory of linguistic semantics. Proceedings of COLING 86. Bonn, F.R.G.: University of Bonn, pp. 338-340. See the full version in: Nirenburg et al. 1985.

Nirenburg, Sergei, and Victor Raskin 1987. The subworld concept lexicon and the lexicon management system. Computational Linguistics 13:3-4, pp. 276-289.

Nirenburg, Sergei, and Victor Raskin 1996. Ten Choices for Lexical Semantics. Memoranda in Computer and Cognitive Science MCCS-96-304, Computing Research Laboratory, New Mexico State University.

Nirenburg, S. and V. Raskin 2001. Ontological Semantics. In preparation.

Nirenburg, Sergei, Victor Raskin, and Boyan Onyshkevych 1995. Apologiae ontologiae. Memo-

randa in Computer and Cognitive Science MCCS-95-281. New Mexico State University: Computing Research Laboratory. Reprinted in: Judith Klavans, Bran Boguraev, Lori Levin, and James Pustejovsky (eds.), Symposium: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity. Working Notes. AAAI Spring Symposium Series. Stanford, CA: Stanford University, 1995, pp. 95-107. Reprinted in a shortened version in: TMI 95: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation. Centre for Computational Linguistics, Catholic Universities Leuven Belgium, 1995, pp. 106-114.

Onyshkevych, Boyan, and Sergei Nirenburg 1995. The Lexicon in the Scheme of KBMT Things. Memoranda in Computer and Cognitive Science MCCS-94-277. Computing Research Laboratory, New Mexico State University, 1994. Reprinted as: A Lexicon for Knowledge-Based Machine Translation, in: Bonnie J. Dorr and Judith Klavans (eds.), Special Issue: Building Lexicons for Machine Translation II. Machine Translation 10:1-2., pp. 5-57.

Quine, Willard Van Orman 1960. Word & Object. Cambridge, MA: M.I.T. Press.

Raskin, Victor 1971. K teorii yazukovykh podsistem /Towards a Theory of Linguistic Subsystems/. Moscow: Moscow State University Press.

Raskin, Victor 1986. Script-based semantic theory. In: Donald G. Ellis and William A. Donohue (eds.), Contemporary Issues in Language and Discourse Processes. Hillsdale, N.J.: Erlbaum, pp. 23-61.

Raskin, Victor 1987. What is there in linguistic semantics for natural language processing? In: Sergei Nirenburg (ed.), Proceedings of Natural Language Planning Workshop. RADC: Blue Mountain Lake, N.Y., pp. 78-96.

Raskin, Victor, and Sergei Nirenburg 1995. Lexical Semantics of Adjectives. Memoranda in Computer and Cognitive Science MCCS-95-288, Computing Research Laboratory, New Mexico State University.

Raskin, Victor, and Sergei Nirenburg 1998. An Applied Ontological Semantic Microtheory of Adjective Meaning for Natural Language Processing. Machine Translation 13:2-3.

Raskin, Victor, and Irwin Weiser 1987. Language and Writing: Applications of Linguistics to Rhetoric and Composition. Norwood, N.J.: Ablex.

Searle, John R. 1969. Speech Acts. Cambridge: Cambridge University Press.

Searle, John R. 1979. Metaphor. In: John R. Searle, Expression and Meaning. Cambridge: Cambridge University Press, pp. 76-116. Also printed in: Andrew Ortony (ed.), Metaphor and Thought. Cambridge: Cambridge University Press, 1979.

Svens n, Bo 1993. Practical Lexicography: Principles and Methods of Dictionary-Making. Oxford: Oxford University Press.

Tarski, Alfred 1956. The semantical conception of truth. In: Leonard Linsky (ed.), Semantics and the Philosophy of Language. Urbana, IL: University of Illinois Press. Reprinted in the 2nd paperback edition, 1972, pp. 13-47.

Tversky, Amos, and Daniel Kahnemann 1983. Probability, representativeness, and the conjunction fallacy. Psychological Review 90:4, pp. 293-315.

Vendler, Zeno 1963. The grammar of goodness. The Philosophical Review 72:4, pp. 446-465.

Webster's 7th Collegiate Dictionary 1963. Merriam-Webster.

Weinreich, Uriel 1966. Explorations in semantic theory. In: T. A. Sebeok (ed.), Current Trends in Linguistics, Vol. III. The Hague: Mouton, pp. 395-477.

Whorf, Benjamin Lee 1956. Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. Ed. By J. B. Carroll. New York: Wiley.

Zajac, R mi , Evelyne Viegas, and Svetlana Sheremetyeva 1997. The Generic Structure of a Lexical Knowledge Base Entry. Research Memorandum, Computing Research Laboratory, New Mexico State University.

Zgusta, Ladislav 1971. Manual of Lexicography. Prague-The Hague-Paris: Academia-Mouton.

Zgusta, Ladislav (ed.) 1980. Theory and Method in Lexicography: Western and Non-Western Perspectives. Columbia, S.C.: Hornbeam Press.

Ziff, Paul 1960. Semantic Analysis. Ithaca, N.Y.: Cornell University Press.