In Proceedings of the AISB'96 Workshop on Multilinguality in the Lexicon, Brighton, UK, April 1-2, 1996

Lexicons in the MikroKosmos Project

Sergei Nirenburg, Stephen Beale, Kavi Mahesh, Boyan Onyshkevych[†]
Victor Raskin[††], Evelyne Viegas, Yorick Wilks[§], Remi Zajac
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003, USA
mikro@crl.nmsu.edu

† Also of US Department of Defense.
†† Also of Purdue University NLP Lab.
§ Also of University of Sheffield.

1 Introduction

Our approach to the lexicon has been driven by both theoretical and practical concerns. From a theoretical viewpoint, we are interested in capturing the core meanings of texts. We believe that lexical semantics is a crucial knowledge component of text meaning derivation, as it describes lexical meanings in their composition and combination[1] properties. A major tenet of our methodology is a separation of what is language-dependent and what is universal in semantic specifications. To help in determining this, our research methodology centrally includes a cross-linguistic perspective. From a practical standpoint, our lexicons are designed to support machine translation (MT) systems, in which the issue of multilinguality is clearly central. Theoretical and practical concerns are combined into the Mikrokosmos (μK) approach, where, within the paradigm of knowledge-based machine translation (KBMT), we build lexicons for different natural languages, relying on a multilingual framework.

All the central lexicons[2] in KBMT are essentially bilingual, in that they provide mappings from a natural language into an artificial knowledge-representation language, and vice versa. The lexicon structure used in the μK approach is universal, in that it does not change from language to language and the artificial language used in each lexicon (whether for Spanish, Japanese, Russian, English, etc.) is the same language of representation, called Text Meaning Representation (TMR) in μK . Thus each language's lexicon provides a mapping for each word sense into a meaning representation in the TMR language, in addition to providing extensive information about the behavior of the word in utterances in that natural language.

In the following sections, we briefly discuss the different issues we would like to raise at the workshop. In section 2, we illustrate how are our lexicons organized, the format we adopted, and what types of information we encode in them. In Section 3, we justify the choices made in Section 2 in terms of content and representations and also directly address the issue of multilinguality of lexical meaning representations. Finally, we describe some tools needed to acquire and check the entries we are building.

[1] By combination we refer to lexical units which do not comply to full compositionality, such as constructions, collocations or idioms.

[2] We consider some special-purpose lexicons as non-central, specifically those which support word segmentation for Japanese or Chinese, name detecting, etc.

2 Organization and Use of Computational Lexicons

It is well known in computational lexical semantics that a sense enumeration approach based exclusively on subcategorization differences is computationally expensive and deficient from a theoretical viewpoint, because it fails to capture the core meaning of words (Boguraev and Pustejovsky, 1990), (Sanfilippo, 1992), (Viegas and Nirenburg, 1995a). Our lexicons are characterized by a mixture of generative capabilities and sense enumeration based strictly on meaning differences. Our lexicons consist of superentries (Meyer et al., 1990), one for each citation form, independently of their part of speech (the verb and noun forms of *walk* are under the same superentry), under which are listed word senses. Each word sense is identified by a unique identificator, or lexeme, such as *walk-V1* or *walk-N1*, (Mel'čuk et al., 1984), (Onyshkevych and Nirenburg, 1994).

2.1 Lexicon Zones in a Lexeme

The information about a lexeme is minimally divided into **zones** corresponding to various levels of lexical information, (Meyer et al. 1990).

`CAT`egory: *Noun, Verb, Pronoun,...*;
`MORPH`phology: for irregular forms and stem changes *mouse* vs *mice*;
`COMMENTS`: administrative information, definition, examples...;
`ORTH`tography: abbreviations, *United States of America* vs *USA*;
`PHON`onology;
`SYN`tactic-`STRUC`ture: essential subcategorisations;
`SEM`antic-`STRUC`ture: the semantics, with selectional restrictions, in terms of its unsaturated TMR;
`LEX`ical-`REL`ations: collocational information;
`LEX`ical-`RULES`: rules that apply to this lexeme[3];
`STYL`istics: information on stylistic factors, such as familiarity, etc..., and sub-zones containing triggers for analysis and generation.

As can be seen, some zones, such as `SEM-STRUC`, are "more multilingual" than others, in that the meaning of *eat* in English, and *manger* in French or *comer* in Spanish, will be shared across languages; whereas some other zones, such as `LEX-REL` encode language-related idiosyncrasies, for instance, information about collocations.

