Towards a Representation of Idioms in WordNet

Christiane Fellbaum
Cognitive Science Laboratory, Princeton University
Rider University
Princeton, New Jersey, USA

1 Introduction

WordNet (0; 0) is perhaps the most widely used electronic dictionary of English and serves as the lexicon for a varity of different NLP applications including Information Retrieval (IR), Word Sense Disambiguation (WSD), and Machine Translation (MT). Despite WordNet's large coverage, which comprises some 100,000 concepts lexicalized by approximately 120,000 word forms (strings) and is comparable to that of a collegiate dictionary, it contains relatively little figurative language. WordNet includes a number of multi-word strings, such as phrasal verbs, but many idiomatic verb phrases like *smell a rat*, *know the ropes*, and *eat humble pie*, are missing. Idioms and metaphors abound in everyday language and are found in texts spanning many genres (see, e.g., (0) for a numerical estimate of the frequency of idioms and fixed expression). Clearly, a dictionary that includes extended senses of words and phrases is likely to yield more successful NLP applications. On the one hand, no system wants to retrieve the string *bucket* from the idiom *kick the bucket*. On the other hand, MT and WSD efforts need to distinguish the sense of *ropes* in phrases like *know/learn/teach someone the ropes* from the sense meaning "strong cords"; selecting the latter sense in any of the idiomatic phrases leads to failure. An IR query is likely to be interested *only* in the "strong cord" reading. When this sense is to be retrieved with the aid of a lexicon intended for multiple applications, the figurative sense must be successfully recognized and excluded from a text that may contain instances of the string *ropes* with both meanings.

In this paper, we consider the possibility of extending WordNet to accommodate figurative meanings in the English lexicon. While much has been written on figurative language, there is no agreement on the boundary between literal and non-literal language, see e.g. (0). Criteria that are commonly accepted include semantic non-compositionality and syntactic constraints on internal modification (such as adjective and adverb insertion) and movement transformations. Our purpose here is not to attempt a clear delimitation or definition of non-literal language, but to examine how extended senses of words and phrases from different syntactic and lexical categories-or conforming to none of the standard categories-are compatible with the network structure of a relational lexicon like WordNet and its particular way of representing words and concepts. Our discussion will focus on, but not be limited to, idiomatic verb phrases.

2 A simple classification

An inspection of idiom dictionary sources such as (0) suggests a three-fold distinction among idioms for our purposes.

3 Constructions

First, some idiomatic constructions are simply too complex to be integrated into WordNet and must be excluded at this point. We have in mind constructions of the kind studied by (0) and (0; 0). Examples are *the more the merrier* and *she can't write a letter, let alone a novel*. These structures comprise discontinuous constituents and morpheme chunks that are governed by special syntactic and semantic rules. Thus, *the X-er the Y-er* allows the insertion of a wide variety of adjectives. Fillmore et al. discuss *let alone* and show that its syntactic properties require an amazing amount of description of facts absent from the standard grammar. A full account of these constructions goes far beyond the lexical level, and therefore we need to

exclude them, at least for now, in a database like WordNet that does not include much syntax and whose relational semantics cannot accommodate the kind of semantic facts observed by Fillmore et al. and Jackendoff.

4 Idioms as a kind of polysemy

By contrast, the second kind of idiomatic structure is unproblematic for WordNet. WordNet contains not only simple verbs and nouns but also more complex verb and noun phrases like *show the way* and *academic gown*. Strings like *stepping stone*, *kick the bucket*, *hit the bottle*, and *come out of the closet* therefore correspond to categories already represented in the database, and can be included when they are considered as particular manifestations of polysemy. Polysemy in WordNet is represented by membership of the polysemous string in different synonym sets; synonym sets (synsets) in WordNet represent concepts that are lexicalized by one or more strings (synonyms). In other words, the synsets contain different words forms with the same meaning, and a word form with more than one meaning appears in as many different synsets as it has meanings.

