

Head-Driven Lexical Representation of Idioms in HPSG

Gregor Erbach*
Universität des Saarlandes
FR 8.7 Allgemeine Linguistik
Computerlinguistik
W-6600 Saarbrücken, FRG
erbach@coli.uni-sb.de
☎ +49 (681) 302-4117
Fax +49 (681) 302-4351

Abstract

This paper is concerned with the lexical representation of idioms. We distinguish idioms from other kinds of multiple-word expressions like collocations, support-verb constructions, and lexicalized metaphors. A lexical representation for idioms in Head-driven Phrase Structure Grammar (HPSG) is proposed, which can account for the syntactic variability found in idiomatic constructions, and for the non-compositional semantics of idioms. Since all idioms, which are not completely fixed, consist of a lexical head and "frozen" complements, the information about an idiomatic expression can be encoded in the subcategorization list of the idiom's lexical head. Since idioms involve selection for particular lexemes, a feature is added to HPSG signs to encode this lexemic information. It will be argued that the Locality Principle of HPSG is too strong because it prohibits the representation of idioms proposed in this paper, and that the Semantics Principle must be modified. We explain the fact that some idioms do not passivize by the fact that no thematic role is assigned to accusative objects like *the bucket* in *kick the bucket*.

Introduction

The unit of lexical description within theoretical linguistics has traditionally been the word (or the morpheme). According to this view, the lexicon is a set of pairs $\langle \text{Word}, \text{Description} \rangle$, where the description may be a syntactic category (as in phrase structure grammar), an elementary tree (as in Tree Adjoining Grammar) or a feature structure (as in the mainstream unification formalisms)¹.

Syntactic theory is concerned with discovering the rules and principles that combine words into phrases and phrases into sentences, and semantic theory seeks to explain how the meaning of a sentence depends on the words from which it is made up, and its syntactic structure.

* This research was supported through IBM Germany's LILOG project. Thanks to colleagues in Saarbrücken and Stuttgart, and especially Erik-Jan van der Linden, for their comments on the ideas presented in this paper.

¹These remarks should not suggest that the lexicon is just an unstructured set. See (Pollard and Sag 1987) for a discussion of structuring the lexicon by a hierarchy of lexical types and relating classes of lexical entries by means of lexical redundancy rules.

In lexicography², on the other hand, and in a linguistic research tradition called phraseology, it has been realized that single words are not necessarily the appropriate units for lexical description. The lexicon must also contain entries which consist of several words, referred to in the literature as idioms, collocations (Ettinger 1977), compound words (Gross 1986), fixed syntagms (Rothkegel 1973), phraseologisms (Korhonen 1987), lexical solidarities (Coseriu 1967), phraseolexemes (Wotjak 1985) and polylexical expressions (Rothkegel 1989).

For the purposes of this paper, we distinguish idioms from other multiple-word expressions, like collocations, support-verb constructions, and conventionalized metaphorical expressions. All elements of these classes share the property that they involve more than one word. They differ in the semantic combination function that determines the meaning of the expression. However, since idioms develop historically from conventional metaphors and collocations, no well-defined boundaries between the four classes exist.

Collocations are multiple-word expressions where the usual compositional semantic functions construct the meaning of the expression from the meanings of its constituent words. Collocations are characterized by the fact that the words involved in them have a tendency to occur in a particular combination, e. g. *dogs bark*, *cats meow*, *birds sing* or *crude oil* are typical collocations.

Support-verb constructions involve a predicative noun, which contributes the semantic relation of the expression, and a semantically deprived support verb (e.g., take, give ...) which contributes the aspect and "Aktionsart" (e. g., inchoative, terminative ...). Examples are *take a walk*, *give a talk*, *make a contribution*.

Lexicalized metaphorical expressions (or analyzable idioms) like *pull strings*, *attack an argument*, *keep tabs on* can be interpreted by making use of metaphorical language conventions (Martin 1991), which transfer the meanings of the constituents to some target domain. Lexicalized metaphors differ from idioms proper in that the meaning of the expression can be modified by modifying its constituent parts (e.g., *pull certain strings*, *keep close tabs on*), and that the metaphorical convention used for the interpretation of the expression can be used in subsequent discourse. Rather than treating lexicalized metaphors like idioms, we propose to list them in the lexicon, with a pointer to the metaphorical language convention needed for their interpretation. The relation between idioms and lexicalized metaphors is discussed by van der Linden (1991).

Idioms (or unanalyzable idioms) are multiple-word expressions the meaning of which is a property of the whole expression, and not the combination of the meanings of its constituent words.

This definition is based on a definition given by Erik-Jan van der Linden (1989, page 134), but we define idioms as "multiple-word expressions" rather than "multi-lexemic expressions" , because we consider both single words and idioms as lexemes.

