Apologiae Ontologiae

Sergei Nirenburg
Computing Research Lab
New Mexico State University
Las Cruces, NM 88003 USA
sergei@crl.nmsu.edu

Victor Raskin[1]
Computing Research Lab
New Mexico State University
Las Cruces, NM 88003 USA
raskin@crl.nmsu.edu

Boyan Onyshkevych[2]
US Dept. of Defense
Attn:R525
Ft. Meade, MD 20755 USA
baonysh@afterlife.ncsc.mil

Abstract. Treatment of meaning in NLP is greatly facilitated if semantic analysis and generation systems rely on a language-neutral, independently motivated world model, or ontology. However, the benefits of the ontology are somewhat offset in practice by the difficulty of its acquisition. This is why a number of computational linguists make a conscious choice to bypass ontology in their semantic deliberations. This decision is often justified by questioning the principles underlying ontologies and by challenging the ontology-based semantic enterprise on the grounds of its ostensible irreproducibility. In this paper we briefly illustrate the expressive power of ontological lexical-semantic descriptions used in the Mikrokosmos machine translation project and make a comparison with some of the non-ontological approaches to lexical semantics. We argue that these approaches, in reality, rely on ontologies in everything but name. We claim that no underlying principles for ontologies are possible and explain why the charge of irreproducibility is not valid.

The central tenet of ontology-based semantics is "grounding" the meanings of lexical units in an independently motivated and interpreted system of symbols, an ontology. An ontology-based lexicon is an important component of the Mikrokosmos multilingual MT project (see, e.g., Onyshkevych and Nirenburg, 1994; earlier versions of the same paradigm were described in Nirenburg *et al.,* 1992 as well as Carlson and Nirenburg, 1990 and Meyer, Carlson and Onyshkevych, 1990). Advantages of lexicons supported by ontologies include support for treatment of synonymy (and near-synonymy), control over uniformity of the grain size of lexical descriptions, enabling the treatment of lexical gaps in translation as well as support for a variety of inferencing mechanisms necessary for lexical and other kinds of disambiguation. It has been widely accepted since Bar Hillel (1960) that without such mechanisms a large number of textual phenomena cannot be fully treated. It might, therefore, seem strange that lexicographers and MT researchers are not flocking to this method (cf. a similar lament in Bateman, 1993, p.83). But, of course, there are important reasons for that, some connected with linguistic tradition, and some others with purely engineering considerations. Very briefly, the major objections can be summarized as follows:

- The Redundancy Objection: it is often possible to resolve semantic ambiguities using non-semantic (syntactic, morphological, etc.) clues[3]. This observation has led to a variety of claims that semantics is redundant. As a practical consequence, recognition supplanted interpretation in system designs, as attempts were made at disambiguation but not at *representation* of the meaning of text.

---

1. Also of NLP Lab, Purdue University, W. Lafayette, IN 47907 USA
2. Also of *Computational Linguistics Program,* Carnegie Mellon University, Pittsburgh PA 15213 *USA*
3. Cf. the many suggestions for syntactic treatment of prepositional phrase attachment, a problem whose solution is universally accepted to require semantics.

- **The Language Dependence Objection:** It is widely assumed that a given ontology will reflect the linguistic bias of its builders and will, thus not be really language-neutral. At a different level, this objection is formulated in terms of "upper case semantics" (McDermott, 1978), that is, the claim that ontological concepts are, in reality, camouflaged English words.
- **The Scientific Method Objection:** the apparent irreproducibility of ontological systems and the related unavailability of reliable and easily teachable heuristics or methodology for either ontological or ontology-based lexical acquisition.
- **The Engineering Objection:** the difficulty and cost of developing, maintaining and updating ontologies and the seemingly easy availability of simple and straightforward English-language frameworks for a minimum feature characterization of lexicon entries lead people to opt for the latter.