For example, the string *fish* occurs as a verb in two different synsets, and has thus two distinct senses in WordNet. One expresses the concept "catch, or try to catch, seafood;" the other sense is "seek indirectly," as in the phrases *fish for compliments* and *fish for information*. Note that such a representation does not in fact attempt to answer the question as to whether or not the second sense of *fish* is indeed an "extended" one or not, but simply treats them as different meanings of the same word form.

Figurative senses can be seen as homophones rather than polysemes in that there is no discernible relation between the "literal" and the "extended" senses. WordNet does not formally distinguish between polysemy and homophony but treats these two phenomena of multiple meanings alike under the label of polysemy.

In all cases of polysemy, membership in two different synsets entails a different location in the semantic network and relatedness to distinct concepts for each sense. Thus, the first sense of *fish* is a subordinate of *catch* and is further related to more semantically specified senses (troponyms) including *flyfish*, *net fish*, *trawl*, and *shrimp*. The second, arguably extended, sense has as its superordinate concept the synset containing the strings *search* and *look for*. The different locations in the network of the two senses of *fish*, together with the difference in the kinds of noun objects they select are the sort of information exploited in NLP applications, and they will suffice in most cases to distinguish the two senses in such cases where the senses are homophones rather than polysemes.

Some phrases consisting of more than one word can be treated in a similar manner. For example, the idiomatic verb phrases *kick the bucket*, *chew the fat*, and *take a powder* can be considered as single units. Their constituents never occur in an order different from the cited one because these idioms are syntactically completely frozen. They not tolerate the insertion of an adjective or adverb, nor do they undergo passivization, clefting, or any movement transformation that would change the order of the individual strings.[1]

The system therefore needs only to recognize the string that is part of the lexicon. If the strings *kick, bucket, powder, fat*, etc., occur outside of the idiom order, they do not receive the idiomatic interpretation and must be considered as carrying different meanings.

Some compound nouns have extended senses as well, such as *stepping stone*, *straight arrow*, and *square shooter*. We classify these as instances of non-literal language, because the head (the rightmost noun) is not the superordinate concept for the figurative reading: a *stepping stone* is not a kind of *stone*; a *straight arrow* is not a type of *arrow*, and a *square shooter* is not a specific *shooter*. By contrast, nouns like *limestone, gravestone*, and *gemstone*, and *sharpshooter* and *trapshooter* are linked to their superordinates senses, one or more senses of *stone* and *shooter*, respectively; similarly, a *broad arrow* is a subordinate of *arrow*. Many NLP applications using WordNet for determining discourse coherence, finding malpropisms (0), and word sense disambiguation (0); (0) identify related word senses by means of

[1] Only the verb changes in that it shows the usual inflectional endings; this should not pose a major problem for English idioms where the verb is virtually always the first constituent in a Verb Phrase (VP) idiom and can thus be easily recognized.

links such as between super- and subordinates. When searching a text, such systems could easily recognize (and discard as potentially related senses) figurative compounds such as *stepping stone* and *straight shooter* because these are not linked to nouns corresponding to their heads.[2]

Moreover, literal and figurative senses are often in very different WordNet files: an *arrow* (and its hyponyms *broad arrow* and *butt shaft*) are classified as noun.artifacts; while a *straight arrow* is found in the noun.person file.

Frozen VP idioms and metaphoric noun compounds can be integrated into the WordNet database and distinguished from literally referring expressions in many cases. But much of what is commonly considered to be figurative language presents more serious problems for a semantic network like WordNet and applications relying on its particular design. The remainder of this paper will be devoted to a discussion of the third category of idioms, which includes verb phrases like *learn the ropes* and *hide one's light under a bushel*. These cannot automatically be integrated into WordNet, but we offer some proposals for adding them to the lexicon.

5 Some challenging idioms for WordNet

The integration into WordNet of many idioms that do not fall into one of the categories discussed above is problematic for a variety of reasons.

