

The Representation of Idioms in WordNet

Anne Osherson
St. Hilda's College
Oxford University

anne.osherson@gmail.com

Christiane Fellbaum
Computer Science Department
Princeton University

fellbaum@princeton.edu

Abstract

WordNet's very extensive coverage does not systematically include idiomatic expression, multi-word units that are semantically compositional to varying degrees. This gap must be filled if WordNet is to be successfully used in applications requiring Word Sense Disambiguation and language pedagogy. Our focus here is on Verb Phrase and sentential idioms. We propose a classification of idioms based on their semantic compositionality and suggest a specific mechanism for integrating them into WordNet.

1 Introduction

We define idioms as multi-word units with specific, conventionalized meanings. A hallmark of idioms is their semantic non-compositionality, i.e., their meanings typically are not the sums of the meanings of their constituents. Idioms fall along a scale of semantic compositionality, ranging from the completely opaque (*buy the farm*, *have a chip on one's shoulder*) to fairly transparent (*point a finger at*, *throw pearls before swine*) expressions. Idioms are frequent in all genres and registers and constitute an integral and large, important part of the lexicon (Jackendoff, 1997). They present a challenge to Natural Language Processing systems, which must recognize the constituents as part of a larger unit and assign the appropriate meaning to the phrase (Fazley et al., 2009; Lin and Sporleder, 2009). A resource like WordNet can help with both tasks. Matching strings in text to WordNet's entries will identify lexically relevant multi-word units, and WordNet's relational structure will provide the level of semantic interpretation that has been shown to be useful in many tasks (see Alonge and Loenneker 2004 and Fellbaum, 1998, 2002 for related work).

2 A Semantic Classification of Idioms

The idioms we have encoded can for the most part be divided between two groups: compositional and non-compositional idioms. In compositional idioms, the meaning of the phrase can be deduced from the sum of its parts, although how literal that meaning is varies greatly (Nunberg et al., 1994; Moon, 1998; Cowie 1998 *inter alia*). This is not the case in non-compositional idioms, and we found that most often, non-compositional idioms take the form of evocative images, which serve to illustrate the idiom's meaning in a more abstract manner. Taken individually, the parts of these idioms may have no connection at all with the idiom's meaning and usage. We examine several classes of idioms for their compositionality and their metaphoric character of their constituents with an eye towards their representation in WordNet.

2.1 Compositional Idioms

Compositional idioms work most often by extending metaphorical significance to real and recognizable events. Their meanings may be culture-specific and thus intuitively accessible to native speakers but not to others.

A subclass of compositional idioms whose idiomatic readings are closely related to their literal ones are those referring to physical behaviors. These can be reactions that accompany an emotional or cognitive response, and the idiomatic reading focuses on this response. Examples are listed in (1).

- (1) *raise eyebrows*
bat an eyelid/eye/eyelash
cut to the quick
hit a nerve
get on your nerves
turn your nose up
take your breath away

bite your tongue/lips
bite off more than you can chew

These idioms are derived from the addition of psychological/cognitive meanings to physical events. The phrase preserves a literal meaning but can also be used in an extended sense. Some of the idioms in this category (2) are less plausible in their literal interpretation than those listed above – these have an added element of exaggeration or figurative language, but because the imagery used is almost exclusively sensory, and therefore implicitly understood, these idioms remain as transparent as those in (1). These expressions below may be even more culture-specific than those in (1) and depend on the speaker and audience being familiar with the customs and ideas behind them.

(2) *make your flesh crawl*
make your blood boil
melt your heart

The first kind of idiom is more likely to undergo variations, due to its transparency and plausibility. As well as the tendency to substitute components of these phrases for similar ones (hence the ability to say “bat an eyelid” as well as “bat an eyelash”) there also seems to be a greater freedom in using the phrases themselves – contextual words can be easily added into the phrase, and the original components are subject to morphological changes (Ernst, 1981; Langlotz, 2006; Fellbaum, 2006, 2007). Indeed, internet searches reveal usages such as

(3) Schiavo case hit a political nerve
http://findarticles.com/p/articles/mi_qn4188/is_20050410/ain13598785/