1. How the Ontology-Based Lexicon Works.

The Mikrokosmos lexical entry consists of 12 zones.[1] Each zone is included for an overt purpose, to support a particular stage in text analysis or generation or to help developers navigate the lexical knowledge base. With respect to basic lexical meaning determination, three kinds of lexicon information are central — subcategorization information (encoded in the SYN-STRUC zone), information on linking syntactic and semantic dependency (encoded using a system of connected variables occupying argument positions in the SYN-STRUC and SEM zones) and basic semantic information (in SEM). As an example, the SEM and ANNO ("annotation") zones from the entry for one of the senses of the Spanish verb *dejar* are shown below.

```
[dejar-v5
        [anno
         [def "varl permits var3 to use (var2) var4 and varl gives up using var4"]
         [ex "dejó su computadora a su hermano"]]
        [sem
         [lex-map
                *permit
                    agent        ^varl
                    beneficiary  ^var3
                    theme        ^var2
                ^var2
                    instance-of  *use
                    agent        ^var3
                    theme        ^var4
                *use
                    agent        ^varl
                    theme        ^var4
                    aspect
                        phase    end
                        telic    no]]]
```

In *dejar-v5* the SEM zone represents in the representation language of our choice what is explained informally in DEF. The actual meaning represented by *dejar-v5* is, of course, equivalent to one of the meanings of the English verb lend.[2] The meaning is

1. A detailed description of the content of the Mikrokosmos lexicon can be found in Onyshkevych and Nirenburg, 1994.

presented in DEF as "permit to use" in order to show explicitly each of the syntax-semantics linking variables. The information about the elidability of var2 (which would lead to an English realization such as "varl lends var4 to var3") as well as the temporal relations are omitted for the sake of simplicity.

For details of the Mikrokosmos formalism, see, for instance, Onyshkevych and Nirenburg, 1994. Here let us mention only that the operator "^" maps a syntactically defined variable into its meaning. The meaning of *dejar-v5* can be glossed as "^varl permits ^var3 to do ^var2" (for example, "The banks permit the companies to use excess office space") and "var2 is the event in which ^var3 uses ^var4" (for example, "The companies use the excess office space"). The meaning also contains a postcondition "^varl cannot use ^var4 for the duration of ^var2" (for example, "The banks cannot use the office space while the companies use it").

The ontological concept ***permit** has the following definition:

```
*permit
        is-a            *directive-act
        agent           *human
        beneficiary     *human
        theme           *event
                                    agent       *permit.beneficiary
                                     precond    *have-authority
                                                              agent       *permit.agent
                                                             . theme      OR
        *permit.beneficiary
                                                                                    *permit.theme
```

During text processing, after an occurrence of *dejar* is identified as *dejar-v5,* the semantic constraints in the concept ***permit** will be used for checking whether the actual fillers of the case role slots (i.e., agent, beneficiary, theme, etc.) conform to the constraints imposed on the values of the latter in the ontology and the lexicon. This is the mechanism for using selectional restriction information in constructing text meaning representations in Mikrokosmos.

2. Alternatives to Ontology

A number of more or less detailed proposals have been advanced in computational lexical semantics over the past seven to ten years. A significant portion of the efforts in this field is devoted to developing formalisms in which to record lexical information. There have also been substantive proposals about the content of the semantic parts of the lexicon entry, including the LCS theory of Jackendoff and his followers (e.g., Jackendoff, 1983, Hale and Keyser, 1986, Atkins *et al.,* 1986, Kegl, 1989, Dorr, 1993), the generative lexicon approach of Pustejovsky and his coauthors (e.g., Pustejovsky, 1991, Pustejovsky *et al.,* 1993) and the work of group most recently associated with the ACQUILEX project in Europe (e.g., Boguraev and Briscoe, 1989, Briscoe *et al.,* 1990, Copestake *et al.,* 1992 and selected chapters from Briscoe *et al.,* eds., 1993).

2. Our ontology does not contain a concept corresponding to the English verb lend because we "do not add a concept if there is already one "close" to it or slightly more general than the one being considered" (Mahesh, 1995).

