THE WEB OF WORDS

Brian Hayes

A reprint from

American Scientist

the magazine of Sigma Xi, the Scientific Research Society

THE WEB OF WORDS

Brian Hayes

A dictionary is more than a book of definitions; it is an index to a language, imposing an order on our inventory of words. Likewise a thesaurus is a table of contents, which takes the same stock of words but organizes them thematically rather than alphabetically. Both kinds of books reveal something about the underlying structure of the lexicon (the set of all words that make up a language). That structure is what mathematicians call a graph—a collection of "nodes" connected by "edges," usually drawn as a web of dots and lines.

When a language is viewed as a mathematical graph, the nodes are words (or sets of words), and the edges are relations between them. Any dictionary will help you to walk from node to node through the graph. For example, in defining the word *elegant*, the *American Heritage Dictionary* offers *delicate* as a synonym; on looking up *delicate*, you find *dainty* among the meanings listed; *dainty* in turn leads you to the entry for *exquisite*; and among the meanings of *exquisite* is *elegant* again. In this way you trace out one of many loops, or cycles, within the graph defined by this particular dictionary.

Exploring small regions of a lexical graph is a familiar process; you do it mentally whenever you grope for the right word. But trying to construct a graph for an entire language is another matter entirely. English has well over 100,000 words, and they are related to one another in dozens or perhaps hundreds of ways. Finding and recording all the connections is a task on the same scale as compiling a large dictionary. Furthermore, it has to be done with great precision and consistency, because the goal is to create a mathematical structure in which the relations between words are so explicit that the graph can be explored and manipulated algorithmically.

The construction of a lexical graph for English has been under way for almost 15 years in a project called WordNet, which now includes some 168,000 words and 345,000 relations among them.

Brian Hayes is a former editor of American Scientist. *From January to June 1999 he is Journalist in Residence at the Mathematical Sciences Research Institute in Berkeley. Internet: bhayes@amsci.org.*

WordNet is the work of George A. Miller and his colleagues in the Cognitive Science Laboratory at Princeton University. (Other contributors are Christiane Fellbaum, Randee I. Tengi and the late Katherine J. Miller.) A book describing WordNet and its applications has recently been published, and the database that defines the lexical graph is available via the Internet and on a CD-ROM, along with software for browsing the graph.

Synonyms, Hypernyms and Other Nyms

In WordNet the emphasis is less on words than on the relations between words. And among the various relations defined in the lexical graph, the most fundamental is synonymy. Words that mean more or less the same thing are grouped into synonym sets, or synsets, much as they are in a thesaurus. The synsets (rather than individual words) then become the basic nodes of the graph.

A finicky logophile might well argue that true synonyms do not exist—that no two words are exactly equivalent. The compilers of WordNet take a pragmatic position on this issue. They classify words as synonyms if there is some class of sentences where one word can take the place of another without substantially altering the meaning. Thus *yell*, *shout* and *holler* have distinguishable nuances of meaning, but in many sentences the words are interchangeable.

Most words have multiple meanings, and synonymy is really a relation between the individual senses of the words. The adjective *light* is an approximate synonym both of *weightless* and of *pale*, but *weightless* and *pale* are not themselves synonyms. Each sense of *light* has to be given its own synset. (In the current version of WordNet *light* has 26 adjective senses as well as 15 noun senses, six verb senses and one adverb sense.)

Synonymy is the glue that binds WordNet together. Nouns, verbs, adjectives and adverbs all have synonyms. Nevertheless, the structure of the lexical graph comes mainly from other kinds of relations, which are somewhat different for each part of speech.

For nouns the most important relations are hypernymy and hyponymy, which organize concepts into a treelike hierarchy. Hypernyms and hyponyms embody the "is-a" or "is-a-kind-of"

relation: A horse is a mammal is an animal is an organism. Thus the word *horse* (or the sense of *horse* referring to a hoofed quadruped) is a hyponym of *mammal*, which in turn is a hyponym of *animal*, and so on. Hypernyms describe the same relation seen from the other end of the telescope: *Animal* is a hypernym of *mammal* and also of *reptile*, *bird*, *fish*, etc. The biological examples are apt here, since the construction of such taxonomic hierarchies is a specialty of the life sciences. In WordNet the sequence of hypernyms for *horse* captures much of the phylogenetic detail a biologist would want to see recorded: horse ⇒ equine ⇒ odd-toed ungulate ⇒ ungulate ⇒ placental mammal ⇒ mammal ⇒ vertebrate ⇒ chordate ⇒ animal ⇒ organism ⇒ entity.