While idioms are diachronically often derived from conventionalized metaphors, a metaphorical interpretation of idioms is no longer possible, and their meaning must be listed in the lexicon, and cannot be made accessible by a pointer to a metaphorical language convention. This paper is focussed on unanalyzable idioms that fall under the above definition.

Some idioms are completely fixed and can be entered into the lexicon as multiple-word entries. They do not allow for syntactic variation or internal modification, as seen in the following examples. The # sign means that the string does not have an idiomatic reading

²Examples for the English language are the *Dictionary of English Colloquial Idioms* (Wood 1979), the *Oxford Dictionary of Current Idiomatic English* (Cowie, Mackin, McCraig 1975) and *A Dictionary of American Idioms* (Boatner and Gates 1975).

- (1) a spick and span / #spick and very span
- b by and large / #by and much larger
- c at any rate / #at any special rate
- d real estate / #real urban estate / urban real estate

The majority of idioms, however, allow for considerably more flexibility, for example the idiom *take into account*. In a normal declarative sentence, there is a direct object between *take* and *into account*. The idiom consists of the verbal head *take* and the "frozen" prepositional object *into account*.

- (2) a John took the proposal into account.
- b John took the proposal that the president's advisor made yesterday into account.
- c John took the proposal that Bill claims the president's advisor made yesterday into account.

While most idioms involve "frozen" arguments of the verb, some idioms involve "frozen" adjuncts, e.g., *take the bull by the horns*.

Most idioms headed by a verb can undergo passivization, but there are some cases, where the idiomatic meaning is lost in the passive (e. g. *kick the bucket*).

- (3) a The proposal was taken into account. (idiomatic)
- b The FBI kept tabs on Jane Fonda. (idiomatic)
- c Tabs were kept on Jane Fonda. (idiomatic)
- d John kicked the bucket. (literal and idiomatic)
- e #The bucket was kicked by John. (only literal)

(Abeillé and Shabes 1990) point out that the heavy NP-shift applies regularly (their example):

- (4) a He took Mary's words into account.
- b He took into account all the nasty things Mary has been telling him when they were living together.

Some idioms can be modified internally, while others cannot.

- (5) a John kicked the bucket.
- b #John kicked the tragic bucket.
- (6) a John took the proposal into account.
- b #John took the proposal into thorough account.
- c #John took the proposal into (several) accounts.
- (7) a The FBI kept tabs on Jane Fonda.
- b The FBI kept close tabs on Jane Fonda.

However, it may be argued that (7) is not really idiomatic because *tabs* has the same meaning without the verb *keep*, for example in a newspaper headline *FBI: close tabs on Jane Fonda*. This is not possible with the phrase *the bucket*, as exemplified by the following impossible headline: *John: the bucket*.

In the idiom *make up one's mind*, the possessive pronoun must agree in gender and person with the subject of *make*. Modification of *mind* and passivization are not possible.

- (8) a He made up his mind.
- b #He made up her mind.
- c #He made up his bright mind.
- d #His mind was made up by him.

In idioms which are headed by a verb, the verb can of course have different inflectional forms:

- (9) a We take the proposal into account.
- b Bill takes the proposal into account.
- c Bill took the proposal into account.
- d After Bill had taken the proposal into account, ...
- e Taking John's proposal into account seemed easy, but ...

While the "frozen" complements involved in support-verb constructions or lexicalized metaphors can be taken up again by pronominalization, the same is not true for unanalyzable idioms.

- (10) a I took a walk today. If the weather is fine, I will take one tomorrow.
- b He pulled a string or two.
- c #If John doesn't kick the bucket today, he will kick it tomorrow.

The above examples illustrate the syntactic aspects of idiomatic expressions, which must be accounted for in a formal theory of grammar.

Semantically, the major characteristic of idioms is that they are meaningful linguistic units whose meaning is not computed by the usual functions for combining the meanings of constituent words. In other words, while idioms can be analyzed syntactically, they cannot be decomposed semantically. The meaning of *take into account* cannot be computed by the usual functions from the usual meanings of *take*, *into* and *account*.

This appears to be in contradiction with the Fregean Principle of Compositionality. However, it can well be argued that idioms are semantically unanalyzable lexical units, and therefore the Principle of Compositionality should not apply to their internal structure³.

Requirements of a formalism for the treatment of idioms

Formal treatments of idioms have been proposed in Lexicon Grammar (Gross 1986), Tree Adjoining Grammar (Abeillé and Shabes 1990), and Lexical-Functional Grammar (Bresnan and Kaplan 1982), Categorical Grammar (van der Linden 1989), and other frameworks.

In all of these approaches, idioms involve a lexical head and some "frozen" complements. All of the analyses assign internal structure to the frozen complements. While the approaches of Gross and Abeillé/Shabes represent idiom patterns, the LFG approach is based on subcategorization for the form of the frozen complement.

Analyses based on subcategorization are economical because they allow lexical rules to account for processes like passivization and inflection.

