WordNet : what about its linguistic relevancy?

Monique Slodzian

Centre de recherches en ingénierie multilingue
Institut national des langues et civilisations orientales (Paris, France)
Monique.Slodzian@inalco.fr

1 Preamble

Princeton WordNet and EuroWordNet belong to a group of projects which aim at providing general lexical resources on a large scale, at an acceptable cost, allowing reusability and common access for humans and computers. Such a practical purpose - how to satisfy the urgent demand for large, machine-tractable lexicons for various applications- had to find a suitable linguistic theoretical frame which, ideally, would answer the expectations of meaning acquisition from lexicons and texts and the constraints of implementation. As a matter of fact, the necessity of reducing drastically the complexity and multiplicity of linguistic facts leads to select one type of semantic paradigm, the one which restate the antique credo that the sign represents the concept and the concept represents the object or referent. And the cognitive semantics paradigm doesn't change the equation : the sign is given as linguistic and the meaning is given as conceptual, but we have the same one-to-one relationship concept vs. word.

But as soon as this naïve conception of meaning is not taken for granted, a question arises: to what extent does the reductionism imposed by the application and authorized by the theoretical frame have an effect on the relevancy of the results? Hence, we suggest that the interest of these projects in terms of natural language comprehension or generation does have heavy limitations. In other words, we assume that the underestimation of the relationship between theory and methodological issues has a price and that the claim for realism doesn't entirely justify theoretical naïvety. To account for what we call simplifications, let us give some examples :

- the isolated word is the basic unit in WordNet (the sign is supposed to have a pre-established meaning by itself or in isolated settings). Consequently, the syntagmatic properties are neglected; syntax, semantics, pragmatics are separated;
- the hierarchical structure of WordNet reinforces the adhesion to a linguistics dealing most exclusively with the signification of the sign (ignoring the textual sense problematic);
- the hypothesis that it is possible to build a general ontology representing the words (and their meanings) of a given language (i.e. English) rests on a logical conception of language which remains dubious from a linguistic point of view. In addition, assuming the universality of such an ontology as an interlingua and pretending that it is possible to merge word senses from different languages

inside this framework is neither a new nor a credible postulate. At least, it reveals an accultural, unhistorical and abstract account of meaning[1];

Obviously, the major theoretical concession allowing such assumptions lies in the epistemological presupposition that meaning doesn't go beyond the isolated sign (or isolated sentences). Unlike continental semantics, cognitive semantics puts that the signified may be assimilated with the concept. Considering language as an instrument of representation and meaning as pre-existing to words, they are founded to advocate that a general ontology or an interlingua constitute a "neutral format" adequate for processing natural language. The benefits of such a conception are evident for the formalization tasks.

Nevertheless, G. Miller and his associates are conscious of the dangers of oversimplifying linguistic facts. Compared with other projects like Mikrokosmos or Pangloss, WordNet objectives remains modest. The authors insist on the necessity to enlarge the stock of derived senses, to reach other phenomena via pragmatics, semantic constraint theories or script techniques. They are convinced that the WordNet model is able to produce relevant results for a large range of applications by solving subproblems in a realistic applied environment, which means that it isn't necessary to control totally the semantic problem[2] .

Put differently, they suggest that by borrowing from computer techniques the layered approach, one can postpone most complex problems which cannot be solved by a sign-based approach.

From this point of view, the theoretical issues remain non-fundamental for the application : any semantics is good as long as one holds on such a procedural vision of meaning. Beyond the WordNet case, it maybe useful to weigh up more carefully the prejudice in favour of a sign-based semantics, which, obviously, cannot suit all the tasks undertaken by the authors.

Prior to discussion, we shall give : a short outline of the two projects (WordNet 1.5 and EuroWordNet) in order to enter in a deeper way into their theoretical issues; a review of their performances, specially in information retrieval, since it constitutes the main application domain for both projects. To be as close as possible to the author's vision, we shall mainly refer to their collective books[3].