2.2 Towards a Multi-purpose Knowledge Base

Acquiring a large-scale computational semantic lexicon is a very expensive enterprise (Viegas and Nirenburg, 1995b), (Viegas et al., 1996); this is why it is advantageous to build lexicons which are reusable for other domains or applications. We need lexicons that are multi-purpose, supporting the three following paradigms:

a **multi-lingual**: French, English, Japanese, Russian, Spanish, etc...,

b **multi-media**: containing linguistic and ontological information for natural language processing as well as phonological information, essentially for speech recognition and production,

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

[3](Viegas et al., 1996) illustrates in detail the design and implementation of lexical rules for **Spanlex** the Spanish lexicon developed for μK , where about 100 morpho-semantic rules applied to the different meanings of 1056 verb citation forms, produced about 35,000 new candidate entries.

3 Multilinguality of Lexical Meaning Representation

In this section we focus our attention on how meaning is encoded in a representation language that is not defined in terms of any one natural language. The lexicon connects with the ontology (see below) and the onomasticon (a special-purpose lexicon of named entities such as cities, corporations, or products), thus becoming the locus of links between lexical units in texts and the language-neutral TMR. The interlingual meaning representation (or TMR) is derived from representations of word meanings in computational lexicons and from representations of world knowledge in ontologies (and possibly episodic knowledge bases). The formalism for the lexical semantic specification in this zone in our lexicon, the `SEM-STRUC` zone is discussed in detail in other sources, such as (Onyshkevych and Nirenburg, 1994), (Onyshkevych, 1995).

Each lexical entry contains a representation of its semantics, represented by using terms from the ontology as primitives (in addition to other non-ontological primitives, e.g., to reflect speaker attitudes and modality). These representations of lexical meaning may be defined using any number of ontological primitives, which we call *concepts*. We give below the example of the syntax and semantics for the Spanish entry *beber* (drink) (Figure 1), using the typed feature structures (tfs) as described in (Pollard and Sag, 1987).

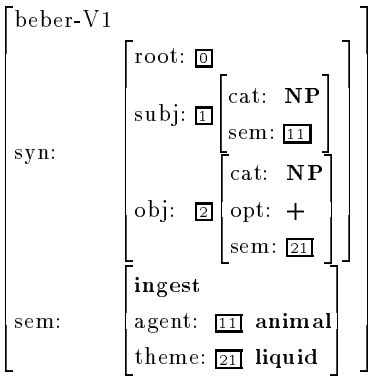

Figure 1: Partial Lexicon Entry for the Spanish lexical item *beber*.

Notice that the concept *beber* maps into is INGEST, which has selectional restrictions included in the ontology, such as ANIMAL and INGESTIBLE for its agent and theme respectively. These selectional restrictions work fine for *eat* but not *drink*, there we constrained the theme of INGEST to LIQUID as shown in the entry above for Spanish. When the meaning is represented using multiple concepts, they are tightly interconnected and constrained as appropriate. Any of the concepts in the ontology (currently numbering about 5,000 in μK) may be used (singly or in combination) in a lexical meaning representation.

The Ontology The set of symbols and possible relationships between them are grounded in a language-independent knowledge source called the *ontology.* The symbols are defined as *concepts* in the ontology. As described, e.g., in (Mahesh and Nirenburg, 1995), (Mahesh, 1995), the ontology is a large collection of information about `EVENT`s, `OBJECT`s and `PROPERTY`s in the world, displayed

in Figure 2. In addition to the taxonomic multi-hierarchical organization, each concept has a number (currently averaging 14) of other local or inherited links to other concepts in the ontology, via relations (themselves defined in the `PROPERTY` sublattice). These links include case-role-like relations linking `EVENT`s to semantic constraints on the allowable fillers of those case-roles (i.e., selectional restrictions), properties (such as `MANUFACTURER-OF`) of things like COMPANYs, etc., as can be seen in Figure 3.

