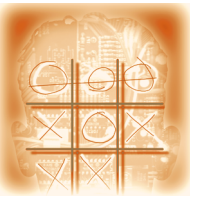
WordNet: A Lexical Database for English

George A. Miller

This database links English nouns, verbs, adjectives, and adverbs to sets of synonyms that are in turn linked through semantic relations that determine word definitions.

Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and machine-readable dictionaries are now widely available. But dictionary entries evolved for the convenience of human readers, not for machines. WordNet[1] provides a more effective combination of traditional lexicographic information and modern computing. WordNet is an online lexical database designed for use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept. Semantic relations link the synonym sets [4].



Language Definitions

We define the vocabulary of a language as a set W of pairs (f,s), where a form f is a string over a finite alphabet, and a sense s is an element from a given set of meanings. Forms can be utterances composed of a string of phonemes or inscriptions composed of a string of characters. Each form with a sense in a language is called a *word* in that language. A *dictionary* is an alphabetical list of words. A word that has more than one sense is *polysemous;* two words that share at least one sense in common are said to be *synonymous.*

A word's usage is the set C of linguistic contexts in which the word can be used. The syntax of the language partitions C into *syntactic* categories. Words that occur in the subset N are nouns, words that occur in the subset V are verbs, and so on. Within each category of syntactic contexts are further categories of *semantic* contexts—the set of contexts in which a particular f can be used to express a particular s.

The morphology of the language is defined in terms of a set M of relations between word forms. For example, the morphology of English is partitioned into *inflectional, derivational,* and *compound* morphological relations. Finally, the lexical semantics of the language is defined in terms of a set S of relations between word senses. The semantic relations into which a word enters determine the definition of that word.

[1] WordNet is a registered trademark of Princeton University, available by anonymous ftp from clarity.princeton.edu

More Than 166,000 Word Form and Sense Pairs

In WordNet, a form is represented by a string of ASCII characters, and a sense is represented by the set of (one or more) synonyms that have that sense. WordNet contains more than 118,000 different word forms and more than 90,000 different word senses, or more than 166,000 (*f,s*) pairs. Approximately 17% of the words in WordNet are polysemous; approximately 40% have one or more synonyms.

WordNet respects the syntactic categories *noun, verb, adjective,* and *adverb*—the so-called open-class words (see Table 1). For example, word forms like "back,'' "right,'' or "well'' are interpreted as nouns in some linguistic contexts, as verbs in other contexts, and as adjectives or adverbs in other contexts; each is entered separately into WordNet. It is assumed that the closed-class categories of English—some 300 prepositions, pronouns, and determiners—play an important role in any parsing system; they are given no semantic explication in WordNet.

Inflectional morphology for each syntactic category is accommodated by the interface to the WordNet database. For example, if information is requested for "went," the system will return what it knows about the verb "go." On the other hand, derivational and compound morphology are entered into the database without explicit recognition of morphological relations. For example, "interpret," "interpreter,'' "misinterpret," "interpretation," "reinterpretation," "interpretive," "interpretative," and "interpretive dancing" are all distinct words in WordNet.

A much larger variety of *semantic relations* can be defined between words and between word senses than are incorporated into WordNet. The semantic relations in WordNet [6] were chosen because they apply broadly throughout English and because they are familiar—a user need not have advanced training in linguistics to understand them. They are shown in Table 1. WordNet includes the following semantic relations:

- *Synonymy* is WordNet's basic relation, because WordNet uses sets of synonyms (*synsets*) to represent word senses. Synonymy (*syn* same, *onyma* name) is a symmetric relation between word forms.
- *Antonymy* (opposing-name) is also a symmetric semantic relation between word forms, especially important in organizing the meanings of adjectives and adverbs.
- *Hyponymy* (sub-name) and its inverse, *hypernymy* (super-name), are transitive relations between synsets. Because there is usually only one hypernym, this semantic relation organizes the meanings of nouns into a hierarchical structure.
- *Meronymy* (part-name) and its inverse, *holonymy* (whole-name), are complex semantic relations. WordNet distinguishes *component* parts, *substantive* parts, and *member* parts.
- *Troponymy* (manner-name) is for verbs what hyponymy is for nouns, although the resulting hierarchies are much shallower.
- *Entailment* relations between verbs are also coded in WordNet.

Table 1. Semantic Relations in WordNet

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: ***N*** = *Nouns* ***Aj*** = *Adjectives* ***V*** = *Verbs* ***Av*** = *Adverbs*

Each of these semantic relations is represented by *pointers* between word forms or between synsets. More than 116,000 pointers represent semantic relations between WordNet words and word senses.

Relational theories of lexical semantics hold that any word can be defined in terms of the other words to which it is related. For example, a definition of the compound noun "sugar maple" might start with its hypernym, "A sugar maple is a maple that . . . ," followed by a relative clause based on meronymy or other semantic relations that specify how sugar maples differ from other kinds of maples. However, not enough semantic relations are encoded into WordNet to support such constructions. Following standard lexicographic practice, definitional glosses are included in most synsets along with the synonyms that represent the sense.