³Zadrozny (1992) has proved that some compositional meaning function can be found for any semantics, for example one could construct a compositional semantics that constructs the meaning of a word from the meanings of its phonemes. The concept of compositionality is formally vacuous unless it is restricted to a particular class of functions. For lack of a formal specification of the class of functions which do not account for idioms, the term "usual function" is used here.

On the other hand, approaches which take (partially specified) trees as primitive units, like that of Tree Adjoining Grammar do quite well in describing the variable syntactic flexibility of idioms.

In the following, an representation of idioms is proposed which involves both subcategorization and the use of partially specified trees.

A Head-driven Approach

The approach presented here takes as a theoretical framework Head-Driven Phrase Structure Grammar (Pollard and Sag 1987)⁴. In HPSG, lexical entries are represented as feature structures of type *lexical-sign*, which have three attributes: PHON (the phonology of the sign, usually just represented by a string of graphemes), SYN (the syntactic information) and SEM (the semantic information). Among the local features of SYN (those features that are not concerned with long-distance dependencies) are HEAD, which contains information about grammatical category, case etc., and SUBCAT which encodes the valency of the sign (a list of signs which the sign needs to combine with).

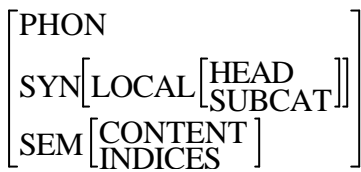


Figure 1: Structure of a lexical sign

Phrases and sentences are represented by feature structures of type *phrasal-sign*, which have an additional attribute DAUGHTERS, by means of which the constituent structure is encoded. The DAUGHTERS feature (abbreviated DTRS) is subdivided into head-daughter (a sign), complement-daughters (a list of signs) and adjunct-daughters (a set of signs).

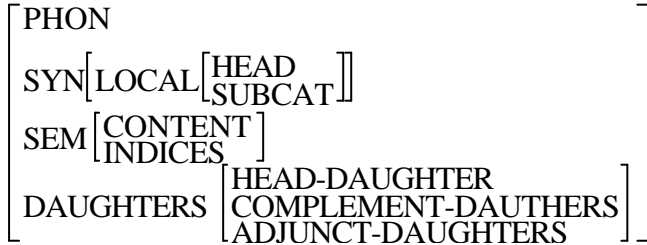


Figure 2: Structure of a phrasal sign

HPSG is a principle-based theory of grammar. A sign is licensed by the grammar if and only if it instantiates one of the lexical or phrasal signs (lexical entries L or rule schemata R) of the grammar and satisfies all universal and language-specific principles P of the grammar. One of the goals principle-based grammar theories like HPSG is to use only a small number of rule schemata, to which the principles of the grammar apply.

English = $P_1 \dots P_{m+n} (L_1 \Delta \dots \Delta L_p \Delta \dots \Delta R_1 \Delta \dots \Delta R_q)$ (Pollard and Sag 1987, page 44)

Among the universal principles are the Head Feature Principle (figure 3), the Subcategorization Principle (figure 4), and the Semantics Principle (figure 13). The Constituent Order Principle is one of

⁴For this paper, we use the 1987 version of HPSG because it is most well-known. For the treatment of idioms in a more recent variant of HPSG, see (Erbach and Krenn, in preparation)

the language-specific principles of HPSG. Principles are formulated as conditional feature structures. The Head Feature Principle states that the head features of a constituent are the same as the head features of its head.

phrasal-sign []

$$\left[\begin{array}{l} \text{SYN|LOC|HEAD } [1] \\ \text{DTRS|HEAD-DTR|SYN|LOC|HEAD } [1] \end{array} \right]$$

Figure 3: Head-Feature Principle⁵

HPSG appears to be well suited for the description of idioms because it provides the necessary descriptive devices: a subcategorization list, and the possibility to describe tree structures by means of the DAUGHTERS attribute.

We will discuss

- the syntactic specification of idioms,
- the semantics treatment of idioms,
- restrictions on passivization,, pronominalization etc.

and propose modifications to the theory of HPSG in order to handle these phenomena.

Syntax

Subcategorization in HPSG is handled by the universal Subcategorization Principle, which states that the SUBCAT list of the head of a constituent is equal to the concatenation of the list of complements realized as siblings to the head with the SUBCAT list of the constituent.

$$[\text{DTRS headed-structure} []] \Rightarrow \left[\begin{array}{l} \text{SYN|LOCAL|SUBCAT } [2] \\ \text{DTRS } \left[\begin{array}{l} \text{HEAD-DTR|SYN|LOC|SUBCAT append}([1],[2]) \\ \text{COMP-DTRS } [1] \end{array} \right] \end{array} \right]$$

Figure 4: Subcategorization Principle

The following is a part of the information of the lexical entry of the verb *kick*. It encodes that the verb has the phonetic form *kick*, the major syntactic category verb, and a SUBCAT list with two NPs.

$$\left[\begin{array}{l} \text{PHON /kick/} \\ \text{SYN|LOC } \left[\begin{array}{l} \text{HEAD|MAJ V} \\ \text{SUBCAT } \langle \text{NP}, \text{NP} \rangle \end{array} \right] \end{array} \right]$$

Figure 5: lexical entry of *kick*

Are idioms phrasal or lexical signs?