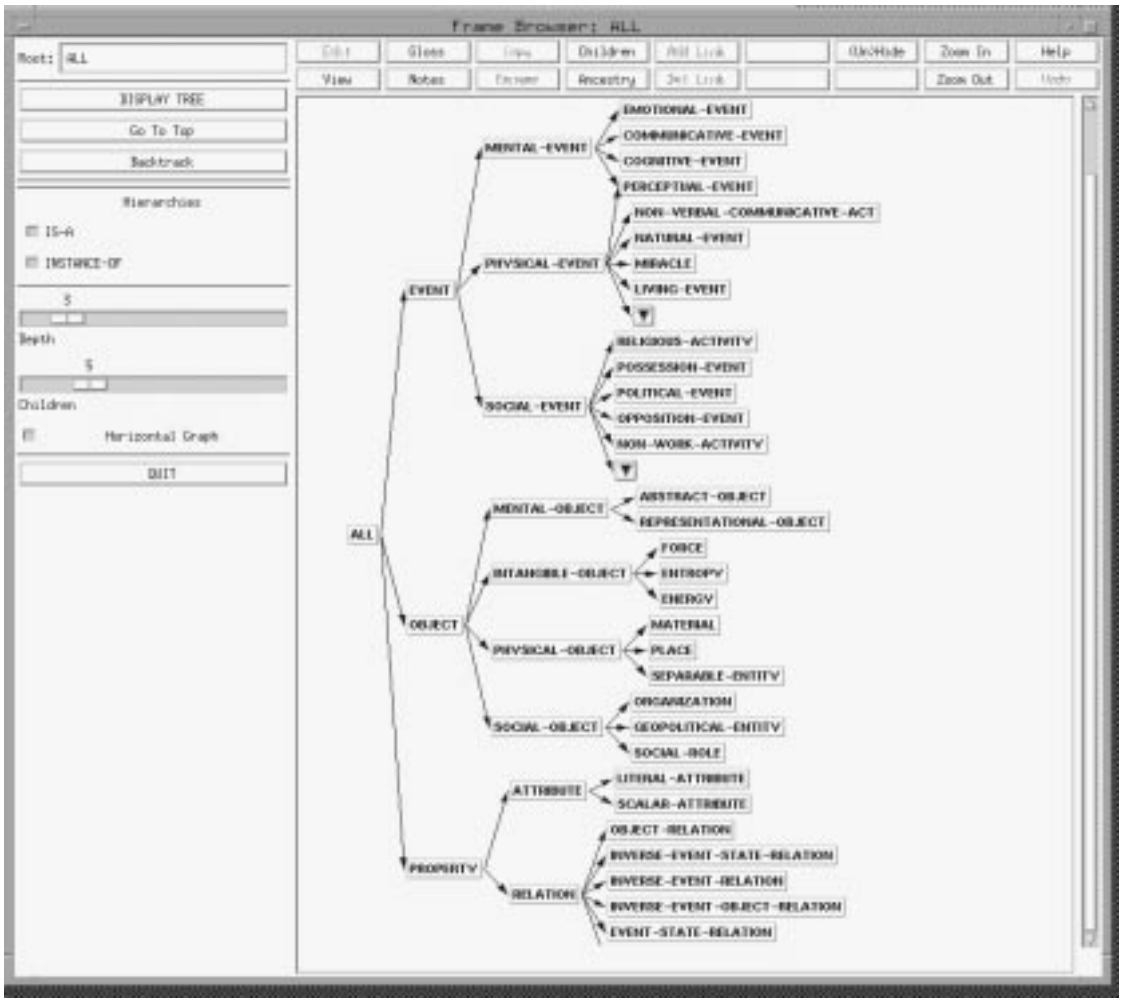

Figure 2: Ontology Top Level.

In interlingual machine translation, the principal reasons for using an ontology are:

- to provide a grounding for representing text meaning in an interlingua;
- to enable lexicons for different languages to share knowledge;
- to store selectional restrictions and other pieces of world knowledge;
- to "fill gaps" in text meaning by making inferences based on the content of conceptual knowledge in the ontology;
- to resolve semantic ambiguities and interpret non-literal language by making inferences using the topology of the ontology to measure the semantic affinity between meanings;

See (Mahesh and Nirenburg, 1995) and (Beale, Nirenburg, and Mahesh, 1995) for more examples of the use of the ontology in the KBMT paradigm.

The ontology parallels μK lexicons in developing the MT system. Word meanings are represented partly in the lexicon and partly in the ontology. *In principle, the separation between ontology and lexicon is as follows: language-neutral meanings are stored in the former; language-specific information in the latter.* Figure 3 illustrates the constraints encoded in the concept INGEST; note that the Spanish word *beber* (*to drink*) was mapped into this concept, except that the selectional restriction specified in the theme of the lexicon entry of *beber* (Figure 1) constrained it to be of type LIQUID.