(4) without the merest bat of a judicial eyelash...
<http://visibleprocrastinations.wordpress.com/2006/12/>

(5) the fact that no original Shakespeare manuscripts survive raise a few scholarly eyebrows.
<http://74.125.93.132/search?q=cache:1BJuwTCAbJUJ:www.ricksteves.com/plan/destinations/britain/strtrfd.htm+raise+scholarly+eyebrow&cd=10&hl=en&ct=clnk&gl=us&client=firefox-a>

But variation is also found for idioms of the second kind:

(6) John Saul can make readers' skin crawl
<http://www.highbeam.com/doc/1P2-4067887.html>

(7) Hundreds point a collective finger at Crew and Board of Ed.
<http://connection.ebscohost.com/content/article/1032458628.html?jsessionid=BAA395F60ABB08BD1078EDE8EA451610.ehctc1>

Since they are based on cultural constructs, which change, evolve, and are forgotten with time (unlike human physical reactions listed in (1) and (2), which we can assume remain fairly constant) these idioms are much more susceptible to changes in use and meaning. Certain idioms have even become completely opaque to native speakers of the language, who learn and use it as a set phrase with a set meaning, but are now completely unaware of its original meaning or etymology.

A related subclass of “cultural-construct” idioms in current usage are still arguably transparent:

(8) *take your hat off*
blow the whistle
blow s.b./s.th. out of the water
point the finger at
pull the trigger

Even if the speaker has never actually taken his hat off in respect, he may recognize the gesture and the meaning behind it. Most people have probably never fired a gut, but would understand the idiomatic meaning of *pull the trigger* (but see Keysar and Bly, 1995). *Blow the competition out of the water* is still widely used, but the reference to battleship warfare is quickly fading from our cultural memory, as is the metallurgy origin of *strike while the iron is hot* (Dobrovolskij and Piirainen, 2005). The meaning here resides in the entire idiom rather than in that of the constituents. Consequently, the entire idiom must be regarded as a lexical item.

The idioms listed in in (9) encode cultural references that are probably no longer current to most modern speakers:

(9) *cry wolf* (from a morality story)
rest on your laurels (Ancient Roman laurel crowns)
wash your hands of... (Biblical: Pontius Pilate)
look a gifthorse in the mouth (few people today acquire horses)

These idioms are likely learned and stored as set phrases and probably exhibit less variation since speakers do not assign meaning to their components. We found no attested examples like *kick the pail*, *check a gifthorse's teeth*, *cry coyote*. Nevertheless, some constituents are referential, e.g., *gifthorse*, and could therefore merit a lexical entry.

2.2 Non- Compositional Idioms

The phrases in (10) use imagery to illustrate the idioms' meanings, which do not arise from the constituents' meanings. Rather, the words paint a picture, which, in most cases, conjures up a concrete situation.

- (10) *spill the beans*
walk on eggshells
beat/flog a dead horse
add fuel to the fire
bury your head in the sand
see the light
a watched pot never boils

In some of the idioms, the constituents may have metaphoric character.

2.3 Idioms with mixed properties

The classification suggested above is only approximate; a number of idioms, including those in (11), cannot be straightforwardly fit into any one category.

- (11) *Rise from the ashes*
Throw s.b. to the wolves
See red
Tear your hair out

For example, the reference to a mythical phoenix may be lost to speakers, but *rise from the ashes* nevertheless evokes the intended meaning. Similarly, *tear your hair out*, which refers an ancient mourning ritual, is no longer culturally relevant; yet the original meaning remains perhaps on the strength of the image. *Seeing red* does not literally happen when the idiom is invoked, but the meaning relies on the understood significance of the color red, perhaps specific to Western culture.

The change from transparency to opacity and the loss of meaning of the idioms' components to contemporary speakers is of course consonant

with the overall tendency of the lexicon (and language) to shift and evolve. Idioms are learned, re-hashed, re-interpreted and used in different contexts, resulting in gradual changes in use and meaning, as with many other aspects of language. Idioms that can be classified in more than one way may simply be in a transitional stage.