A sign in HPSG theory can be a lexical entry (lexical sign) or a grammar rule (phrasal sign). Since idioms involve more than one word, it would be a natural move to represent them as phrasal signs.. The following would be the "kick-the-bucket-rule".

⁵Coreference is indicated by shared variables, which are written as boxed numbers, e.g., [1].

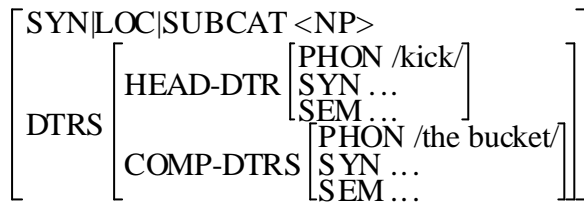


Figure 6: kick-the-bucket-rule

Such an rule-based approach is not only in contradiction with the spirit of keeping the number of rules as small as possible - moreover, it does not allow to handle processes like inflection or passivization by the lexical rules that apply to non-idiomatic entries. Instead, metarules would have to be introduced to capture the same regularities. For these reasons, we favour an analysis of idioms based on subcategorization of the "frozen" complements.

In the case of idioms which involve a lexical head and some of its complements, we will represent all information about the idiom in the lexical entry of the head, which means that we must represent the information about the complements in the SUBCAT list of the head.

First we consider idioms, which involve a verb and a fixed complement (e. g. *take into account*, *kick the bucket*). For these cases, it would seem natural to exploit the PHON feature of HPSG, which would play a role similar to the FORM feature proposed for LFG. In figures 7 and 8, V, BSE, PAS, PP, NP abbreviate the feature bundles that correspond to Verb, Base Form, Passive, Prepositional Phrase and Noun Phrase, respectively, and parentheses around a SUBCAT element denote optionality.

take: V , BSE , SUBCAT < [PHON /into account/] NP, NP >

Figure 7: lexical entry of *take* (*into account*)

Lexical rules like that of passivization can apply to the entry of *take*, thus deriving the SUBCAT list for the passive form shown in figure 8.

taken: V , PAS , SUBCAT < (PP[by]) , [PHON /into account/], NP >

Figure 8: SUBCAT list of *taken* (*into account*)

However, evidence from German suggests that the PHON feature is not appropriate to represent information about the frozen complements, because the phonological information may change if the frozen complement receives a nominative case by passivization, instead of accusative in the active form.

- (11) Sie machten ihm den Garaus.
Der Garaus wurde ihm gemacht.

This example shows that idioms involve selection of lexemes rather than the selection of phonologies. In order to account for this, we introduce a new head feature LEXEME into HPSG. LEXEME is a head feature because in the case of modifiable frozen complements, it is the lexical head of the frozen complement that is selected by the head verb of the idiom. In case of the idiom *kick the bucket*, the fact that only the form *the bucket* can occur is ensured by specifying the head lexeme as *bucket*, the syntactic features definite and singular; modification is excluded by requiring the empty set as the value for adjunct daughters.

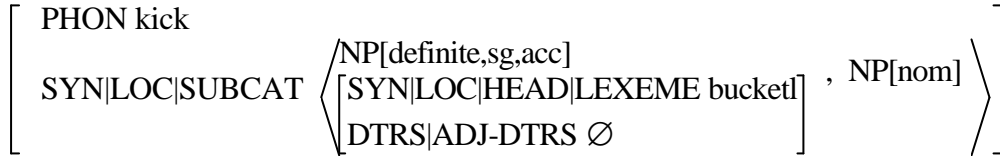


Figure 9: lexical entry for *kick (the bucket)*

In order to represent *take into account* with the mechanism proposed here, the phrase *into account* must be represented as a partial tree with the head preposition *into* (reflected by the head feature [PFORM into], and the complement daughter *account*, specified by [LEXEME account] and [LEX +].

