VP Idioms in the Lexicon: Topics for Research using a Very Large Corpus

Christiane Fellbaum
Cognitive Science Laboratory, Princeton University
Princeton, New Jersey, USA
and
Berlin-Brandenburgische Akademie der Wissenschaften

Abstract

This paper discusses the status and representation of a particular type of VP idiom in the lexicon and in lexical databases.

1 Introduction

This short paper presents some preliminary considerations for a research program that is built around a very large corpus of German (Digitales Woerterbuch der Deutschen Sprache, DWDS) compiled and housed at the Berlin-Brandenburgische Akademie der Wissenschaften(Cavar et al., 2000). "Woerterbuch", i.e., "dictionary," is in fact a misnomer, as the DWDS is at present a corpus and not a dictionary; at least not yet. But we intend to formulate and structure the results of our corpus investigations as lexicographic information that, together with the appropriate corpus data, will constitute a valuable source for the linguistics and NLP communities.

Our work focuses, in part, on Verb Phrase idioms like *let the cat out of the bag* and *have a chip on one's shoulder.* Linguists and psycholinguists have written much about idioms, but a large-scale, comprehensive, and diachronic study has not yet been undertaken. The size (about 500 million words) and coverage (spanning the entire 20th century) of the well-balanced DWDS corpus will permit a comprehensive study of this part of the lexicon. But beyond the the semantic, syntactic, and distributive properties of idioms, we must consider their place and role in the lexicon. Structurally, the VP idioms we focus on are verbs with their syntactically appropriate complements. But unlike phrases that are composed by speakers according to the rules of syntax and semantics, idioms tend violate these rules. Characteristically, they are syntactically fixed (more or less) and semantically noncompositional. Like simple words, they express a single concept. Their exceptional characteristics require that speakers learn and store them as lexical items. It turns out that VP idioms pose a particular challenge for builders of lexical resources who strive to capture and represent linguistically relevant information.

2 Words and Concepts in the Lexicon

Lexicons give information only about those concepts in the world that are paired with a word, i.e. lexicalized. Traditional paper dictionaries, bound to a phonology-based organization, do not reveal the structure of the lexicon or patterns of systematic lexicalization. Electronic lexical resources, which are subject to fewer design constraints, can show up regularities in the way concepts and words are mapped; they can also reveal interesting properties of the lexicon such as the parallelism in the syntactic and semantic behavior of verbs.

WordNet is an example of an electronic dictionary, designed like a semantic network. It organizes words into the kinds of hierarchical structures familiar from ontologies, and thereby reveals the extent to which concepts and words are mapped systematically, as well as the location of structural lexical gaps. Moreover, WordNet's design has been shown to be useful for a number of NLP applications.

We do not wish to claim that WordNet is the optimal way to represent the lexicon of a language or that it realistically and accurately reflects the way speakers store and access word meanings. But it is fair to say that the fact that the bulk of the English lexicon could be successfully represented by means of WordNet's paradigmatic relations makes it a reasonable and valid model.

WordNet's design and coverage permits one to explore the structure of the lexicon. Looking at WordNet's verb component, one can ask whether the patterns resulting from the purely semantic organization reflect the kinds of patterns found in the syntactic-semantic classification of (Levin, 1993). Our focus here will be on a particular kind of idiomatic verb phrases, and the question how they fit into the verb lexicon as projected in a semantic network.

2.1 Verbs in WordNet

Verbs, like nouns and adjectives in WordNet, are organized into unordered sets of cognitive synonyms, or synsets. Synsets are interconnected by various semantic relations, primarily the "manner" relation, which links synset pairs whose members fit the formula "to X is to Y in some manner." A given verb is related to other verbs that are very similar in meaning but are either more or less semantically elaborated along a given dimension (super- and subordinates, respectively). For example, *move* is a very general concept, and such subordinates as *walk*, *swim*, and *fly* add specific information pertaining to the manner of motion. These in turn are related to verbs denoting even more specific manners, such as *amble*, *backstroke*, and *glide*, respectively. WordNet shows that almost all English verbs can be arranged into tree structures, or hierarchies, based on increasing semantic complexity (Fellbaum, 1998a).

3 Idioms

Some English VP idioms, like *buy the farm* and *have a bun in the oven* serve a euphemistic function and can be simply treated as synonyms of their literal counterparts, *die* and *be pregnant*, respectively. However, a number of idioms are not easily integrated into WordNet's manner hierarchies. They seem to defy the conventions and patterns of lexicalization shared by the bulk of the verb lexicon(Fellbaum, 1998b). Among these are idioms are *walk the plank*, *turn the other cheek*, and *go begging*.[1] These phrases have neither (near-) synonyms nor do they have super- or subordinate concepts.