Another relation among nouns is summed up in the phrase "has a" rather than "is a." This is the relation between parts and wholes; words that represent parts or members are called meronyms, and those that denote wholes or groups are holonyms. The distinction between "is a" and "has a" can be subtle, but there are examples that make it clear. Consider the noun *meal:* Its hyponyms are words such as *breakfast*, *lunch* and *dinner*, but its meronyms are *appetizer*, *salad*, *dessert* and so on.

The same kind of treelike organization can be imposed on verbs, although verb trees tend to be somewhat stunted and shrubby compared with noun trees. The relation analogous to hyponymy in nouns has been dubbed troponymy in verbs. For example, the intransitive verb *walk* is a troponym of *go* or *move* or *locomote*; in other words, walking is a way of moving. And *walk* in turn has troponyms such as *shuffle*, *amble*, *swagger* and *march*. For some verbs there is also a relation analogous to meronymy in nouns, defin-
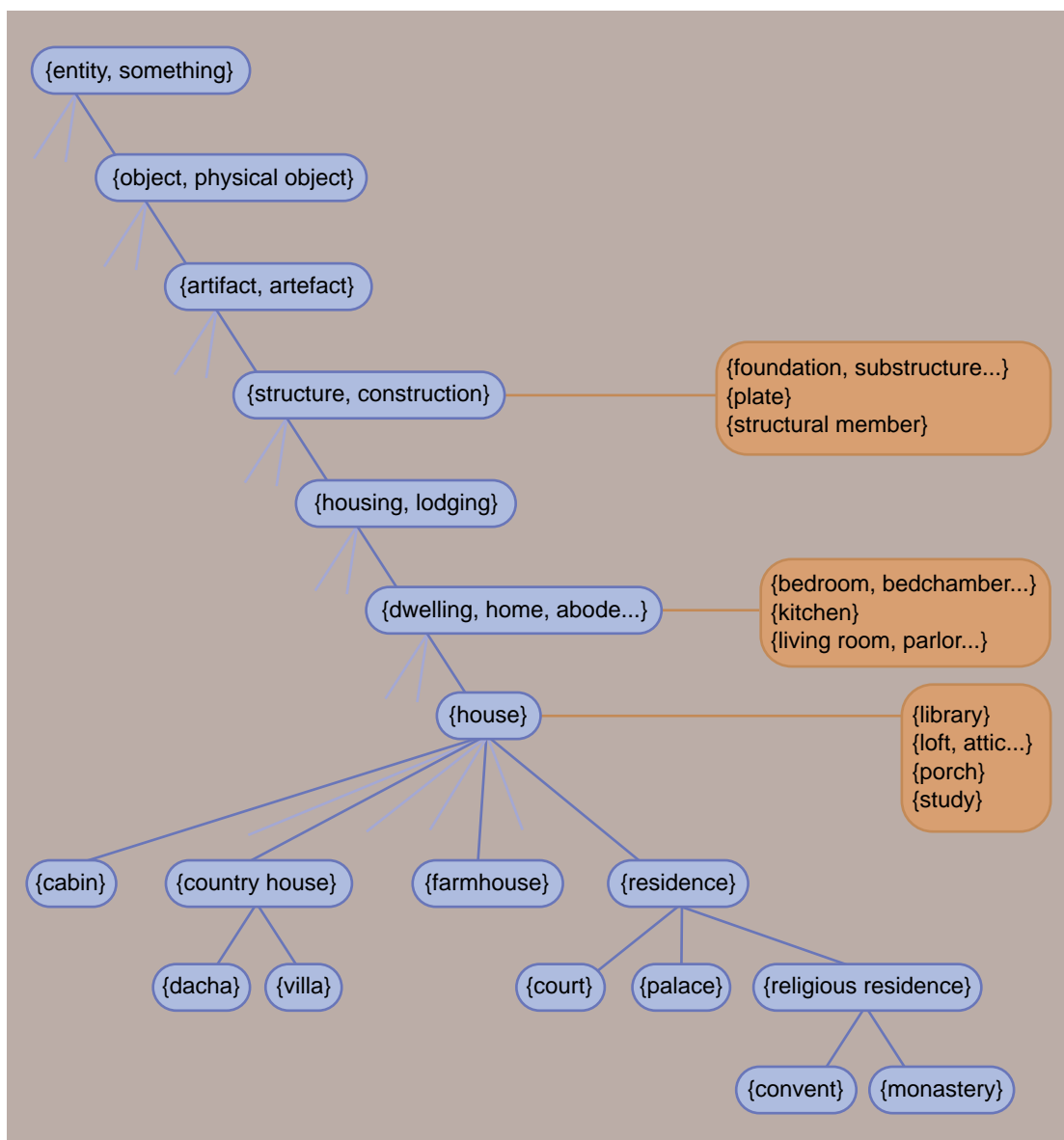

