EuroWordNet: A Multilingual Database with Lexical Semantic Networks

Piek Vossen (editor)
(University of Amsterdam)

[Reprinted from *Computers and the Humanities*, 32(2–3), 1998]
Dordrecht: Kluwer Academic Publishers, 1998, 179 pp; hardbound, ISBN 0-7923-5295-5, Dfl 170.00, $92.00, £58.00

Reviewed by
Graeme Hirst
University of Toronto

WordNet, the on-line English thesaurus and lexical database developed at Princeton University by George Miller and his colleagues (Fellbaum 1998), has proved to be an extremely important resource used in much research in computational linguistics where lexical knowledge of English is required. The goal of the EuroWordNet project is to create similar wordnets for other languages of Europe. The initial four languages are Dutch (at the University of Amsterdam), Italian (CNR, Pisa), Spanish (Fundacíon Universidad Empresa), and English (University of Sheffield, adapting the original WordNet); later Czech, Estonian, German, and French will be added. The results of the project will be publicly available.[1]

Like the original Princeton WordNet, the new wordnets—that's now a generic term—are hierarchies in which each node is a **synset**: a word sense, with which one or more synonymous words or phrases is associated. The synsets are connected by relations such as hyponymy, meronymy, and antonymy. However, some improvements have been made to the original design of WordNet. New relationships, including relationships across parts of speech, have been introduced. For example, the verb *adorn* and the noun *adornment* are related by XPOS_NEAR_SYNONYMY; hyponymy and hyperonymy[2] across parts of speech are also permitted. Semantic roles of verbs are marked; for example, the noun *student* is related to the verb *teach* by ROLE_PATIENT; the inverse relationship is called INVOLVED_PATIENT. Another new relationship, both within and across parts of speech, is causality, which may further be marked as intentional or nonfactive; for example, *to redden* CAUSES *red*; *to search* CAUSES (nonfactive, intentional) *to find*. Meronymy is much more fine-grained than in Princeton WordNet, with a number of new kinds of part-whole relationships.

The most important new development, however, is multilinguality: the use of a common framework to build the individual wordnets and integrate them in a single database in which an **inter-lingual-index** (ILI) connects the synsets that are "equivalent" in the different languages. EuroWordNet thus becomes a multilingual lexicon and thesaurus that could be used in applications such as multilingual text retrieval and (rather basic) lexical transfer in machine translation. The project has sought to

1 Distribution and licensing will be carried out by the European Language Resources Association, http://www.icp.inpg.fr/ELRA. The home page for the project itself is http://www.hum.uva.nl/~ewn.
2 EuroWordNet has adopted the single term *hyperonymy* (with an *o*) for both nouns and verbs, in preference to Princeton's *hypernymy* of nouns and *troponymy* of verbs.

carefully respect the differing conceptualizations and lexicalizations of each individual language while creating a language-independent framework for the multilingual database.

A key element of multilinguality was ensuring that the conceptual core of each of the wordnets would be similar in its coverage. A set of 1,024 **base concepts** common to all the languages (though not necessarily lexicalized in all of them) was developed iteratively, starting with an algorithmic identification of potential base concepts in each language, merging the results by selecting those proposed by at least two of the four languages, and carefully tuning the results. The base concepts include *increase, steal, health,* and *cause pain.* A language-independent **top ontology** was then built from 63 very abstract **top concepts**; these are used as semantic features to classify the base concepts.[3]

The inter-lingual-index serves to link "equivalent" synsets in the various wordnets. The synsets of Princeton (English) WordNet were used as the initial set of index records, as this wordnet is a particularly fine-grained one. New records can be added where necessary. Cross-linguistic equivalence is generally taken as synonymy, denoted EQ_SYNONYMY; but where a synset in one language has no unique direct equivalent in the ILI, it may be linked to several ILI records simultaneously as an EQ_NEAR_SYNONYM, or linked to a record by EQ_HAS_HYPERONYM or EQ_HAS_HYPONYM if it is more specific or less specific, respectively, than any ILI record.

The book *EuroWordNet* is a collection of six papers on the EuroWordNet project by editor Vossen and his colleagues.[4] Five of the papers describe the structure of the EuroWordNet database and the methods employed to construct it, and the sixth speculates on the possible benefits of using EuroWordNet in cross-language text retrieval. In addition, presumably for completeness, a short seventh paper by Christiane Fellbaum outlines the original Princeton English WordNet.