A most interesting feature of all these, otherwise distinct, approaches is that they all avoid the concept of language-neutral ontology in their theoretical framework, while, in reality introducing elements of metalinguistic apparatus which play the same role as the ontology. The catch is that these elements typically remain of indeterminate theoretical status and carry insufficient information for many processing decisions.

As an example, it might be useful for us to create a treatment for this sense of *dejar* in the generative lexicon framework of Pustejovsky (1991), Pustejovsky *et al.* (1993). Based on the formalizations available in the above papers as well as the entry for build in Pustejovsky (1994), we can conclude that *dejar*-v5 would be characterized as a process with an event structure (probably, a transition followed by a state), an argument structure identifying though not naming the case roles, with some reference to their semantic characteristics in qualia form, plus the qualia structure for the verb itself. The entry would contain semantic information, including some selectional restrictions. Pustejovsky's own examples are also very instructive. Here is the generative lexicon entry for the English noun *tape* (Pustejovsky *et al.,* 1993, p. 335):

tape (x, y)	
CONST	**information(y)**
FORMAL	**physobj(x),2-dimen(x)**
TELIC	**contain(S,x,y)**
AGENTIVE	**write(T,x,y)**

Leaving aside the question of the composition and cardinality of the set of the qualia structures (CONST, FORMAL, TELIC, AGENTIVE), the status of values for these properties remains unclear. These cannot be words of English (because two of them are certainly not). Therefore they must be a part of some other symbol system which is not explicitly introduced. Isn’t it natural to conclude that they are all part of an implicit and undeclared ontology? Moreover, an ontology which seems to be a set of unconnected semantic features? By contrast, all the properties and their value sets in our approach are defined in an overtly specified ontology of multiply interconnected complex, structured elements. In other words, we know why we use the concepts we use, where they come from, with which other concepts they connect and how, what to do when we need a new concept, and how to keep the characterizations consistent.

There is a price to pay for the privilege of having all properties, objects, events and value sets used in specifying lexical meaning based on an explicit ontology. The price is the need for a considerable, complex, and costly acquisition and maintenance effort. Interesting attempts have been made in practical NLP applications, especially in machine translation, to try to avoid paying this price. Traditionally, in machine translation this position has led to the so-called transfer approach, where the ultimate goal has been not to understand the meaning of the source text but rather to find word-sense to word-sense correspondences between a given pair of languages. It has been argued that in many cases devices simpler than a full semantic analysis would be sufficient for such a goal. Indeed, the seven entries of *dejar* seem to offer a near ideal example of such a semantics-free approach[1] to disambiguation: it turns out that an analyzer will be able to determine which of the meanings is used in the text by using syntactic, morphological and other non-semantic clues from the various lexicon entry zones which we, due to space constraints, did not show in the example.

In some other projects, the legitimate desire to avoid acquisition difficulties has led to a redefinition of the notion of ontology as an hierarchically organized set of word senses within one language (see Knight, 1993; and, to some extent, Farwell *et al.,* 1993), which are connected directly to corresponding word senses in other languages. There is no essential difference between this approach (which claims to be interlingual) and the transfer approach because, in the former, the hierarchical structures of the language-dependent ontologies of word senses (whose nodes contain little information other than subsumption and translation equivalents) end up being essentially unused.

One example of the benefits of ontology-based description is in resolution of referential ambiguities, a notoriously difficult problem which does not lend itself to a syntax-based or statistics-based solution. Consider the sentence:

The banks lent the companies excess office space which they didn't need for three months.

In order to select the appropriate Spanish pronoun for *they,* it is necessary to determine its reference, because *banco* is masculine and *compania* is feminine in Spanish. (Of course, if both candidate referents were of the same gender, the strategy advocated in non-semantic approaches would suffice.) The entry for *dejar*-v5 contains information that the lender relinquishes the use of the object lent. It is this knowledge element which allows for the resolution of this particular ambiguity.