Figure 1. Small portion of the WordNet lexical graph is centered on the synset representing the first sense of the noun *house*. Nodes and edges shown in blue are part of the hypernym/hyponym hierarchy of "is-a" relations; graph elements in tan represent the holonym/meronym hierarchy of "has-a" relations.

ing the component parts of an action. The verb *step* fills this role for *walk*, since walking entails taking steps.

A Theory of Knowledge

The tree structures within the lexical graph make WordNet into a rough taxonomy of the world. The database codifies a theory of knowledge. Concepts accorded a place near the root of a tree are identified as central, basic, primary; those near the leaves are marginal or peripheral. But who determines the branching pattern of the noun and verb trees? Are the fundamental categories of thought inherent in the language, or are they inventions of the lexicographer?

For the noun hierarchy, WordNet posits 11 "unique beginners"—nouns or noun phrases that have no hypernyms. The unique beginners are *entity*, *abstraction*, *psychological feature*, *natural phenomenon*, *activity*, *event*, *group*, *location*, *possession*, *shape* and *state*. Miller does not argue that these particular choices are the only possible ones, but neither are the 11 categories entirely arbitrary or idiosyncratic. They reflect an analysis by Philip N. Johnson-Laird of Princeton of the classes of nouns that can be modified by various adjectives. The categories also satisfy another important criterion: The hierarchy has a place for every English noun.

In practice, no single, immutable hierarchy can possibly capture the structure of the entire lexicon. Conflicts over the classification of words emerge not only at the root but at all levels of the tree. Consider the small subtree of nouns that denote close family relations. The generic term *relative* can serve as the root of the subtree, but what are its immediate hyponyms? In one scheme the *relative* node has two subordinate nodes, *kinswoman* and *kinsman*. At the next level the subordinates of *kinswoman* include *mother*, *sister*, *daughter*; those of *kinsman* include *father*, *brother*, *son*. But the same words could equally well be organized another way, giving *relative* three hyponyms, namely *parent*, *sibling* and *child*; then each of these nodes divides by gender into *mother* and *father*, *sister* and *brother*, *daughter* and *son*. As it happens, WordNet is inconsistent in its treatment of the lexical family tree. *Sister* and *brother* follow the first model: They are listed as hyponyms of *kinswoman* and *kinsman* respectively. But *mother* and *father* are hyponyms of *parent*, and similarly *daughter* and *son* are grouped under *child*.

Of course neither solution is clearly right or wrong. Whether the more natural division is by gender or by generation depends on context, and people have no trouble keeping both hierarchies in mind at the same time. Furthermore, both schemes lose some of their appealing symmetry when they are extended to other relatives, such as *cousin*, which carries no indication of gender in English. (But I can't resist noting that *cousin* derives from the Latin *consobrinus*, which originally referred only to a cousin on the mother's side, and which derives in turn from *soror*, sister.)