An XWindows interface to WordNet allows a user to enter a word form and to choose a pull-down menu for the appropriate syntactic category. The menus provide access to the semantic relations that have been coded into WordNet for that word. For example, if "leaves" is entered, a noun menu for "leaf" and a verb menu for "leave" are available. The noun menu includes options for synonyms, hyponyms, hypernyms, sisters, meronyms, and holonyms for "leaf"; no antonyms for "leaf" are available. If synonyms of the noun are requested, the window display three synsets, along with their immediate hypernyms:

- Leaf, leafage, foliage—the main organ of photosynthesis in higher plants; plant organ—a func-

tional and structural unit of a plant
- Leaf, folio—a sheet of written or printed matter; sheet, piece of paper, sheet of paper used for writing or printing
- Leaf—hinged or detachable flat section, as of a table or door; section, segment—one of several parts that fit with others to constitute an object

Other choices from the menus would result in other displays of lexical information. A command line interface to the database is also available.

Contextual Representations

Polysemy is a major barrier for many systems that accept natural language input. For example, two different senses of an English word form may translate into totally different words in another language. Therefore, systems for machine translation should be able to determine which sense the author had in mind. In information retrieval, a query intended to elicit material relevant to one sense of a polysemous word may elicit unwanted material relevant to other senses of that word. For example, in computer-assisted instruction, a student asking the meaning of a word should be given its meaning in that context, not a list of alternative senses from which to pick.

WordNet lists the alternatives from which choices must be made. WordNet would be much more useful if it incorporated the means for determining appropriate senses, allowing the program to evaluate the contexts in which words are used. This unmet requirement is a goal for further development.

Choosing between alternative senses of a polysemous word is a matter of distinguishing between different sets of linguistic contexts in which the word form can be used to express the word sense. People are quite skillful in making such distinctions [1]. For instance, people who are told, "He nailed the board across the window," do not notice that "board" is polysemous. Only one sense of "board" (or of "nail") reaches conscious awareness. How people make such distinctions is not well understood.

An algorithm for sense identification must distinguish sets of linguistic contexts, raising the question of how much context is required. The limits of a linguistic context can be defined arbitrarily, but we prefer to define it in terms of sentences. That is to say, two words co-occur in the same context if they occur in the same sentence. Given this definition, sense identification is a matter of distinguishing among sets of sentential contexts. Miller and Charles [5] proposed that a contextual representation associated with each sense characterizes sentential contexts in which a given word can be used to express that sense. Therefore, the empirical problem is to determine what contextual representations should look like.

The usual way computational linguists have coped with polysemy has been to limit the domain of discourse. For example, the noun "flight" has eight senses in WordNet, but when the domain of discourse is limited to air travel, only one of the eight is likely to occur. Therefore, *topical context* (the vocabulary used to discuss a well-defined topic) provides some of the information needed for a contextual representation. However, results obtained by Leacock, Towell, and Voorhees [3] indicate that topical context can identify senses correctly only about 80% of the time. People seem to make more use of local context—the exact sequence of words immediately preceding and following the polysemous word. How best to characterize the contexts associated with word senses remains an open question.

Semantic concordances are being prepared to provide a basis for empirical studies of sense identification [7]. A semantic concordance is a textual corpus and a lexicon combined so that every substantive word in the text is linked to its appropriate sense in the lexicon. For example, words in passages from the Brown Corpus [2] are linked to their senses in WordNet, providing a test bed for proposed sense-identification systems. However, this semantic concordance is still too small to provide representative samples of contexts indicative of the different senses of polysemous words. Supplementing WordNet with a textual database remains an ongoing project.

Acknowledgments

Preparation of this article was supported in part by grants from the Office of Naval Research, the Advanced Research Projects Agency (Information and Technology Office), the Linguistic Data Consortium, and the James S. McDonnell Foundation.

References

1. Charles, W. G. The categorization of sentential contexts. *J. Psycholinguistic Res. 17*, 5 (Sept. 1988), 403–411.
2. Francis, W. N., and Kucera, H. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, Mass., 1982.
3. Leacock, C., Towell, G., and Voorhees, E. M. Towards building contextual representations of word senses using statistical models. In *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text* (Columbus, Ohio, June 21) ACL/SIGLEX, 1993, pp. 10–20.
4. Miller, G. A., Ed. WordNet: An on-line lexical database. *International Journal of Lexicography 3*, 4 (Winter 1990), 235–312.
5. Miller, G. A,. and Charles, W. G. Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*, 1 (Feb. 1991), 1–28.
6. Miller, G. A., and Fellbaum, C. Semantic networks of english. In B. Levin and , S. Pinker Eds. *Lexical and Conceptual Semantics*. Blackwell, Cambridge and Oxford, England, 1992, pp. 197–229.
7. Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop* (Princeton, NJ, March 21–23). 1993, pp. 303–308.

About the Author:
GEORGE A. MILLER is the James S. McDonnell Distinguished University Professor of Psychology Emeritus at the Cognitive Science Laboratory, Princeton University. He wrote *The Science of Words*, published in 1991 by the Scientific American Library. A book describing WordNet and its applications is scheduled for publication by The MIT Press in 1996.

Author's Present Adress: Cognitive Science Laboratory, Princeton University, Princeton, NJ 08544-1010; email: geo@clarity.princeton.edu