In the end, all approaches to machine translation that avoid semantics face difficulties when it is impossible to establish a one-to-one word sense correspondence across any two language-oriented ontologies. The latter are of little use when supralexical context must be taken into account, as in resolution of reference ambiguities. They can also lead to false positive solutions when lexical units are not decomposable to the level of separate words or whenever figurative language is used. Last but not least, in the cases of residual ambiguity (one-to-many word sense correspondences), it will be necessary to write special disambiguation rules which often will not be possible in the absence of elements of world knowledge not usually provided. By contrast, the ontology-based approach insists on a rich and proactive semantic characterization of lexical meaning, to support the automatic derivation of text meaning representations.

3. Reproducibility of Ontology

Our ontologies are arrived at in the same empirical way as generative grammars. As Chomsky put it, “it is unfortunately the case that no adequate formalizable techniques are available for obtaining reliable information concerning the facts of linguistic structure” (1965, p. 19), thus declaring theory formally undiscoverable. Similarly, no algorithmic procedure for acquiring ontologies is available either. And therefore the question of the theory of ontology loses its only claim to relevance.

---

1. In reality, the situation is more complex, as there is an entire spectrum of semantic involvement in text analysis. See Nirenburg and Levin, 1992 for a discussion. Also note that semantic knowledge can be introduced partially by positing an uninterpreted set of semantic features (e.g., animate or physobj) and including their values directly in lexicon entries. This seems to be the approach taken in the ULTRA MT system (Farwell *et al.,* 1993) and Pustejovsky (1991).

Our approach has been criticized (e.g., by Briscoe, 1994) for not stipulating overt principles on which our ontology is based. There are two possible responses to this criticism. First, as argued above, that no such principles are available. Secondly, exactly the same line of criticism applies to all extant metalinguistic frameworks, including those of our critics. This means that if the enterprise of ontology acquisition is considered unsound because it is theoretically undetermined, all the non-ontological and pseudo-non-ontological approaches we have mentioned above, are unsound as well for exactly the same reason.

While there is no algorithmic procedure for discovering ontology, this is not to say that nothing is known about the methodology of ontology acquisition. Chomsky could afford to make his pessimistic statement quoted above and stop at that because both Chomskian and post-Chomskian traditions sanction the use of descriptions of language fragments for arguing theoretical issues rather than the goal of comprehensive language description. Facing the latter task and mindful of Chomsky's admonition, we adhere to the "combination of weak methods" paradigm formulated by Wilks (Wilks *et al.,* 1992, p. 38) for knowledge acquisition in NLP. It does not follow that our ontologies are built in a haphazard and *ad hoc* fashion. They are formalized using a specially defined representation language and are organized around property inheritance hierarchies which facilitate (though, of course, do not guarantee) data consistency.

Reproducibility may indeed be a legitimate "goal of scientific experiments" (ibid). We believe, however, that our paradigm is not that of a scientific experiment thus defined. Just as it is possible to write any number of different but equally adequate grammars for the same set of linguistic data, there must be multiple but equally acceptable ontological models of a given world (cf. Hobbs, 1989, pp. 16-17). Thus, there is no single "true" or "correct" ontology waiting to be discovered, and our experiments are not similar to those in the natural sciences, which attempt to describe existing entities and phenomena. We deal with artificial constructs. We do not even have to claim the psychological reality of an ontology along the lines of "natural kinds" (in Kripke 1972), or its correspondence to mental structures in the human mind (e.g., Lakoff 1988). We can, however, offer an empirical verification of the ontologies we acquire. What we are after is a working tool, something that enables us to develop a lexicon that makes a quality translation possible. Our ontologies can be refuted if they fail to support NLP applications. For instance, when an MT system based on one of our ontologies systematically provides substandard translations, the quality of the ontology may be one of the culprits.