These tangled hierarchies are not the only oddities lurking in the lexical trees. One might reasonably suppose that hyponymy, meronymy and other relations between synsets would be transitive. In many cases they are. A mouse is a mammal, a mammal is an animal, and sure enough a mouse is an animal. Often, however, language fails to obey the Aristotelian rules. A house has a door and a door has a knob, but most people would find it odd to say that a house has a knob.

In citing these inconsistencies and peculiarities I don't mean to suggest that the graph-theoretical approach to the lexicon is fundamentally wrong. On the contrary, I would argue that these bugs are features! It seems to me that WordNet is most illuminating just where the construction of the graph runs into difficulties. This is where we may have a chance to learn something about the underlying structure of the language.

Discourse in theoretical linguistics proceeds largely by example and counterexample, by constructing sentences that the prototypical native speaker would or wouldn't find acceptable. Much of this discourse has been carried on with rather small specimens of language—with "toy" grammars and lexicons that generate only a tiny subset of all possible sentences. But building a lexicon of 50 or 100 words and leaving the rest of the language as an exercise for the reader risks missing something important. Indeed, it is characteristic of graphs that some properties do not emerge until the last nodes and edges are added.

Black and White

The treelike organization of both nouns and verbs in WordNet leads naturally to the hypothesis that all words are best catalogued in such structures. The treatment of adjectives therefore comes as a surprise. In WordNet adjectives do not grow on trees; instead they come in matched pairs of opposites—*black* and *white*, *clean* and *dirty*, *fast* and *slow*, *good* and *bad*. And on reflection, the pairing of antonyms does seem like the natural organizing principle for this class of words. In free-association tests, many adjectives strongly evoke their opposites, suggesting that we file them mentally in symmetrical pairs.

There is something else odd about adjectives. Other relations in WordNet are between meanings or concepts, which are conveniently represented by synsets, but the antonymy of adjectives seems to be a relation between specific words. *Hot*, *sultry*, *torrid* and *sweltering* may all belong in the same synset, but the antonym *cold* is strongly associated with only one of these. If you ask people "What's the opposite of *hot*?" you'll get an immediate answer, but "What's the opposite of *torrid*?" is a harder question.

The solution adopted in WordNet is to organize adjectives in clusters around focal pairs of antonyms. Thus *hot* and *cold* stand face-to-face like gang leaders, each surrounded by a throng of allied words, *fiery* and *blistering* on one side,

frigid and *chilly* and *frosty* on the other. If you ask WordNet for the antonym of *torrid*, it responds: "indirect (via *hot* → *cold*)."

This scheme for organizing adjectives was not planned when the WordNet project began in 1985. The need to provide antonym pointers between specific word forms was something discovered while building the graph, and it was not an altogether welcome discovery. Up to then it had been assumed that all edges of the graph would extend between synsets; the database format had to be altered to accommodate the antonym pointers.

WordNet includes some 3,500 clusters of adjectives arranged in antonymous pairs. Most of them fit neatly into the bipolar plan. Indeed, many of the adjectives can be graded, or arranged along a one-dimensional continuum. *Hot* and *cold*, for example, have *tepid* at the neutral point between them, with *warm* and *cool* occupying less extreme positions. A few adjectives, however, refuse to conform. *Angry* is the chief example mentioned by WordNet's authors. *Angry* is the focus of a cluster of related words, which can be graded according to intensity from *annoyed* to *furious*, but English seems to offer no antonym to *angry*. The asymmetry is a reminder that although language is a human invention, it is not an engineered product; it doesn't have to be consistent.

WordNet includes adverbs as well as adjectives but finds little to say about them. They are grouped into synsets, and some of them are linked to adjectives from which they are derived, but there are no taxonomic trees or bipolar pairs. Other parts of speech—the prepositions, conjunctions, pronouns and other "little" words—are omitted altogether.

Six Degrees of Lexicography

Looking on language as a graph invites the kinds of questions that graph theorists ask. Many of these questions have to do with connectivity—with the number of edges linking pairs of nodes.