2 Brief Outline of WordNet 1.5 and EuroWordNet

As a rule, conceptual semantics proposes two alternative ways for processing linguistic data, either by linking the word-concept relationship or the hierarchical relationship (one concept, many words which act as synonyms).

2.1 The WordNet projects belong to the second trend since they stand for a hierarchical organization of the lexicon founded on psycholinguistic arguments. In fact, the authors state that arranging concepts in terms of class inclusion seems to capture an important principle of knowledge representation in human lexical memory.

In representing concepts by synonym sets, the WordNet projects come under both a thesaurus and an electronical dictionary conception. As Fellbaum puts it: *like a standard dictionary, WordNet includes not just single words, but also compound nouns(…)and collocations (…). Unlike a standard dictionary, WordNet does not take the word, or lexeme, as its elementary building block. Instead, WordNet resembles a thesaurus in that its units are concepts, lexicalized by one or more strings of letters, or word form. A group of words that can all refer to the same concepts is dubbed a synonym set, or synset*[4].

Each SYNSET represents a concept, which is often explained through a brief definition. In other words, the relationship "the word W refers to the concept C" shows itself in the fact that M belongs to the synset of C. As a result, two words are synonyms if they designate the same concept.

Let us note that the hybrid conception (thesaurus and/or dictionary) is indicative of some confusion on the status of synonymy, given sometimes as a conceptual relation, sometimes as a lexical relation. In our view, this is characteristic of a deeper problem in so far that conceptual semantics puts that the signified may be assimilated to the concept. As a practical consequence, instead of taking into account linguistic phenomena inside a concrete production, it aims at the representation of already normalized lexical contents. In so doing, WordNet sticks to lexicography tradition. Its major asset from this point of view is quantitative : it can offer more acceptions around a central signification. The main issue is to reduce polysemy by multiplying senses.

WordNet doesn't bring into question the fact that it deals with artefacts (the isolated sign , definitions, polysemy are artificial phenomena).

This conception explains the centrality of synonymy as well; nouns, verbs and adjectives are organized in large sets representing the underlying concept.

These sets are interconnected via semantic relations where hyperonymy/hyponymy is prevailing, given the predominance of nouns. The representational paradigm is more at ease with nouns than with adjectives and adverbs. Obviously, WordNet reflects this preference for nominal items, presumably labelling the concepts resulting from human perception.

Let us say for the moment that WordNet 1.5 presents itself as a successful hypermedia realization in the field of lexicography, enlarging and systematizing the data and making these data accessible for human or automatic applications. At the same time, it doesn't go beyond the lexicographic semantic conception. Maybe, the necessity to reinforce the overall architecture leads sometimes to intensify the artificial sides of traditional dictionaries.

2.2. EuroWordNet aims at constructing a large multilingual database with wordnets in several European languages (7 at least) structured along the same lines as the Princeton WordNet. The European wordnets are stored in a central lexical database system and each meaning is linked to the closest synset in WordNet 1.5. Thus, English becomes the parameter of all languages. Synsets linked to the same WordNet 1.5 synset are supposed to be equivalent. Equivalence relations between the synsets in different languages and WordNet 1.5 are made explicit in the Inter-Lingual Index (ILI), which is an unstructured list of meanings, taken from WordNet 1.5, where each ILI-record consists of a synset, an English definition giving the meaning.

To these definitions can be added domain labels (e.g. Sport, Military, Hospital, Traffic,…) or top concepts (e.g. Object and Substance, Location, Dynamic and Static,…).

Obviously, EuroWordNet reinforces the conceptual options of WordNet 1.5. By requiring an overall neutral ontology (Top Ontology), it goes a step further, making its semantical foundation more fragile. At the same time, it emphazises the philosophical options of Princeton WordNet : *words express concepts, and the lexicon is constrained by the kinds of concepts that are available to us by virtue of our perception, and interaction with the world around us. These limitations of our conceptual inventory may be innate or universal. The majority of lexicalized concepts are shared among languages, although most, if not all, languages have words for some concepts that are not lexicalized in other languages*[5] .