6 Formal problems

First, there are formal problems. Some idiom strings have surface forms that do not conform to any of the syntactic categories included in WordNet. For example, many idioms must occur with a negation: the VP *give a hoot* loses its (figurative) meaning in the absence of negation; the same is true for the VP *hold a candle to*. The negation must therefore be considered part of the idioms. But a verb phrase headed by negation is not a constituent recognized in WordNet.

[2]In this respect, idiomatic compounds resemble exocentric compounds like *low-life* and *scofflaw*, which are not kinds of *lives* or *laws*, either, nor are they found in the vicinity of these concepts in the semantic net.

Consider also the string *eat one's cake and have it, too*: here, two verb phrases are adjoined and are often followed by an adverb. Moreover, the second clause contains a pronoun coreferent with the noun in the first clause. Again, such a string does not fit in with WordNet's entries. Some idioms are entire sentences. *Wild horses could not make me do that* and *the cat's got your tongue* are not compatible with any of WordNet's noun, verb, adjective, or adverb component. WordNet does not contain sentences, and at present we see no way of integrating these into the lexical database. The problem should be addressed in the future, because an NLP system would simply attempt to treat each constituent in these idioms separately, with undesirable consequences.

In some cases, idioms whose syntactic shape does not correspond to any of the categories in WordNet could be accommodated nevertheless when they are synonymous with strings that are represented in an existing synset. For example, the negation-headed phrase *not in a pig's eye* and the clauses *when hell freezes over* and *when the cows come home* are all synonymous with *never*, which is included among WordNet's adverbs. If such strings are completely frozen, as they tend to be, they can be included as synonymous members of existing WordNet synsets and the fact that they do not conform to any of WordNet's syntactic categories can be ignored. Such idioms do not pose problems for automatic processing because they do not admit of any phrase-internal variation or modification.

Another formal (syntactic) problem pertains to the fact that the fixed parts of many VP idioms are not continuous. For example, a number of expressions contain nouns that resemble inalienable possessions, such as body parts, and a possessive adjective that is bound to the subject. Examples are *hold one's light under a bushel*, *blow one's stack*, and *flip one's wig*. In other idioms with a similar structure, the possessive is not bound to the subject but refers to another noun (*got someone's number*). And expressions like *cook one's goose* allow for both bound and unbound genitives.

These idioms cannot be treated as single strings because the genitive slot can be filled by any of the possessive adjectives, or by a noun in the case of the unbound genitive. One so-

lution would be to enter these strings into the lexicon with a placeholder, such as a metacharacter, in place of the genitive. This would make for a somewhat unfelicitous entry. But a rule could be added to a preprocessor for a syntactic tagger that allowed the placeholder be substituted with either a pronoun from a finite list (for the bound cases) or any noun from WordNet (for the unbound cases); the preprocessor would then be able to recognize the idiom as a unit and match the WordNet entry and the actual string. Currently, we do not have a preprocessor that is able to recognize discontinuous constituents, but given the large number of VP idioms and their frequency in the language, the development of such a tool seems desirable.[3]

7 What kinds of concepts are these?

In the previous section, we considered idioms whose syntactic form does not comply with any of the categories N(P), V(P), Adj(P), or Adverbial(P) represented in WordNet or whose syntax poses problems for the creation of a neat dictionary entry. However, such idioms could easily be added to the lexical database when they are synonymous with strings that fit into WordNet's design and organization. But many such syntactically idiosyncratic idiom strings raise a second problem having to do with their conceptual-semantic rather than their syntactic nature. They express concepts that cannot be fitted into WordNet's web structure either as members of existing synsets or as independent concepts, because there are no other lexicalized concepts to which they can be linked via any of the WordNet relations. In fact, if one examines idioms and their glosses in an idiom dictionary, one quickly realizes that almost all idioms express complex concepts that cannot be paraphrased by means of any of the standard lexical or syntactic categories. Consider such examples as *fish or cut bait*, *cook one's/somebody's goose*, and *drown one's sorrows/troubles*. These idioms carry a lot of highly specific semantic information that would probably get lost if they were integrated into WordNet and attached to more general concepts.