The ultimate in connectivity is a clique, which is a graph where every node is directly linked to every other node. But a clique is not a plausible architecture for a lexical graph. (For one thing, there are not enough words to uniquely name all the relations between words.)

At the opposite end of the connectivity spectrum is a graph with no edges at all, just isolated nodes. Again the lexical graph can look nothing like this. Suppose a word occupied such a lonely outpost. With no relations to any other words—no synonyms, no hypernyms, no antonyms—what could it possibly mean? What could one say about it—or say with it? Even pairs of nodes linked only to each other are problematic. They would be no more useful than the dictionary—famous in the lore of lexicography although I don't know if it really exists—that defines *furze* as *gorse* and then defines *gorse* as *furze*.

A more realistic question is whether the lexical graph consists of a single connected component.
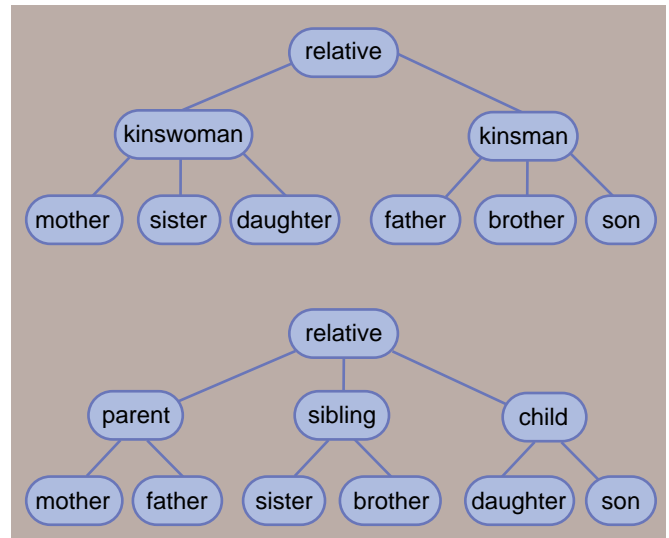

Figure 2. Words for common relations can be organized by gender or by generation; WordNet 1.6 mixes the two architectures.

Can you find at least one continuous path from any given node to any other? For WordNet in its present form the answer is clearly No. With the exception of some adverb-adjective links, there are no edges between different parts of speech, and even within the noun and verb hierarchies the graph breaks into several disconnected pieces. But the mind's lexical graph is surely richer in relations than WordNet and may well be connected. Trying to find a plausible path of relations between two randomly chosen words is easy enough that it doesn't even make a very good parlor game. (And suppose you found a pair of words that stumped everybody, that seemed to have nothing whatever in common. They would then be related by their shared membership in the unusual class of words that are not otherwise related.)

In the end the question of connectivity comes down to what kinds of relations between words qualify as edges of the graph. Within the mental lexicon there are surely many more kinds of links than WordNet admits. Consciously or unconsciously, we form word associations based on common etymology (*river* and *arrive*), based on assonance or rhyme (*slumber* and *encumber*), based on pairing in familiar phrases (*law* and *order*). None of these relations are likely candidates for inclusion in WordNet, but a few other kinds of edges could be important additions to the graph.

Selection rules operating between parts of speech might be the most valuable enhancements. As noted above, adjectives can supply useful clues to the classification of nouns. For example, the adjectives *living* and *dead* can describe only biological organisms (except in figurative uses). In a similar way, many verbs are restricted to certain classes of subjects and objects. You can *count sheep* or *noses*, but you can't *count water*; you can *open* a *door* but not a *contradiction*. Would adding such constraints to the graph be a practical undertaking? That depends on whether or not the con-

straints follow the hyponymy and troponymy hierarchies. If the selection rule for the object of *eat* could be encoded by creating a single link to the node for *food*, which would then subsume all the hyponyms of *food*, the number of edges added to the graph would be manageable. But English is surely not quite that tidy; in the worst case, inserting a separate edge between every verb and each of its potential objects would bring an exponential explosion in the number of edges.