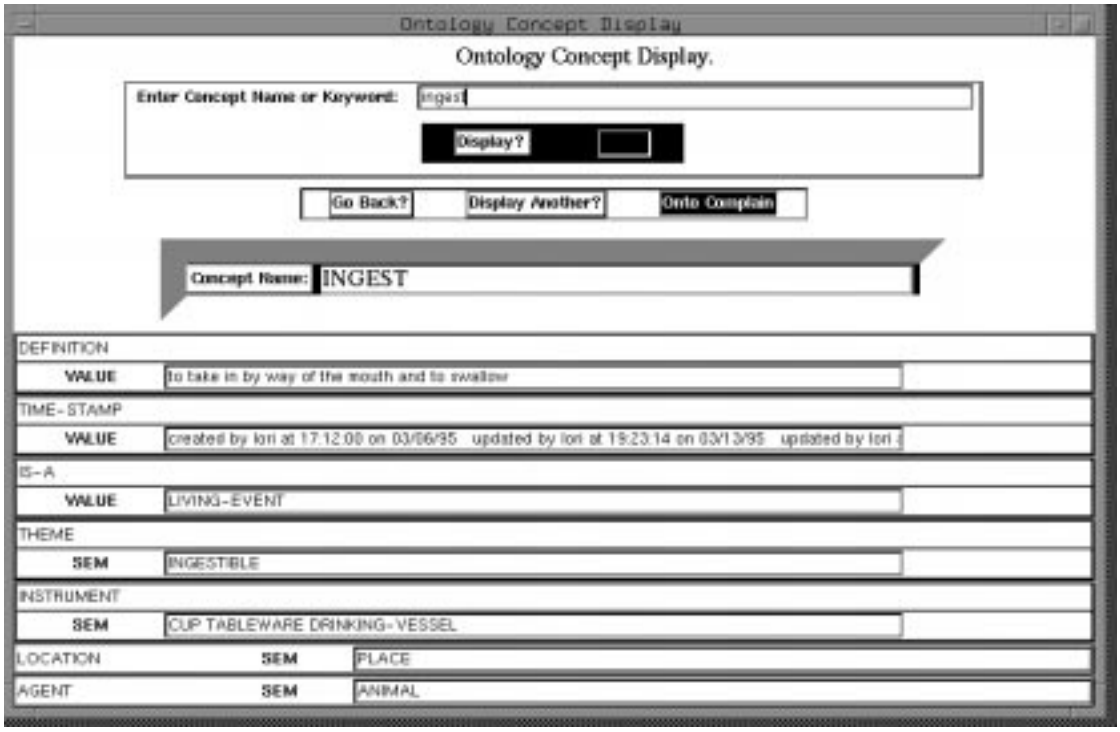

Figure 3: Ontological Description for INGEST.

In a multilingual situation, however, it is not easy to determine this boundary. As a result, ontology and lexicon acquisition involves a process of daily negotiations between the two teams of acquirers, as is described in section 5.1.

In each lexical entry, the syntax-semantics interface links argument structure and local syntactic context with elements in the meaning representation. In the trivial cases, the syntactic subject might be linked to the semantic agent, and so on. However, this mapping is often very complex, and exists for nouns, verbs, adjectives, prepositions, and so on.

It is important to note that there need not be any correlation between syntactic category and semantic or ontological class. For example, although many verbs are `EVENT`s and a number of nouns are represented by concepts from the `OBJECT` subtree (such as the class of artifacts), frequently this is not the case. This is particularly the case with words derived via Lexical Rules, (LRs). Many LRs change the syntactic category of the input form; in our model the semantic category is often

preserved in many of these LRs. For example, the verb *destroy* may be represented by an `EVENT`, as will the noun *destruction* (with a different linking in the syntax-semantics interface, of course). Similarly, *destroyer* (as a person) would be represented using the same event with the addition of a `HUMAN` as a filler of the agent case-role. This built-in transcategoriality is a very natural aspect of the interlingual approach to MT, avoiding many of the category mismatches and misalignments that plague other paradigms in MT.

4 The Practical Principle of Effability

While our ontology is language-independent and thus usable in multilingual projects without any changes, conventional wisdom would claim that the actual lexical entries are language-dependent. We have discovered that this is not completely true. The superentries are, for the most part, language-dependent, but the individual lexical entries, one for each meaning, travel freely from language to language. The theoretical basis for that is the principle of effability (Katz, 1978, Frege, 1963, Tarski, 1956, Searle, 1969, Raskin and Nirenburg 1995).