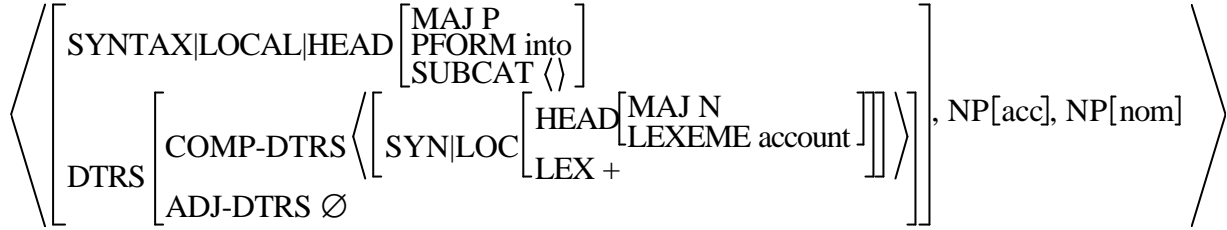


Figure 10: subcat list for *take (into account)*

Of course, adverbial modification is still possible, as in the sentence *John never took the facts into account*.

In the formal account of idioms presented here, we treat frozen adjuncts, as in *take the bull by the horns*, as subcategorized for.

Cases of idioms or conventionalized metaphor which can be modified internally, like *pull strings*, are easier to handle because modification need not be explicitly excluded.



Figure 11: lexical entry for *pull (string(s))*

SUBCAT lists like those in figures 9 and 10 violate the Locality Principle of HPSG (Pollard and Sag 1987, pages 143-144), which states that

the SUBCAT elements of lexical signs specify values for SYNTAX and SEMANTICS, but crucially not for the attribute DAUGHTERS.

The motivation behind the Locality Principle is that verbs do not select for particular phrasal configurations, but only for local properties like category, case etc. Even with the new head feature LEXEME, which allows the selection of particular lexemes, it is still necessary to use the attribute DAUGHTERS to block modification of the "frozen" complements. It should be pointed out that a semantic treatment of idioms is not a solution, because it would not be able to distinguish *kick the bucket* from *kick the pail*. According to Pollard and Sag, the Locality Principle "makes strong empirical claims about the full range of lexical dependency phenomena", however, it restricts lexical

dependency to non-idiomatic expressions. Therefore, it appears that the Locality Principle is too strong a constraint on lexical entries.⁶

Semantics

In HPSG semantics is in general handled by unification rather than functional application. This is best shown by the lexical sign for the literal reading of *kick*. The notation $NP_{[1]}$ denotes an NP with the semantic value (value of path SEM|CONT) $[1]$. In the lexical entry for *kick*, it is lexically specified that the semantics of the object NP (denoted by the variable $[2]$) is assigned as the KICKEE argument to the relation kick, and the semantics of the subject NP (denoted by the variable $[1]$) as the KICKER argument.

$$\left[\begin{array}{l} \text{PHON } /kick/ \\ \text{SYN|LOC|SUBCAT } \langle NP_{[2]}, NP_{[1]} \rangle \\ \text{SEM|CONT } \left[\begin{array}{l} \text{RELN kick} \\ \text{KICKER } [1] \\ \text{KICKEE } [2] \end{array} \right] \end{array} \right]$$

Figure 12: lexical entry for *kick*

For non-quantified expressions, the semantics of a projection of *kick* is identical with the semantics of *kick*, where the KICKER and KICKEE arguments are instantiated with the semantic values of the subject and object, respectively.

The Semantics Principle (Pollard and Sag 1987, page 110) deals with both quantified and non-quantified expressions.

$$[DTRS_{\text{headed-structure}} []] \Rightarrow \left[\begin{array}{l} \text{SEM} \left[\begin{array}{l} \text{CONT successively-combine-semantics}([1], [2]) \\ \text{INDICES collect-indices}([3]) \end{array} \right] \\ \text{DTRS } [3] \left[\begin{array}{l} \text{HEAD-DTR|SEM|CONT } [1] \\ \text{COMP-DTRS } [2] \end{array} \right] \end{array} \right]$$

Figure 13: Semantics Principle (Pollard and Sag 1987, page 110)

collect-indices is a function that returns the set union of all referential indices of the daughters.

combine-semantics is a function that deals with quantification and takes two arguments: a head and its complement. combine-semantics returns the semantic value of the head in case the complement is not a quantifier expression, and returns a quantified expression otherwise.

combine-semantics(A,B) =
 if A has type *circumstance* and
 B has type *quantifier*
 [QUANT A]
 then return [SCOPE B]

⁶In a more recent version of HPSG, the effect of the locality principle is achieved by making the subcat list a list of feature structures of type SYNSEM rather than type sign. SYNSEM is an attribute that conflates the SYN and SEM attributes, but not PHON or DAUGHTERS. This move would make our representation of idioms in HPSG impossible, and should therefore be reconsidered.

else return A

successively-combine-semantics is the recursive version of combine-semantics, which takes as arguments a head and a list of complements.

```

successively-combine-semantics(A,L) =
if length(L) = 0
then return A
else
  successively-combine-semantics(combine-semantics(A,SEM|CONT of first(L), rest(L)).