The shared characteristic of each of these idioms is that they are semantically more complex than most literally referring verbs, which follow the kinds of lexicalization patterns clearly revealed by a semantic network. *Walk the plank* encodes several sub-events (walk down a ship's plank, jump off, and drown). *Turn the other cheek*, glossed in standard dictionaries as "refrain from retaliating after having suffered an apparent injustice" makes implicit reference to a prior event. *Go begging* ("not be wanted or needed") contains an implicit negation.[2]

(Fellbaum, 2002) argues that it is characteristic of many, perhaps most, idioms to express concepts that are semantically rich and highly complex. Crucially, the semantic complexity of idioms does not follow from their syntactic complexity: intransitive verbs like *die* and *run away* have idiomatic synonymous phrases with a transitive verb: *kick the bucket* and *take a powder*, respectively. Conversely, idioms like *shoot the breeze* and *chew the fat*, which contain a verb and an object, are glossed in dictionaries simply as intransitive "chat." The constitutents of idioms commonly do not refer; the meaning of the constituents gets lost as the idiom gradually enters the language as a lexical unit.

(Fellbaum, 2002) examines and classifies different kinds of idioms. The typology includes idioms that incorporate an aspectual component, such as *call it a day/quits* ("stop working or doing what one is doing"); *get on the stick* ("start being active"). A number of VP idioms include a secondary predicate: *set tongues wagging*, *go out on a limb*, *keep one's lips sealed/zipped*. Some idioms link two VP with a Boolean operator, such as *fish or cut bait* and *have one's cake and eat it*.

VP idioms appear to encode specific common events and states that speakers can refer to without the "work" of composing the message on-line. These ready-formed messages are more complex than those of the lexical items that constitute the bulk of the verb lexicon. As a result, they cannot be cast into a straightforward semantic relation with other concepts expressed by verbs.

[1] The idioms I am focusing on here consist of a verb and at least one NP and/or a PP selected for by the verb.

[2] The glosses cited in this paper are taken from the *American Heritage Dictionary* and (Boatner et al., 1975).

4 Negation

Negation is usually expressed by means of a free or bound morpheme such as *not, no, never, un-, non-*. The lexicon itself contains few underived items whose meaning include a negative. But in building the verb component of WordNet, a handful of such verbs stood out because they did not fit into the hierarchical design. They include *avoid, refrain, prevent, lack, fail, ignore, neglect, miss, refuse.* These verbs, in contrast to most others, refer to non-events or non-states. They do not denote specific (non-) actions or events, but function a bit like auxiliary verbs or negative operators and must co-occur with a main verb: *fail/neglect/refuse to V*; *prevent X from V-ing*, *avoid V-ing*, etc.

Most of these verbs are polysemous. When they select for an NP object, their meaning derives from the semantics of the noun, specifically what might be considered its telic role. *Miss the bus/train/plane* means "miss or fail to ride the bus/train/plane"; *miss a class* means "miss or fail to attend a class." *Refuse a present* means "refuse to accept a present"; *prevent*, when selecting for event nouns, means "prevent the happening or occurrence" of the event.

Unlike these verbs, the great majority of verbs serve to refer to actual events and states. In a semantic network like WordNet, which is build around the semantics of ordinary verbs, verbs like *miss* and *fail* do not fit, as they do not express semantic elaborations of other concepts. However, some of these verbs can be represented in terms of semantic opposition relation.

4.1 Idioms with Overt Negation

Browsing through any idiom dictionary turns up a surprising number of VP idioms that contain a negation of some kind:

(1) not have the foggiest/faintest idea (know nothing)

give one the time of day (totally ignore somebody)

not give a damn/hang/hoot/shit (not care at all)

not hold a candle to (not compare favorably to)

(2) never darken one's door (never show up again)

never say die (not be discouraged)

(3) cut no ice with (not impress)

hold no water (not stand up to critical examination)

be neither fish nor fowl (lack specific characteristics)

Others admit both a negative and a Negative Polarity Item (NPI):

(4) not/hardly bat an eyelash (not/hardly show surprise/fear)

not/hardly believe one's eyes (not trust one's eyesight)

Some lexemes seem to exist only as NPI in idioms, and the phrase cannot be interpreted literally in the absence of the negative:

(5) You don't know beans/boobkahs/diddly-squat.