[3] A related phenomenon is that of phrasal verbs, many of which allow particle movement. In the cases where the verb head and the particle are not contiguous, they cannot currently be adjoint by the preprocessor and they are therefore not matched to an entry in WordNet.

The problems for WordNet posed by syntactically or semantically idiosyncratic idioms would be reduced if these could be broken up, that is, if the individual content words in the idioms could be treated as referring expressions and be assigned meanings that are similar to concepts already represented in the lexicon. Some traditional dictionaries decompose a number of such idioms and attempt to give an interpretations to their individual parts. This may seem justifiable particularly in cases where the idioms are syntactically variable, indicating that speakers assign meanings to some of their components. For example, the American Heritage dictionary defines one sense of the noun *ice* as "extreme unfriendliness or reserve." This entry seems motivated by the apparent semantic transparency of the noun (in contrast to strings like *bucket* in *kick the bucket*, which seems to have no referent at all, let alone a transparent one). But synsets of the kind *ice, extreme unfriendliness or reserve* seem undesirable for a computationally viable dictionary like WordNet, because *ice* cannot be used freely and compositionally with the proposed meanings. This is evident in sentences like the following:

I felt/resented his unfriendliness/reserve/*ice

His unfriendliness/reserve/*ice melted away

Our laughter broke the *unfriendliness/reserve/ice

A language generation system (or a learner of English) relying on WordNet's lexicon could not be blocked from producing the ungrammatical sentences above, if they are exploiting on the close similarity and usage of the members of the synset. Moreover, automatic attempts at word sense disambiguation that rely on syntactic taggers could probably not identify the correct sense of *ice* in this phrase, because they could not recognize that the noun is a part of an idiom if the dictionary entry contains this noun in isolation, outside of its idiomatic context. Only when one entry for *ice* lists the specific environment (*break* and the definite determiner) can a program recognize the idiom and assign the proper meaning.

Consider a second example. The American Heritage Dictionary contains an sense of *ropes* that is glossed as "specialized procedures or details." This sense of ropes is the one in

the expressions *know/learn/get/teach the ropes*. To assume a compositional reading here seems more justified than in the case of *ice*, because this idiom is more flexible than *break the ice* and can undergo some internal modification as well as passivization (*he never learnt the ropes/he taught Fred the ropes/Fred was taught the ropes*). Moreover, *ropes* co-occurs with more verbs than just one. In fact, the verbs for which it can serve as an argument are compatible with the meaning assigned to *ropes* by the American Heritage Dictionary. A word sense disambiguation system that relied on the semantics of the contexts of the ambiguous word (such as the verbs a noun co-occurs with), would probably choose the correct sense of *rope*, because the contexts of "specialized procedures" or "details" do not seem to overlap with the contexts in which *ropes* is found with the sense of "strong cords."

Yet despite their shared verb contexts, the distribution of *ropes* is far more narrow than that of *specialized procedures or details*. Again, a language generation system or a learner of English might overgenerate and produce incomprehensible sentences like *I forgot the ropes* or *Tell me the ropes*. Therefore, an optimal solution might be to enter the idiom as a string but with a placeholder instead of the verb; a separate rule in the lexicon would list the verbs that are compatible with the idiomatic reading of the string.

The proposed solution for the idioms like *teach/learn/get the ropes* and those that contain a possessive genitive might suggest a huge amount of work. However, a survey of English idioms suggests that most are frozen and could therefore simply be entered as entire strings, without the need for specifying a list of selected verbs.