```

In order to handle the semantics of idioms we need to account for the fact that the meaning of an idiom is not derived from the meaning of its constituents. Since the compositional semantics of HPSG is encoded in the lexical head, this semantic behavior can be easily encoded, as illustrated in the following entry for *kick the bucket*.

$$\left[\begin{array}{l} \text{PHON kick} \\ \text{SYN|LOC|SUBCAT} \left\langle \begin{array}{l} \text{NP[definite,sg,acc]} \\ \text{[SYN|LOC|HEAD|LEXEME bucket]} \\ \text{DTRS|ADJ-DTRS } \emptyset \end{array} \right\rangle, \text{NP}_{[1]} \right\rangle \\ \text{SEM|CONT} \left[\begin{array}{l} \text{RELN die} \\ \text{VICTIM}_{[1]} \end{array} \right] \end{array} \right]$$

Figure 14: lexical entry for *kick (the bucket)*

For the sentence *John kicked the bucket*, the semantic index of *John* is unified with the VICTIM role of the relation DIE, while the semantic index of *bucket* does not play a role in the semantics of the sentence, resulting in the structure shown in figure 15.

$$\left[\begin{array}{l} \text{SEM} \left[\begin{array}{l} \text{CONT} \left[\begin{array}{l} \text{RELN die} \\ \text{VICTIM}_{[1]} \end{array} \right] \\ \text{INDS} \left\{ \begin{array}{l} \text{VAR}_{[1]} \\ \text{REST} \left[\begin{array}{l} \text{RELN naming} \\ \text{NAME /john/} \\ \text{NAMED}_{[1]} \end{array} \right] \end{array} \right\} \end{array} \right] \end{array} \right]$$

Figure 15: desired semantic representation of *John kicked the bucket*

However, if we take a look at the Semantics Principle, we will see that this is not the resulting structure, because of the functions **collect-indices** which collects the indices of all referential expressions in the sentence, and **combine-semantics** (and its recursive version successively-combine-semantics) which deals with quantification.

Since *the bucket* is a quantified referential expression both functions are applied yielding the following structure.

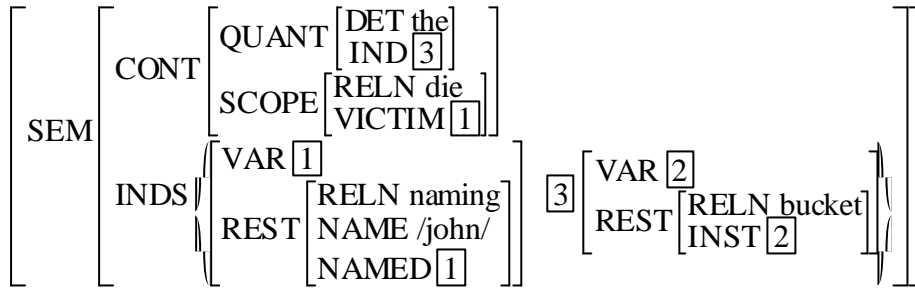


Figure 16: undesirable sign for *John kicked the bucket*.

This sign is undesirable for two reasons: First it includes a counter-intuitive vacuous quantification (There is a definite bucket, such that John dies). The second problem is that an index for the bucket is introduced, which can serve as an antecedent to a pronoun, as in (12).

(12) John kicked the bucket. It was rusty.

This text sequence should allow only the literal reading of *kick the bucket*, because in the idiomatic reading, there is no bucket which can be referred to by the pronoun *it*.

In order to overcome this problem, we modify the semantic principle in such a way that it only collects quantifies and indices of those complements that are assigned a thematic role by the verb. The lexical entry for verbs must be modified in such a way that verbs assign thematic roles to their complements.

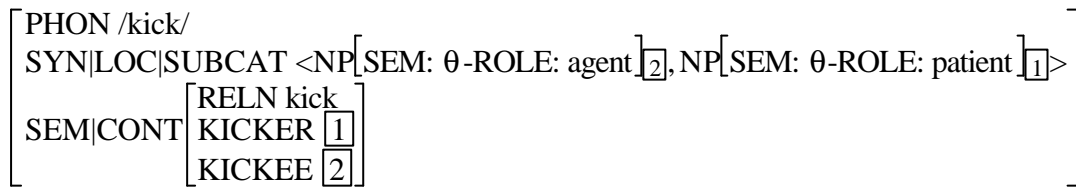


Figure 17: lexical entry for *kick*

There is independent evidence for thematic roles on the arguments of verbs: they are needed also for the statement of word order principles (Uszkoreit 1987). In order to account for the fact that idioms like *kick the bucket* do not assign a thematic role to their frozen complements, the value nil is used for the θ-role argument.