The papers were all written in late 1997 (or perhaps earlier) while the project was still under way and the wordnets were not complete. So while all the methodological and design decisions had been made, and could be usefully reported to the research community, the EuroWordNet database itself was far from complete and little in the way of evaluation or application could have been carried out. Consequently, the paper on applying EuroWordNet in text retrieval is largely speculative and sometimes written in the future tense—indeed, it describes itself as a "proposal"—and might better have been reserved for publication at a time when the project had actually been performed. The paper on the ILI also often lapses into the future tense.

The volume was previously published as a special issue of the journal *Computers and the Humanities,* and is paginated both as a book and as a journal fascicle (which means that page references in the overlap of the two ranges are ambiguous). It differs from most such special issues in being tightly focused on a single project, with frequent cross-references between the papers. In fact, one might rather think of it as a book that was prepublished as a journal—certainly, many readers will approach it as a unitary

3 While one understands the spatial metaphors inherent in the individual elements of the EuroWordNet terminology, they are confusing when combined, and one ends up wondering just how a wordnet is supposed to be oriented. In particular, if top concepts are indeed metaphorically at the top, then base concepts must be metaphorically in the middle, with all other concepts hanging below them; but on the other hand, if base concepts are metaphorically at the base, then other concepts must be metaphorically on top of them *except* for the top concepts, which are somehow underneath them.

4 Except for Vossen's introduction, all of the papers have many authors and many of the authors are involved in several of the papers. The complete set of authors is: Geert Adriaens, Antonietta Alonge, Francesca Bertagna, Laura Bloksma, Nicoletta Calzolari, Irene Castellon, Salvador Climent, Pedro Díez-Orzas, Julio Gonzalo, Elisabetta Marinai, Antonia Marti, Carol Peters, Wim Peters, German Rigau, Horacio Rodríguez, Adriana Roventini, Felisa Verdejo, and Piek Vossen.

work, to be read from start to finish—and it is its value in this regard that merits this review. However, this point also serves to make the absence of an index, which should be found even in journal-issue reprints, that much less forgivable. And tables of reference material, such as a glossary of terms and a complete list of the EuroWordNet lexical relationships, are given as appendices to individual papers, not at the back of the book, so the reader has to either remember which paper has which reference material or perform a search.

The quality of the writing and editing are reasonable, the notation is mostly consistent, and the book is largely free of typos. A couple of exceptions: On page 25, the word *finger* should be *Barbary ape*; on page 89, the underlining promised in example 8 is missing. While the papers are fairly well cross-referenced, they remain independent papers rather than interdependent chapters of a book; the need for each to stand alone as a journal paper leads to some annoying redundancy for the reader who approaches the work as a linear unit. The final paper, for example, begins by defining EuroWordNet as if the reader had never heard of it before. The papers also vary widely in the emphasis that they give to the methodology and process of designing and building the database relative to the emphasis given to the actual resulting design and content. The former is interesting and needs to be documented, but the latter is what most readers will want to concentrate on; yet the reader who turns to the section on base concepts (p. 52ff) will find a long description of the method by which the set of base concepts was derived but very few examples of actual base concepts (fortunately, they can be found in the section on the top ontology (p. 64ff)). And it is unclear why the papers are in the order that they are; for example, the paper on the inter-lingual-index, which seems to belong in about the middle of the book, appears only at the end, after the paper on applying EuroWordNet to cross-language text retrieval, which presupposes knowledge of the inter-lingual-index.

It is unfortunate, therefore, that this book is likely to become the definitive reference on EuroWordNet. It has preempted the publication of a better-organized, fully integrated, properly indexed book. Nonetheless, there is much of interest in *EuroWordNet* for the computational lexical semanticist, and much in EuroWordNet that will be of immediate practical use for anyone requiring lexical resources in any of the languages covered.

Reference

Fellbaum, Christiane (editor). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Graeme Hirst has used Princeton WordNet in research on measures of semantic distance and on real-word spelling-error correction. Hirst's address is Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4; e-mail: gh@cs.toronto.edu; URL: http://www.cs.toronto.edu/~gh