What this means is that the individual meanings of a polysemous word in one language will be present in the lexicons of other languages as well, but not necessarily bunched together in one superentry. The effability principle states, technically, that "each proposition can be expressed by some sentence in any natural language." In our work, we extended it to the individual lexical entries, so that our modified principle of effability is as follows: "for most meanings expressed by a word in one language, there is a word in another language which expresses this meaning, though it may express different meanings, too."

Consider the English word *good*. Its basic meaning in μK is represented as a high range on the scale expressing a speaker's evaluative attitude and scoped over the noun syntactically modified by the adjective. This meaning will be expressed by a word in just about any language, and, therefore, this lexical entry, without any change in its semantic part, will go into the lexicons of those languages. There will occasionally be a meaning expressed by a word in only one or very few languages, but most meanings are pretty language-universal.

Now, the superentry for the English adjective *good*, the set of all individual meanings of the word, is most likely unique. Such meanings of *good* as the good moral character of a person or the validity of a check or the reliability of a promise are not likely to appear in the superentry of the same words in other languages, which all express the basic evaluative meaning of *good*.

The extended principle of practical effability is a useful simplification grounded in the idea about variable grain size of description determined by a concrete application. This principle makes it possible to "translate" individual lexical entries from language to language, thus making lexical acquisition in new languages much faster. We have tested it for several thousand English adjectives, whose individual meanings were translated quickly and accurately into Spanish by a bilingual graduate research associate, who did not have – and did not need – any special descriptive-semantic skills.

5 Semi-automatic Acquisition and Testing of Multilingual Computational Lexicons

We now turn to the tools developed in μK for acquisition and validation of lexicon entries.

5.1 Acquisition Tools

Since we need to support the acquisition of lexicons in a variety of languages, and our lexicon format is language-independent, we provide support tools for lexicon acquisition that are user-configurable to support any language. The enabling underlying stratum of our tool suite is a multi-lingual

display, input, and output widget that supports display and input of a wide range of languages in the native script, using UNICODE or national representation codesets. The user is able to configure the lexical acquisition tool to use the language-specific syntactic category tag set, a list of syntactic categories, syntactic phrase or branching structure inventory, syntactic class inheritance hierarchy, syntax-semantics interface class inheritance hierarchy, etc. What makes the tool possible, however, is the fact that not only is the lexicon format language-independent, but that the (lexical) semantic specification and format for the syntax-semantics interface are language-independent as well. The lexicons for all languages are defined using the same set of primitives (they all use the same ontology plus a few additional non-ontological primitives), thus the tool suite used to support browsing, searching, or editing the ontology do not change. So far, the tool suite has been used to support the input of lexicon entries for Spanish, Japanese, English, and Russian. Figure 4 illustrates the acquisition process of static knowledge, such as the lexicon and the ontology, which are being used dynamically by the semantic analyser: lexicon acquirers have access to various on-line resources, such as **corpus search**, **look-up dictionary**, **ontology browser** tools, whereas the primary source of information for concept acquisition is the flow of requests for additions and changes from lexicographers, testers, and analyzer developers. Domain experts, other ontologies, domain dictionaries, and published standards (e.g., SICM, 1987) also provide useful information when available.

This set of tools is being shared across geographical, disciplinary, and project group boundaries on a daily basis.

5.2 Validation Tools

As the size of computational lexicons begins to increase, and as more and more automated methods are used to populate them, the need for new testing methodologies grows. Testing lexicons can be divided into the following three areas:

- Form (syntax). Tools are needed to detect improperly formed lexicon entries. For example, misplaced parentheses and quotes, although uninteresting at a theoretical level, can cause many headaches practically.

- Meaning (semantics). The meaning of a lexicon entry should be consistent within the entry, within the lexicon as a whole, and with respect to any other knowledge sources, such as ontologies, that are used in the lexicon. This consistency needs to be checked within each zone of the lexicon. For instance, inside the syntax zone, the syntactic structures indicated should be consistent with the grammar of the language. In the semantics zone, meaning assignment and constraint checking should be consistent with meaning definitions in the ontology. Assigning an AGENT to a PLACE in some lexicon entry, for example, is suspect, as is constraining the AGENT of a KILL event to be a MONKEY.[4]

- Correctness (pragmatics). The lexicons need to fulfill their purpose, either to analyze input text or to correctly generate text from input semantic representations (or both). It is impossible to automatically test the correctness of a lexicon in this regard over a large corpus. However, we have created a set of tools that will automatically create input test sentences (or test semantic structure inputs) that can be used to evaluate specific lexicon entries. This helps in two ways. First, simply seeing a sentence that fits the syntax and semantics present in the lexicon entry can often highlight errors. For example, after extending a base lexicon by automatically generating thousands of nominalizations from verbal forms, it would be advantageous to be able to quickly scan a list of sentences in order to identify erroneous forms.