*You really know beans/boobkahs/diddly-squat.

*I found out beans/boobkahs/diddly-squat.

All these idioms lose their figurative meaning or become entirely meaningless without the negation. Their paraphrases show that the negation is part not only of the surface encoding but of the meaning.

4.2 Covert Negation

A related type of idiom lacks any form of negative morpheme in its surface form but contains a negation as part of its meaning, as revealed in the paraphrase:

(6) go begging (not be wanted or needed)

turn a blind eye to (refuse to see or recognize)

fall on deaf ears (go unheeded; be ignored)

fall through the cracks (pass unnoticed, neglected, or unchecked

turn a deaf ear (refuse to hear or listen to)

close the door on (prevent any further action or talk about a subject)

turn one's back on (deny, reject)

beg the question (avoid and not answer a question or problem)

Idioms with such "covert negation" are quite numerous in English. Like those with an overt negative morpheme, they pose a problem for integration into a semantic network, as they break the regular patterns of lexicalization via increasing elaboration.

Because the language has only a few lexicalized verbs that include a negative meaning component, and because these verbs resemble auxiliaries with their low information content, idioms may fill a type of lexical gap, providing preformed messages with negations. The events or states they negate are overwhelmingly cognitive or emotional.

5 States

A strikingly large number of English VP idioms, including the following, express states:

(7) foam at the mouth (be very angry)

hang loose (be calm)

have a chip on one's shoulder (be hostile or combative)

whistle in the dark (be brave)

hang one's head (be ashamed)

push up daisies (be dead)

kick around (be alive)

hold water (be true or irrefutable)

hang by a thread/hair (be in doubt or threatened)

fill/fit the bill (be appropriate)

fly in the face/teeth of (violate, go against)

Properties of entities are regularly referred to by combining a copula like *be* or *become* with an AP, NP, or PP, as in *the child was already big* and *he became old.* English does not provide for the regular expression of such states by means of a single verb, though the event of entering states (or the causation of entering states) is lexicalized regularly by verbs of change like *grow* and *age.* Idioms may fill gaps here, too, by providing lexical items that express specific states and thus obviating the need for the composition of the string.

We have not sampled a sufficiently large part of the lexicon to be able to say whether such idioms tend to predominantly lexicalize states from particular semantic fields like emotions and cognition, although these seem to be frequent.

In the course of our work with WordNet, our attention was drawn to these idioms because they, like those with a negative meaning component, do not fit into the English lexicon, which as a rule does not lexicalize states. These idioms pose a challenge to systematic lexicographic description and the design of computationally viable lexicons, which we hope to meet.

6 Future Work

Searching the DWDS, we plan to address several questions. First, what part of the language is constituted by idioms? Second, are German VP idioms semantically similar to English idioms? In particular, are there as many idioms with a negative meaning component that is expressed either overtly or covertly, and idioms denoting states? What is the nature of these "preformed messages", i.e., why does a phrase become an idiom? Do idioms express the same kinds of concepts encoded crosslinguistically? Finally, given that both idiomatic and non-idiomatic VPs that denote states or contain a negative meaning component do not fit easily into the kind of network or lexical ontology that is exploited in NLP research, how could they be better represented in a way that makes them useful to the computational community?

These questions have, to my knowledge, never been explored with the help of a large corpus that would permit a realistic sampling of idiomatic expressions in natural contexts. We hope to be able to provide some answers by exploiting the DWDS.

7 Acknowledgements

This work was supported by a Wolfgang-Paul-Preis from the Alexander von Humboldt Foundation and grant number IIS-ITR 0112429 from the National Science Foundation.

References

Maxine Tull Boatner, John Edward Gates and Adam Makkai. A dictionary of American id-

ioms. Barron's Educational Series, Woodbury, NY, 1975.

Damir Cavar, Alexander Geyken and Gerald Neumann. Digital Dictionary of the 20th Century German Language. Proceedings of the Language Technologies Conference IS-2000 Ljubljana, Slovenia.

Christiane Fellbaum. *WordNet: An electronic lexical database.* MIT Press, Cambridge, MA, 1998

Christiane Fellbaum. Towards a Representation of Idioms in WordNet. Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems Montreal, CA: COLING/ACL 1998.

Christiane Fellbaum. VP Idioms in a Lexical Ontology. Abstract, Workshop "Ontological knowledge and linguistic coding" DGfS Meeting 2003, Munich, Germany.

Beth Levin. *English Verb Classes and Alternations.* Chicago University Press, Chicago, IL, 1993.