By now, EuroWordNet being available only in small fragments, a comparison between the different modules seems too early. Therefore, we shall come back to EuroWordNet principles in section 4 only, in the general discussion.

3 Applications in information retrieval

WordNet and EuroWordNet are designed for supporting a large range of applications, including fact extraction and summarization, automated indexing and machine translation. At a first stage, the most essential function is helping the user with query formulation through synonym relationships between words and hierarchical relationships (plus other relationships like meronymy and inference) between concepts.

The authors assume that through these relationships, it is possible to enhance access to collections of texts. Synonyms can substitute for key words, and hierarchical relationships (such as hyponym/hyperonym) being reciprocal, it is possible to maximize word supply with descendants plus parents.

In a paper called *Using WordNet for text retrieval*[6], E. Voorhees describes in detail two types of investigations. In the first one, WordNet synsets, as concepts opposed to words, are used to represent the content of documents. In the second one, WordNet is the source of words that are added to the user's query.

3.1. The first investigation comes under concept matching techniques. As far as synsets are considered as concepts, WordNet structure has been used for several experiments in order to calculate the similarity between a query and a document. Such an approach has been implemented for disambiguation in retrieval procedures. Vorhees gives a penetrating comment on the results of the MED collection experiment, showing how problems occur : *the query, requesting documents on separation anxiety in infants and pre-school children, retrieves 7 relevant documents in the top 15 for the standard run and only 1 relevant document 15 for the "110 run". The problem is selecting the sense of separation in the query. WordNet contains eight senses of the noun separation. With so few other words to use in the disambiguation processing of the query, the*

process selects a sense of separation that is never selected for any document, and retrieval performance suffers accordingly.
We shall suggest that the major problem here is related to the fact that synsets are built abstractly, according to a theory of meaning based on sign-based linguistic. The attempts to make up for the irrelevancy of the general meaning thesis show their limits. As a matter of fact, the "hood" principle developed by G. Miller illustrates this kind of strategy : *a hood (…) is an area in WordNet in which a string is unambiguous".* Other artefacts have been imagined, like helping filtering the irrelevant senses by domain tagging. Obviously, this is not convincing for general lexicon. As a result, as indicated in the Vorhees paper, *the problem of missing some good matches remains even if the sense resolution procedure works perfectly.*

As a good illustration of the problem, let us quote the following example : *the nouns nail, hammer, and carpenter are all good hints that the intended sense of board is the "lumber" sense. However, within WordNet a nail is a fastener, which in turn is a device, so nail would help select the "control panel" sense of board. Similarly, a hammer is a tool, which is an implement, which is an article of commerce, so hammer would help select the "dining table" sense of board. Finally, a carpenter is a worker, which is a person, which is both an agent and a life form, which are both things. Thus, carpenter would not help select any sense of board. This analysis indicates that specialization/generalization relations are unlikely to contain sufficient information to choose among fine sense distinctions.* Such a deficiency due to the ontological approach will be discussed in next section.

3.2. In the second experiment, WordNet is used as a word source which will help query expansion. *Instead of trying to select a single synset to represent a concept -an error-prone process where a single error can severely degrade retrieval effectiveness- use WordNet as a source of additional words to supplement the user's query. The goal is to add to a query all the different words that can be used to express the query's concepts.*

Most obviously, when the selection of synsets is handmade (with a control of good starting concepts), given a synset, there is a large choice of words to add to the query. However, as Vorhees says : *a poor choice can be worse than not expanding.*
As for the possibility of a completely automatic procedure to expand short queries, the results seem poor. In fact, we are faced here with the same question than in 3.1, concerning the relevancy of the ontological approach.
If the inability to automatically resolve polysemy limits the benefits of using WordNet, however some experiments seem to show that interesting results can be obtained in extraction procedures to help coarse-level analysis techniques. Let us quote the "Harvard as an Institution" Experiment developed by Grefenstette[7] whose conclusion is : *it seems that by taking a WordNet synset as a filter through a large text, we have extracted a rather coherent corpus.*

As a conclusion, WordNet certainly is a successful hypermedia system which maximalizes the Anglo-American lexicography heritage by its size (100 000 synsets in WordNet 1.6), systematicity, accessibility and portability. As a lexical database, it offers broad coverage of general lexicon in English. WordNet has been employed as a resource for many applications in information retrieval. It doesn't seem to be entirely suitable for these applications. Is it possible to improve these deficiencies by adding more lexical relations and more word senses?