3 Representing Idioms in WordNet

Currently, only few idioms are included in WordNet. For example, *kick the bucket* and *buy the farm* are members of the synset that also includes *die* (WordNet synsets do not respect connotational or stylistic differences but are purely based on denotational equivalence). Treating idioms as “long words” in this manner is convenient in the case where the idiom is not composed of constituents that have a meaning, i.e., metaphors. But in many cases, the components of idioms can be said to be lexical items (form-meaning pairs) in themselves. For example, in *spill the beans*, the verb arguably carries the meaning “reveal” and *beans* refers to “secret or confidential information.” Speakers assign such meanings to the idiom components, as can be seen by the fact that they modify them or substitute semantically similar items (see papers in Fellbaum 2007). We need to reflect, first, the metaphorical status of such words and, second, the fact that their use is limited to the particular context of an idiom.

An on-line idiom dictionary (<http://www.usingenglish.com/reference/idioms>) was manually examined and over 200 idioms were paraphrased and analyzed for their compositionality by a native English speaker. For each idiom, we identified those components that have a literal meaning. We then manually determined with WordNet synset expressed that meaning and recorded its sense key. The following are representative examples.

Old flames die hard

It's very difficult to forget old things, especially the first love.
flames – **feeling%1:03:00::** the experiencing of affective and emotional states
die - **fade%2:30:00::** disappear gradually

Rain on your parade

If someone rains on your parade, they ruin your pleasure or your plans.

rain on - **spoil%2:41:02::** make a mess of, destroy or ruin

parade – **plan%1:09:00::** a series of steps to be carried out or goals to be accomplished

fun%1:04:00:: activities that are enjoyable or amusing

See the light

When someone sees the light, they realise the truth

see - **realise%2:31:00::** perceive (an idea or situation) mentally

light - **truth%1:26:00::** conformity to reality or actuality

In many cases, some constituents have a literal meaning, as is the case with *keep* in the following example:

Keep at bay

If you keep someone or something at bay, you maintain a safe distance from them.

keep - **keep%2:42:07::** continue a certain state, condition, or activity

bay – **distance%1:07:00::** the property created by the space between two objects or points

distance%1:12:00:: indifference by personal withdrawal

Given these data, the idioms can be straightforwardly integrated into the WordNet database.

3.1 Idiom-specific Relations in WordNet

Wordnet is currently being converted from its text-based format to a relations (SQL) database. One distinct advantage of the new format over the old one will be that it allows the addition of a large number of relations among synsets and synset members. The integration of idioms into WordNet is therefore unproblematic.

The first step is to create a new synset containing the idiom component. In most cases, the synset will have only a single member, though in some cases, different word form, restricted to different idioms, may express the same concept. For example the members of the synset{cat, beans} both refer to a secret in the context of the idioms *let the cat out of the bag* and *spill the beans*. Next, a link will be established between these synsets and the synsets containing the appropriate word form(s) with literal readings; in this

case {secret}, which were manually identified as described above. Finally, the synsets (or synset members if the synset contains more than one) containing the idiom constituents must be inter-linked by another kind of pointer, which assures that only those idiom components that co-occur in the context of a specific idiom are related. For example, {beans} and {spill} are linked, as are {cat} and {bag}, but {beans} and {bags} are not, as they do not share the same distribution.

3.2 A Remaining Problem: Prepositions

The proposed approach to integrating idioms into WordNet will help the automatic recognition of idiomatic strings in a text by identifying the meanings of idiom components. However, we are not able at present to handle prepositions that are specific to an idiom. WordNet does not include prepositions among the major parts of speech that are encoded. Part of the reasoning was that prepositions have a status somewhere between content and function words.

Clearly, the choice of a particular preposition in an idiom is not entirely arbitrary. In some phrases, like *sweep something under the rug/carpet*, which conveys the sense of hiding in a way that *sweeping something on the rug/carpet* could not (and carpet/rugs are a kind of *covering*), the choice of preposition is crucial. In other cases, however, the choice of preposition can be much looser without changing much of the meaning of the phrase - it could *rain on your parade* or *over your parade*, and you could *look on the bright side* or *at the bright side*, and even *look for the bright side*, all of which convey the same meaning. These idiosyncracies make it difficult to treat all prepositions in the same way - they don't all need to be clearly defined, but neither can they simply be left out.