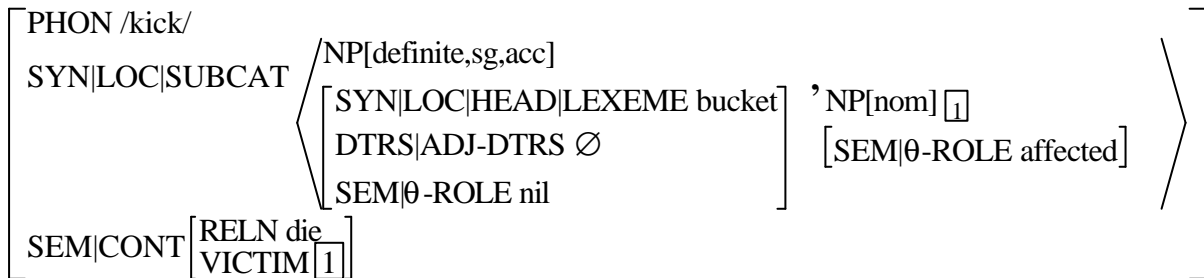


Figure 18: final lexical entry for *kick (the bucket)*.

The functions **combine-semantics** (and its recursive version successively-combine-semantics) and **collect-indices** are modified in such a way that they only combine or collect those referential NPs which are assigned a thematic role by the head. This is ensured by the additional condition that the thematic role of the complement must not be nil.

combine-semantics(A,B) =

```

if      A has type circumstance and
      B has type quantifier
  B δ [q-role nil]
    [QUANT A]
then return [SCOPE B]
else return A

```

The function **collect-indices** is modified such that it returns the set union of all referential indices of the daughters **which are assigned a thematic role**.

The next problem we want to address is that of possessive phrases within idioms which must agree in gender and person with another subcategorized phrase. Since gender and person agreement in English are semantic, this problem is treated in the section on semantics. In the idiom *lose one's mind*, the possessive pronoun must agree in gender and person with the subject of *lose*, and *mind* must agree in number with the subject, as illustrated by the following examples.

- (13) a John lost his mind.
 b #John lost her mind.
 c #John lost their mind.
 d They lost their minds.
 e #They lost their mind.

In the following lexical entry, POSP abbreviates whatever formal analysis is adopted for possessive phrases. The POSSESSOR role of the possessive pronoun is coreferential with the semantic index of the subject. The correct form of the possessive pronoun is then selected on grounds of semantic compatibility. The fact that a pronoun is chosen instead of another possessive phrase is guaranteed by the binding theory.

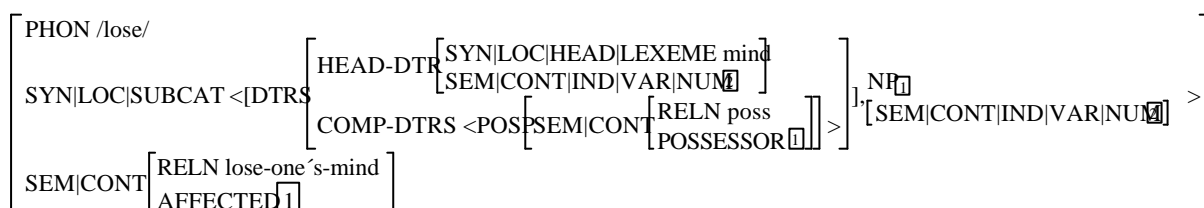


Figure 19: lexical entry for *lose one's mind*

Passivization of idioms

The introduction of thematic roles that was needed in order to make the Semantics Principle work properly for idioms can also account for the fact that some idioms like *kick the bucket* do not passivize.

In HPSG, passivization is handled by a lexical rule which takes the base form of a transitive verb as input and returns the past participle form of the verb. The accusative object of the base form is made the subject of the past participle form, and the subject of the base form becomes the optional BY-object.

There is evidence that the possibility of passivization is related to verb classes and possibly thematic role assignment. The following classes of verbs, called complex intransitives by Bach (1980), cannot passivize:

• Verbs of possession

John has a car

* A car is had by John

• **Lexical reflexive as object**

Mary broke a leg.

* A leg was broken by Mary.

• **Expletive pronoun as object**

I can't stand it any longer.

* It can't be stood by me any longer.

• **Predicative or copulative verbs**

John resembles Bill.

* Bill is resembled by John.

The cheese weighs 15 pounds.

* 15 pounds are weighed by the cheese.

A discussion of the thematic roles involved in these examples is beyond the scope of this paper, but it seems quite plausible that these examples can be handled by ensuring that the passivization lexical rule only applies to subjects and objects that are assigned certain thematic roles.

For idioms, however, the case is easier. We explain the observation that *kick the bucket* does not passivize by the fact that the lexical rule cannot apply to subjects and accusative objects that are not assigned any thematic role. In the case of *take into account*, passivization is possible because the accusative object does have a thematic role. The same is true of some support verb constructions, and lexicalized metaphors.

Conclusion

We have proposed a formal representation of idioms which accounts for their syntactic and semantic behavior. Idioms, collocations, lexicalized metaphors and support-verb constructions can be given a uniform syntactic treatment, but have different semantic analyses.