Another type of VP idiom that does not readily fit into WordNet is that whose meaning can be glossed as *be* or *become Adj*. These idioms have the form of a VP but express states: *hide one's light under a bushel* and *hold one's tongue* mean "be modest" and "be quiet," respectively; *flip one's wig*; *blow one's stack/a fuse*, and *hit the roof/ceiling* all mean "become angry," and *get the axe* means *be fired/dismissed*. Similarly, the phrase *one's heart goes out (to)* can be glossed by means of the verb *feel* and the adjective phrase "sorry or sympathetic (for)." Such idioms pose a problem for integration into WordNet, not because of their form but because of the kinds of concepts they express. In WordNet, verbs (including copular verbs) and adjectives are strictly separated because they express distinct kinds of concepts. This separation is of course desirable and even necessary when one deals with non-idiomatic language, where the meaning of a phrase or sentence is composed of the meanings of its individual parts. Copular or copula-like verbs like *be* and *feel* combine with a large number of adjectives and there is no point in entering specific combinations into a lexicon.[4]

While the separation of verbs and the adjectives they select accounts for the large number of possible combinations allowed in the language, it also means that there exist no concepts like "feel sorry/sympathetic (for)" or "become angry" in WordNet, and idioms like *one's heart goes out (to)* and *hit the roof* are presently excluded from the lexicon. Yet these strings need to be added if the lexicon is to serve NLP applications of real texts, where idiomatic language is pervasise. Expressions of the kind listed above can simply be added as subordinates of *be* without causing a change in the structure of the lexicon. They would stretch the meaning of troponymy, the manner relation that organizes the verb lexicon, in that it is somewhat off to state that "to be angry is to be in some manner." However this seems to be the only way to accommodate such idioms, which express concepts of the kind not found in the literal language.

8 Summary and conclusions

We considered the nature of idiomatic expressions in the light of their potential integration into WordNet. Some idioms pose formal, syntactic problems and express complex concepts that are not expressible by means of the standard lexical and syntactic categories, including those represented in WordNet. Other idioms are formally unremarkable but express concepts that cannot easily be connected to any of the concepts in the semantic network. Perhaps one function of idioms (and one reason for their frequency and their persistence over time) is to

[4] There are some de-adjectival verbs that express specific concepts with meanings "be or become Adjective," such as *pale* or *redden*. Idioms that express the same concepts as such verbs could be added as synonyms, but these cases are very few.

provide for the pre-coded lexicalized expression of complex concepts and ideas that do not exist as units in the language and would have to be composed by speakers. Their frequent occurrence in the language seems to show that many idioms refer to salient concepts and must be considered an important part of the lexicon. We have made some proposals for their integration into WordNet that should benefit in particular the kinds of NLP applications that rely on this lexical resource.

References

Boatner M T, Gates J E, and Makkai A (1975). *A Dictionary of American Idioms.* Woodbury, NY: Barron's Educational Series.

Fellbaum C (1998, Ed.) *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Fillmore C, Kay P, and O'Connor, C (1988). Regularity and idiomaticity in grammatical constructions. *Language* 1988:64, 501-568.

Hirst G and St-Onge D (1998). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In: Fellbaum, C. (Ed.).

Jackendoff R (1995). The Boundaries of the Lexicon. In: *Idioms: Structural and Psychological Perspectives*, M. Everaert, E. J. van den Linden, A. Schenk, and R. Schreuder, Eds. Hillsdale, NJ: Erlbaum. 133-165.

Jackendoff R (1997). Twistin' the night away. *Language* 73, 534-559.

Leacock C and Chodorow M (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In: Fellbaum, C. (Ed.).

Miller G A (1990, Ed.) *WordNet: An On-Line Lexical Database.* International Journal of Lexicography, 3.4.

Moon R (1986). *"Time" and idioms.* In: ZuriLEX '86 Proceedings, EURALEX International Congress, Snell-Hornby M. (Ed.). Francke Verlag. 107-160.

Voorhees E (1998). Using WordNet for Text Retrieval. In: Fellbaum, C. (ed.).