Opinions about the irreproducibility of ontology have been based largely on the claim that no two people would be able to agree on what any particular node or path in the ontology should look like. Our experience in practical medium- to large-scale ontology acquisition tends to refute this claim. The cited difficulties are usually connected with the learning curve of the team, during the initial stages of the task. As the learning curve flattens, a general consensus develops among the team members, which makes it possible to maintain and expand the ontology in a fairly uniform way. The Mikrokosmos project and its predecessors have employed over 30 people over the period of about 8 years in the ontological effort, and, in spite of the considerable normal staff turnover, the approach has remained fairly uniform. Obviously, we have developed a number of premises and methodologies,[1] some of which have found its way into our acquisition tools. One of our most challenging tasks is to explicate these premises and methodologies[2] and automate or semi-automate certain steps in ontology acquisition. In Mikrokosmos this process is well under way, but its description will form the subject for other papers.

4. Conclusions.

The semantics-based approach to translation is **not redundant** in that it provides the detail of knowledge necessary for resolving the problems posing problems for non-semantic approaches. Furthermore, by organizing our semantic primitives into an ontology, not only do we assure a consistent terminology, but also explain the meaning of these semantic primitives by establishing systematic paradigmatic relations among ontology elements. Finally, the ontology provides a rich search space for complex inferencing processes routinely needed to disambiguate figurative language as well as notoriously difficult phenomena such as nominal compounds in English and reference resolution in all languages.

As argued at length in Nirenburg and Goodman (1990), the charge concerning the inevitable language dependency of any ontology is based on a logically simple but unrealistic maximalist view of interlingua. We stress that the interlingua, as the ontology, is not a phenomenon waiting to be discovered but rather an artifact incrementally constructed with the goal of asymptotically approaching the ideal of a completely **language-neutral** ontology. The richer the set of ontological properties and primitives, the more expressive the text meaning representation, and the higher the chances of this mechanism being adequate for the description of a variety of lexical meanings across languages. The ultimate assessment of the language neutrality of ontology is empirical: if an ontology supports a very successful multilingual machine translation system, then it is *ipso facto* sufficiently neutral. The same argumentation, in fact, renders the claims of "upper case semantics" spurious. In any case, the Mikrokosmos ontology relies on neither English syntax nor lexical semantics for any of its reasoning.

We have addressed the issue of **reproducibility** of ontologies and concluded that, in all those cases when such an enterprise was necessary in any of our projects, we were able successfully to carry out this task.

Finally, the engineering objection to ontology-based approaches is, indeed, pertinent. We believe that the benefits of incorporating even an incomplete ontology in an application system far outweigh the risks and costs. The alternative, to abandon the capability for complex reasoning that ontologies support, *a priori* admits failure to resolve some of the common disambiguation tasks inherent in machine translation. To **ame-**

---

1. It is on the basis of one such methodology that we can systematically reduce dozens of word senses in the dictionaries to a manageable number of basic senses. Thus, by taking into account the ontological concepts to which the entries are related as well as the combination of constraints, we shrank the 54 word senses of *dejar* in the Collins Spanish - English Dictionary to the seven entries presented in the paper.
2. We have made some successful initial steps toward this goal (Attardo, 1994; Raskin, 1994; Mahesh, 1995).

liorate the engineering risks and costs, MT projects such as Mikrokosmos develop and deploy a plethora of increasingly automated tools to facilitate and validate the knowledge acquisition process.

Acknowledgments

Research on this paper was supported in part by Contract MDA904-92-C-518 from the U.S. Department of Defense. Many thanks to Evelyne Viegas for checking our Spanish. A longer, earlier version of this paper appeared in the working notes of the AAAI Spring 1995 Symposium on Lexical Semantics.

References

Atkins, B. T., J. Kegl, and B. Levin (1986). "Explicit and Implicit Information in Dictionaries", Lexicon Working Papers 12. Center for Cognitive Science, MIT, Cambridge MA.

Attardo, D. (1994). "MIKROKOSMOS TMR Builder's Guide", NMSU CRL, Internal Memo. URL: http://crl.nmsu.edu/users/mikro/Home.html

Bar Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers,* 1, pp. 91-163.

Bateman, J. A. (1993). "Ontology construction and natural language", in *Proceedings of the International Workshop on Formal Ontology.* Padua, Italy, pp. 83-93.

Boguraev, B. and E.J. Briscoe (1989). *Computational Lexicography for Natural Language Processing.* London: Longman.