The absence of prepositions in WordNet does not allow us at present to represent their semantic contribution to the idioms. In cases like *take a leaf out of s.o.'s book* (copy s.th. from s.b. to one's advantage), reanalysis was the solution. We analyze *take_out* as a unit (a phrasal verb) and mapped it to an existent sense in WordNet.

4 Conclusions

There are several advantages to representing idioms in this way. Besides closing a serious gap in WordNet's coverage, we distinguish opaque and transparent idioms on the one hand, and partially from fully transparent ones on the other

hand. In addition, the proposed representation takes into account the fact that idioms are not fixed strings or “long words.” Rather, they show a remarkable range of syntactic and lexical modification, including lexical substitution (Langlotz, 2006; Fellbaum 2006, 2007)). This flexibility stumps applications that rely on lexical databases treating idioms as long strings, making it impossible to recognize syntactically or lexically “deviations” from the standard citation form that sets an often arbitrary norm (Sag et al., 2001). By contrast, our proposal potentially enables automatic systems to recognize and interpret idioms in all their varieties.

Statistical approaches that measure the collocational properties of idiom components can perform well in recognizing idiomatic phrases (Church and Hanks, 1990). For example, they recognize that *cat* and *bag* frequently cooccur, and they can do so independently of syntactic variation or modification. However, they are unlikely to recognize idioms where a lexeme has been substituted, but see Herold (2007) for a solution.

Acknowledgments

This work has been supported by grant CI-ADDO-EN 0855157 from the National Science Foundation under the American Recovery and Reinvestment Act. We thank Jordan Boyd-Graber and Bettina Burgett for helpful discussions.

References

Antonietta Alonge and Birte Lönneker. 2004. The Heart of the Problem: How Shall we Represent Metaphors in Wordnets? Proceedings of the Second International WordNet Conference. Brno, p. 10.

Kenneth W. Church, K. and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16:22–29.

Alan. P. Cowie. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press

Dmitrij Dobrovolskij and Elizabeth Piirainen. 2005. *Figurative Language: crosscultural and crosslinguistic perspectives*. Amsterdam: John Benjamins.

Thomas Ernst. 1981. Grist for the linguistic mill: Idioms and "extra" adjectives'. *Journal of Linguistic Research*, 1:51–68.

Afsaneh Fazly, Paul Cook and Suzanne Stevenson. 2009. Unsupervised type and token identification

of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Christiane Fellbaum. 1998. Towards a representation of idioms in WordNet. In S. Harabagiu (ed), *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems*. Montreal: COLING/ACL 1998, pp. 52–57.

Christiane Fellbaum. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In S. Busemann (ed), *Proceedings of KONVENS 2002*. Saarbrücken, Germany: DFKI, pp. 7–11.

Christiane Fellbaum. 2006 (ed.) Corpus-based Studies of German Idioms and Light Verbs. *International Journal of Lexicography* 19.4.

Christiane Fellbaum. 2007 (ed.) *Idioms and Collocations*. Birmingham, UK: Continuum Press.

Axel Herold. 2007. *Corpus Queries*. In: Fellbaum (2007, ed.) 54–63.

Ray Jackendoff. 1997. Twistin' the Night away'. *Language*, 73:534–559.

Boaz Keysar and Bridget Bly. 1995. Intuitions of the transparency of idioms. Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34, 89–109.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Andreas Langlotz. 2006. *Idiomatic Creativity*. Amsterdam: John Benjamins.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford Studies in Lexicography and Lexicology. Oxford: Clarendon Press.

Geoffrey Nunberg, Ivan Sag and Thomas Wasow. 1994. Idioms. *Language* 70, 491–538.

Ivan Sag, Timothy Baldwin, Francis Bond, Anne Copestake and Dan Flickinger. 2001. Multiword Expressions: A pain in the neck for NLP. *LinGO Working Papers*, No. 2001-03.

Linlin Li and Caroline Sporleder. 2009. Classifier Combination for Contextual Idiom Detection Without Labelled Data. *EMNLP*.