The semantic properties of idioms can be encoded in the lexical entry of the head of the idiom. Problems relating to quantification, referential indices and passivization can be solved by introducing thematic roles for complements (which are needed independently for the statement of word order constraints).

By presenting a formal treatment of idioms in HPSG, we think we have come one step closer to making HPSG a truly universal theory of grammar.

Problems that need to be addressed are the efficient processing of idioms and the resolution of ambiguity between the literal and the idiomatic reading. Solutions to these problems have been proposed by Stock (1989) and van der Linden (1991).

References

[Abeillé and Shabes 1990]

Abeillé, Anne and Yves Shabes. Non Compositional Discontinuous Constituents in Tree Adjoining Grammar. Symposium on Discontinuous Constituents, Tilburg University, January 1990.

[Bach 1980]

Bach, Emmon. In Defense of Passive. In: *Linguistics and Philosophy* 3, pp. 297 - 341.

[Boatner and Gates 1975]

Boatner, M. T. and J. E. Gates. *A dictionary of American Idioms*. (Revised edition: A. Makkai). Woodbury, New York and London. 1975.

[Bresnan and Kaplan 1982]

Bresnan, Joan and Ronald Kaplan. *Lexical-Functional Grammar: A Formal System for Grammatical Representation*. In: Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge MA. 1982.

[Coseriu 1967]

Coseriu, E. *Lexikalische Solidaritäten*. In: *Poetica* 1, pp. 293 - 303.

[Cowie, Mackin and McCraig 1975]

Cowie, F., R. Mackin and L. McCraig. *Oxford Dictionary of Current Idiomatic English*. Oxford University Press, London. 1975.

[Erbach and Krenn, in preparation]

Erbach, Gregor and Krenn, Brigitte. *Idioms and support-verb constructions*. To appear in: John Nerbonne, Klaus Netter and Carl Pollard (eds.): *German Grammar in HPSG*, CSLI Lecture Notes, Stanford, CA.

[Ettinger 1977]

Ettinger, St. *Kollokationen und Sprachpraxis*. In: *Lebende Sprachen* 22, pp. 151 - 152.

[Gross 1986]

Gross, Maurice. *Lexicon-Grammar - The Representation of Compound Words*. Proceedings of COLING 1986, Bonn.

[Korhonen 1987]

Korhonen, J. (ed.). *Beiträge zur allgemeinen und germanistischen Phraseologieforschung*. Department of Germanic Languages, University of Oulu (Finland). 1987.

[Martin 1991]

James H. Martin. *MetaBank: A Knowledge Base of Metaphoric Language Conventions*. In: Dan Fass, Elizabeth Hinkelman and James Martin (eds.). *Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts, Implicature*. Sydney, 1991.

[Pollard 1989]

Pollard, Carl. *The Syntax-Semantics Interface in a Unification-Based Phrase Structure Grammar*.

[Pollard and Sag 1987]

Pollard, Carl and Ivan Sag. *Information-based Syntax and Semantics. Volume 1: Fundamentals*. CSLI Lecture Notes Number 13, Stanford CA, 1987.

[Rothkegel 1989]

Rothkegel, Annely. *Polylexikalität: Verb-Nomen-Verbindungen und ihre Behandlung in EUROTRA*. EUROTRA-D Working Papers No. 17. Saarbrücken 1989.

[Rothkegel 1973]

Rothkegel, Annely. Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse. Niemeyer, Tübingen FRG. 1973.

[Stock 1989]

Stock, Oliviero. Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind. Computational Linguistics 15:1, 1989.

[Uszkoreit 1987]

Uszkoreit, Hans. Word Order and Constituent Structure in German. CSLI Lecture Notes Number 8, Stanford CA, 1987.

[van der Linden 1989]

van der Linden, Erik-Jan. Idioms and Flexible Categorical Grammar. In: Everaert, M. and van der Linden, E.-J. (eds.). Proceedings of the First Tilburg Workshop on Idioms, ITK, Tilburg, Netherlands, 1989.

[van der Linden 1991]

van der Linden, Erik-Jan. Idioms, Non-literal Language and Knowledge Representation. In: Dan Fass, Elizabeth Hinkelman and James Martin (eds.). Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts, Implicature. Sydney, 1991.

[Wood 1979]

Wood, F. T. Dictionary of English Colloquial Idioms (Revised by R. J. Hill). The Macmillan Press Ltd. London. 1979.

[Wotjak 1985]

Wotjak, B. Zu Inhalts- und Ausdrucksstruktur ausgewählter somatischer Phraseolexeme. In: Deutsch als Fremdsprache 22, pp. 216 - 223, 270 - 277.

[Zadrozny 1992]

Wlodek Zadrozny. On compositional semantics. Proceedings of COLING, Nantes.