Briscoe, E.J., A. Copestake, and B. Boguraev (1990). "Enjoy the paper: Lexical Semantics via Lexicology," in *Proceedings of COLING-90,* Helsinki.

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

Briscoe, T. (1994). Comments at Workshop on Lexical Semantics, University of Pennsylvania, Philadelphia, November.

Carlson, L. and S. Nirenburg (1990). "World Modeling for NLP." CMU CMT Technical Report 90-121.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge MA: MIT Press.

Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni, and E. Marinai (1992). "Multilingual Lexical Representation." ACQUILEX Working Paper 43.

Farwell, David, Louise Guthrie, and Yorick Wilks (1993). "Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System" in *Machine Translation* vol. 8:3, pp. 127-146.

Hale, K. and S. J. Keyser (1986). "A View from the Middle", Lexicon Working Papers 10. Center for Cognitive Science, MIT, Cambridge MA.

Hobbs, J. (1989). "World Knowledge and Word Meaning", in Y. Wilks, ed. *Theoretical Issues in Natural Language Processing.* Hillsdale NJ: Lawrence Erlbaum.

Jackendoff, Ray (1983). *Semantics and Cognition.* Cambridge MA: MIT Press.

Kegl, J. (1989). "The Boundary between Word Knowledge and World Knowledge", in Y. Wilks, ed. *Theoretical Issues in Natural Language Processing.* Hillsdale NJ: Lawrence Erlbaum.

Knight, Kevin (1993). "Building a large ontology for machine translation", in *Proceedings of the ARPA Human Language Technology Workshop,* Princeton NJ.

Kripke, S. (1972). "Naming and Necessity", in D. Davidson and G. Harman, eds. *Semantics of Natural Language,* Dordrecht: Reidel.

Lakoff, George (1988). "Cognitive Semantics" in *Meaning and Mental Representations,* Umberto Eco, Marco Santambrogio, and Patrizia Violi, eds. Bloomington IN: Indiana University Press.

Mahesh, K. (1995). "Guidelines for the MIKROKOSMOS Ontology Developer", NMSU CRL, Internal Memo. URL: http://crl.nmsu.edu/users/mikro/Home.html

McDermott, D. (1978). "Tarskian semantics, or no notation without denotation!" in *Cognitive Science,* vol 2:3.

Meyer, I., B. Onyshkevych and L. Carlson. (1990). "Lexicographic Principles and Design for Knowledge-Based Machine Translation". CMU CMT Technical Report 90-118.

Nirenburg, S., J. Carbonell, M. Tomita and K.Goodman. (1992). *Machine Translation: A Knowledge-Based Approach.* San Mateo, CA: Morgan Kaufmann.

Nirenburg, Sergei and Kenneth Goodman (1990). "Treatment of Meaning in MT Systems," *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language.* Linguistic Research Center, University of Texas at Austin.

Onyshkevych, B. and S. Nirenburg (1994). "The Lexicon in the Scheme of KBMT Things". NMSU CRL MCCS-94-277.

Pustejovsky, James (1991). "The Generative Lexicon" in *Computational Linguistics,* 17:4.

Pustejovsky, James, S. Bergler, and P. Anick (1993). "Lexical Semantic Techniques for Corpus Analysis" in *Computational Linguistics,* vol. 19:2.

Pustejovsky, James (forthcoming). "Linguistics Constraints on Type Coersion", in P. Saint-Dizier, E. Viegas, eds. *Computational Lexical Semantics.* Cambridge: Cambridge University Press.

Raskin, V. (1994). "Methodology for MIKROKOSMOS lexicon acquisition", NMSU CRL Internal Memo. URL: http://crl.nmsu.edu/users/mikro/Home.html

Wilks, Y., L. Guthrie, J. Guthrie, and J. Cowie (1992). "Combining weak methods in large-scale text processing," in P. Jacobs, ed. *Text-Based Intelligent Systems.* Hillsdale NJ: Lawrence Erlbaum.