Even though WordNet does not currently include selection-rule information, Philip Resnik of the University of Maryland has used the WordNet database, along with a large corpus of English prose, to derive probabilistic selection rules. For example, he notes that if you choose a sentence at random from the corpus, the subject is more likely to be a hyponym of *person* than a hyponym of *insect*. But if the verb in the sentence happens to be *buzz*, the probability of finding an *insect* in the subject position rises considerably. The magnitude of the change in probability conveys information about the strength of the selection rule.

Another possible elaboration of WordNet would address what Roger Chaffin of the University of Connecticut has called "the tennis problem." Miller writes: "Suppose you wanted to learn the specialized vocabulary of tennis and asked where in WordNet you could find it. The answer would be everywhere and nowhere. Tennis players are in the *noun.person* file, tennis equipment is in *noun.artifact*, the tennis court is in *noun.location*, the various strokes are in *noun.act*, and so on. Other topics have similarly dispersed vocabularies. At least part of the dissatisfaction with a purely hierarchical organization of nouns can be attributed to this neglect of co-occurrence relations."

Browsing WordNet

WordNet is a product of hand-crafting. Essentially all the words were manually entered into lexicography files, with simple textual markers to indicate the various relations between synsets. For example, the character @ denotes a hypernym, ~ is for hyponyms and ! marks antonyms. The entire structure of the lexical graph is implicit in these files, but the relations are not very readily accessible in this form. The files are therefore compiled with a program called the Grinder, which produces a database in which each relation is encoded as a position, or offset, within a file. Long lists of numerical offsets are not very congenial for human readers, but they are easily traversed by a computer program.

WordNet is supplied with a point-and-click browser interface. As with an ordinary on-line dictionary or thesaurus, you can type in a word and get a listing of synsets and an explanatory gloss for all the word's senses, but this overview is only the beginning of what the browser offers. Depending on the part of speech, you can then climb up or down in a lexical tree, search for co-ordinate terms (siblings at the same level of the tree), find antonyms, rank synonyms by their frequency or similarity, list meronyms, and so on.

The browser serves well for casual exploration of WordNet, but more serious work with the lexical graph generally requires writing specialized software to read the database files. Several such projects have been undertaken both within the Cognitive Science Laboratory at Princeton and elsewhere. Philip Resnik's work on selection rules is mentioned above. Another major endeavor is a series of "semantic concordances"—texts linked to the lexicon in such a way that every substantive word is tagged with the appropriate sense from the WordNet database. One of the texts tagged in this way is Stephen Crane's novella *The Red Badge of Courage*.

Still another application, described by Ellen M. Voorhees of the National Institute of Standards and Technology, uses WordNet to improve the accuracy of text retrieval from document databases. By encoding the content of a document in terms of WordNet synsets instead of individual words, a query can specify particular senses of words. If this approach succeeds, it might solve the problem of searching the Web for *coke* and finding soft drinks and illicit drugs when what you're looking for is carbon.

In my own view, the construction of WordNet would be of interest even if the finished product had no applications at all. It is a linguistic counterpart to the human genome program. Just as we all inherit DNA but cannot ordinarily peer into our own genes, we all possess some mental representation of a lexical graph but know little about its overall structure.

Access to WordNet

The WordNet database and browser programs can be downloaded from the Web site http://www.cogsci.princeton.edu:80/~wn/main/. Versions of the browser software are available for the Macintosh, for Microsoft Windows and for various flavors of Unix. The same materials are also offered on a CD-ROM distributed by the MIT Press.

Bibliography

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge, England: Cambridge University Press.

Evens, Martha Walton (ed.). 1988. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge, England: Cambridge University Press.

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

Pustejovsky, James (ed.). 1993. *Semantics and the Lexicon*. Dordrecht: Kluwer Academic Publishers.

Resnik, Philip. 1998. WordNet and class-based probabilities. In Fellbaum 1998, pp. 239–263.

Sparck Jones, Karen. 1986. *Synonymy and Semantic Classification*. Edinburgh: Edinburgh University Press.

Voorhees, Ellen M. 1998. Using WordNet for text retrieval. In Fellbaum 1998, pp. 285–303.

Wilks, Yorick A., Brian M. Slator and Louise M. Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge, Mass.: The MIT Press.