4 Discussion

4.1 The isolated word principle

In dealing with simple lexical items, WordNet has little chance to find their preferential contexts. The question of the minimal semantic unit is raised by WordNet. Thus, when it is said that the French complex expression *emploi du temps* is a last resort for *schedule*[8], or that the concept of "pet" is absent from several languages, we suspect the authors to be influenced by the one word-one concept equation.

And the long considerations in Miller's article about the perfect paradigm of "robin" leads us to the following reflexion : by taking the isolated word as "a primary storage site", WordNet is condemned to stick to a taxonomic vision, which means dealing with content classes cut from real language mechanisms, as they occur in oral and written communication.

As far as the word is considered as the privileged entry-point, theorists are tempted to justify this convenient choice by philosophic or psycholinguistic reasons.

As a matter of fact, the psycholinguistic argument that isolated words constitute the most natural access to lexicon has been seriously challenged[9] since it has been proved to be easier, on the contrary, to access words in context, due to the semantic and thematic information brought by the text. A different view on polysemy, as an artefact coming from sign-based linguistics, arises from those considerations. For a challenging account on the isolated word paradigm and the limits of linguistic compositionality see Catherine L. Harris[10].

4.2 The pre-existing meaning credo

The ontological approach of meaning doesn't allow to represent the lexical signified in their real functioning. In terms of relevancy, it means that in real texts the signified interplay, often leaving unactuated the general lexical relations and, eventually, the predicted senses within WordNet. For example, the English expression, *apple of one's eye*, has nothing to do with the linear taxonomy of "fruit". For an overall exposure of textual semantics issues, see Rastier[11].

4.3 The paradigmatic/syntagmatic issue

Vorhees suggests (p 301) *that the paradigmatic relations contained within WordNet together with the text to be disambiguated do not supply the information required for this sense resolution task*

In our view, such a criticism is somewhat sterile because too general. Moreover, it nurtures the hope that via pragmatics, it will be possible to add more context, ad infinitum. We already have discussed the point.

As Rastier notes, the paradigmatic option of cognitive semantics leads to a strict semasiological approach, *resting upon the prelinguistic prejudice, born from the philosophy of language, that to one word corresponds one signified; and as this is obviously not the case, one must find for it a preferential signified, or more precisely a basic conceptualization (counterpart to the literal sense in vericonditional semantics)*[12].

Indeed, capturing and adding new senses stand as the main focus of WordNet as well as of traditional lexicography. But from the conclusions on the benefits of using WordNet in information retrieval, we know that too many word senses don't help the resolution task, on the contrary.

4.4. The ontological paradigm

The hypothesis of a general ontology describing the world's objects in a more or less symbolic way comes under a logical conception of language which has been discussed for two centuries[13]. It would be presumptuous to bring new comments about such a traditional debate.

Let us observe that WordNet deals with the concept of primitive in a very naïve way. Language items taken as primitives (e.g. GO, PATH, etc.) fall under the question of confusing primitives and linguistic lexemes drawn from a particular language (as if by chance, English). As Langacker argues, why not MOVE instead of GO? Developing further his criticism, he observes that *Katz, Jackendoff and Fillmore all advocated "Motion" as a semantic primitive, as well as "Time" and "Location", without addressing the possibility that motion can be defined as the change of location over time*[14].

It is worth noting that WordNet authors adopt without much discussion Lyons propositions. For a number of linguists, Lyons ontological assumptions are *unashamedly those of "naïve realism", such as that the external world contains a number of individual persons, more or less discrete physical objects and places (or spaces)* [15].