[4] We recognize that some languages might have a specific word for felonious monkeys; these tools will simply point out **possible** problems and let the user determine if any action needs to be taken.

Figure 4: Acquisition Process of Static Knowledges.

This type of testing ensures that the lexicon items will be useful and will be applied in the correct situations. Second, after generating the test items, the lexicons can be used to process them, with the results also subject to review. This ensures that the lexicon item produces the intended results.

Tools such as these are especially useful when multilingual lexicons are automatically created (or enhanced) from a source lexicon in a different language. For instance, (Viegas and Beale, 1996) describe the conversion of a Spanish lexicon into an English one. The tools described above quickly expose errors introduced in this type of conversion, a process that would take much more effort if done manually.

6 Conclusion

In this abstract, we have briefly sketched, theoretical concerns about multilinguality, including the necessity of, adopting a cross-linguistic perspective, a lexical semantic approach which tries to capture the core meaning of words. We have discussed how to represent lexicon entries, in terms of content and format. We also argue for multi-purpose lexicons, which must support multi-lingual, multi-media and multi-process paradigms. We sketched that multilinguality lies more along the semantics or meaning of words than their syntax or even pragmatic behaviour. Finally we addressed the practical issue of acquiring and validating the entries we build using "multilingual" tools. We hope to be able to present our work in substantially greater detail, with abundant examples and, if possible a demonstration of our ontology, lexicons and the tool suite.

References

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

Boguraev, B. and J. Pustejovsky (1990) *Knowledge Representation and Acquisition from Dictionary.* Coling Tutorial, August 16-18, 1990, Helsinki, Finland.

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

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.

Mahesh, Kavi, and Sergei Nirenburg (1995). A situated ontology for practical NLP. In the Proceedings of *IJCAI'95* Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, August 19-21.

Mahesh, K. (1996). *Ontology Development: Ideology and Methodology.* Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University.

Mel'çuk, I., N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja and A. Lessard (1984) *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I. Montréal*: Presses de l'Université de Montréal.

Meyer, I., Onyshkevych, B. and L. Carlson (1990) *Lexicographic Principles and Design for Knowledge-based Machine Translation.* Technical Report CMU-CMT-90-118, Carnegie Mellon University.

Nirenburg, S., V. Raskin and B. Onyshkevych (1994) *Apologiae ontologiae.* Memoranda in Computer and Cognitive Science MCCS-95-281. New Mexico State University: Computing Research Laboratory.

Onyshkevych, Boyan, and Sergei Nirenburg (1994). The lexicon in the scheme of KBMT things. Memoranda in Computer and Cognitive Science MCCS-94-277. Las Cruces, N.M.: New Mexico State University.

Onyshkevych, B. (1995). *A Generalized Lexical-Semantics-Driven Approach to Semantic Analysis.* Dissertation Proposal, Carnegie Mellon University, Program in Computational Linguistics.

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

Raskin, V. and S. Nirenburg (1995). *Lexical Semantics of Adjectives: A Microtheory of Adjectival Meaning.* Memoranda in Computer and Cognitive Science MCCS-95-288. Las cruces, N.M.: New Mexico State University.

Sanfilippo, A. (ed.) (1992) *The (Other) Cambridge Acquilex papers.* Technical Report No. 253, University of Cambridge Computer Laboratory, New Museums Site.

Searle, John R. (1969). *Speech Acts.* Cambridge: Cambridge 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.

Viegas, E. and S. Nirenburg (1995a) The Semantic Recovery of Event Ellipsis: its Computational Treatment. Proceedings of the Workshop on Context and Natural Language, *IJCAI 95*, Montréal, 1995.

Viegas, E. et S. Nirenburg (1995b) Acquisition semi-automatique du lexique. In proceedings of *Langage - Lexicologie - Traductique* (LTT96), Lyon, France.

Viegas, E., B. Onyshkevysh, V. Raskin and S. Nirenburg (1996) *Submit* to *Submitted* via *Submission*: On Lexical Rules in Large-Scale Lexicon Acquisition. To Appear in Proceedings of *ACL'96*).

Viegas, E. and S. Beale (1996) Multilinguality and Reversibility in Computational Semantic Lexicons. Submitted to *INLG 96*, Sussex, UK, 1996.