Compared to the meaning-text model proposed by Mel'cuk et al., the WordNet ontological engagement seems to be a regression since the former model doesn't postulate any metalanguage and holds on notations describing semantic, syntactic and collocational properties relevant for lexical sections belonging to given languages. *According to Mel'cuk there is no reason to believe, on the assumptions of the MTM, that the elementary meaning of different languages correspond, that is, there are any substantive semantic at all*[16].

As a conclusion, we shall draw attention to the fact that a large part of NLP works claims the feasibility of disambiguating natural language via general ontologies. Some of those projects even go beyond WordNet ambitions since they explicitly aim at defining a methodology for representing the meaning of natural language texts. Their strategy rests upon the same hypothesis: a drastic simplification of meaning mechanisms and a large use of pragmatics, that is a mixture of micro-theories. Such an issue, while legitimized both by the lexicographic tradition and practical considerations, is jeopardized by its metaphysical premises.

By claiming that *"there is a level of concepts (the ontological level) that have a special status in our conceptual system and that there is a structural constraint on these concepts (they form a strict hierarchy)*[17], the Keil-Sommers proposal has arouse excessive enthusiasm in the NLP and A.I. communities. Maybe, it is time to review some certitude on behalf of realism since linguistic relevancy is at stake.

Firstly, the theoretical turn will hold in a radical rethinking of what a word sense is. We claim that it can be defined only by other linguistic units and not by a concept. This presumes usage in context. In our view, only a textual differential semantics is able to cope with those preconditions. As a consequence, a regional ontology paradigm will substitute for a general one. Prior to its building, the task and the domain corpus ought to be defined concretely. Methodological frames are being experimented for knowledge engineering[18].

[1] See Rastier, F., (1999), Cognitive semantics and diachronic semantics : the values and evolution of classes, in *Historical Semantics and Cognition*, Andreas Blank and Peter Koch ed., n°13, Mouton de Gruyter, Berlin-NY.

[2] Vossen, Piek, EuroWordNet : Construction d'une base de données multilingue autour de réseaux de mots pour les langues européennes, *Lettre d'information d'ELRA*, février 1998, p.7

[3] *WordNet, An Electronic Lexical Database*, Christiane Fellbaum ed., (1998), The MIT Press, Cambridge, Mass.

EuroWordNet, A multilingual Database with Lexical Semantic Networks, Piek Vossen ed., (1998), Kluwer Academic Publishers, Netherlands.

[4] Fellbaum, C., (1998), *EuroWordNet* (ibidem), p.210.

[5] Fellbaum, C., *WordNet* (ibidem), p.8

[6] Voorhees, E., *WordNet*, ibidem, pp 285-303

[7] Grefenstette, G., 1994, *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Boston/London/Dordrecht. Pp 116-123.

[8] Fellbaum, C., *WordNet*, ibidem, p.211

[9] Van Patten &Koutas, (1991),in *Understanding word and sentence*, Simpson., G. ed., Amsterdam, North Holland (pp 129-184)

[10] Harris, C.,L., (1994) in *Continuity in linguistic semantics*, C.Fuchs &B.Victorri ed., Benjamins, Amsterdam, pp 205-225

[11] Rastier, F., (1991*), Sémantique et recherches cognitives*, Paris, PUF

[12] Rastier, F. (1999)

[13] Eco,U., (1987), *The Open Work*, Cambridge, Mass., Harvard University Press

[14] Langacker, R.W.,(1987), *Foundations of Cognitive Grammar*, Stanford Univ. Press, vol.1, P167.

[15] Goddard,C. (1994), Semantic Theory and semantic Universals in *Semantic and Lexical Universals*, Goddard and Wierzbicka ed, Benjamins

[16] Goddard, ibidem, p.20

[17] Keil, F.C. (1979*) Semantic and conceptual development : An ontological perspective*. Cambridge, MA, Harvard University Press

[18] Bachimont, B., (2000), Conception et réalisation d'ontologies en ingénierie des connaissances, in *Ingénierie des connaissances, Evolutions récentes et nouveaux défis*, Charlet/Zacklad/Kassel/Bourigault ed., Paris, Eyrolles