1

Idiomatic Constructions in HPSG

Susanne Riehemann
Stanford University
sr@csli.stanford.edu
Draft of September 1997

In this paper I present an approach to idioms in the HPSG framework. Building on earlier work by Copestake 1994, it employs phrasal types that specify the semantic relationship between the idiomatic words involved. Underspecified Phrasal Semantics (UPS) does not require separate lexical entries for the words occurring in idioms, and treats a wide range of data more successfully than previous alternatives. The approach allows for the variability that some idioms exhibit, while being able to express what is fixed. It provides a solution to a problem discussed in McCawley 1981 which involves idioms occurring distributed over a main clause and a subordinate clause. The available psycholinguistic evidence seems to be consistent with the approach, and it is intuitive to view an idiom like *spill the beans* as a whole, comprising the three words *spill*, *the*, and *beans*, without giving these words an independent existence outside the idiom. The alternative view of the idiom as a special form of *spill* that happens to occur only together with the words *the* and *beans* does not provide a representation for the idiom as a whole, and requires a mechanism to ensure that the parts of the idiom that are being subcategorized for cannot occur by themselves. The reason that this has been the predominant view is that it was thought to be impossible to deal with the variation data in a phrasal approach.

The UPS approach can also deal with semantically decomposable and non-decomposable idioms (Nunberg et al. 1994).[1] Examples of decomposable idioms are *pull strings* and *spill the beans*, where *spill* means

[1] I use this different terminology to emphasize the analytic perspective of being able to distribute the meaning over the parts of the idiom rather than the synthetic one of combining the parts to build up the meaning.

something like ‘reveal’ and *beans* means something like ‘secret’; while typical non-decomposable idioms are *saw logs* and *kick the bucket*, which means something like ‘die’ and is a one-place relation in which *bucket* plays no role. So ‘non-decomposable’ has nothing to do with whether one can guess the meaning of an idiom or its metaphorical motivation. Instead it means that parts of the meaning of the idiom are associated with parts of the idiom.[2]

1.1 Problematic Properties of Idioms

Idioms have two main properties that are hard to account for in various approaches: they tend to be syntactically variable, and they sometimes involve fixed items that go beyond simple head-complement relationships.

Most idioms are variable to some extent and cannot be seen as simple strings of words. Nevertheless, it is desirable to list an idiom only once, and use independently existing mechanisms of the grammar to derive the variations. In particular, inflectional information for idiomatic usages of words is identical to that of nonidiomatic usages, and should not have to be repeated. Therefore any approach has to establish some sort of link to the literal lexical entries.

The pieces of many idioms can appear separated from each other in passive, raising, and topicalization constructions:

(1) Once the cat was let out of the bag everyone was happy.

This is not a problem for any word-level approach because lexical rules could apply normally. But it rules out phrasal approaches which fix the phonology or syntax of the phrase. In the past, this has often led to the assumption that idioms must be represented at the word level.

But some idioms occur in variations that do not seem to correspond to any lexical rules:

(2) a. Put the cat among the pigeons.
b. The cat is among the pigeons.

Since no particular verb is part of this idiom, there is no lexical entry at the word level where the relevant relationship could be stated, and it cannot be expressed syntactically at the phrasal level, either.

In (3) the idiom is *tie up loose ends*, but *loose ends* is not the com-

[2] Note that for the purposes of this paper it is not relevant whether all idioms can be classified into these two types without problems—there might be some variation depending on which paraphrase is picked, and not all speakers perceive these idioms the same way. But as long as idioms of both types exist, it is necessary to have an account of them.

plement of *tie* in some of these examples, and there is no way to derive them by lexical rule.

(3) a. I've got some loose ends to be tied up.
b. I'm tying up a few loose ends.
c. A few loose ends need tying up.

Also, as McCawley 1981 observed, parts of idioms can be spread over a main clause and a subordinate clause:

(4) a. The strings that Pat pulled got Chris the job.
b. I objected to the close tabs the FBI kept on Sandy.

This is a problem even for a word-level semantic approach, since the relative pronoun does not meet the subcategorization requirement—the *index* is shared between it and the modified noun, but not the semantic *relation*.[3]

Sometimes adjectives can be inserted syntactically as in:

(5) a. He kicked the proverbial bucket.
b. I'll keep a close eye on his progress.

Some adjectives can even internally modify parts of idioms, as in (6). This is possible only with decomposable idioms, in which the individual words carry parts of the meaning.

(6) a. He cut through a lot of red tape.
b. They were skating on very thin ice.

Another problem for word-level approaches is that idioms sometimes involve fixed items that go beyond mere head-complement relationships. They can include adjectives and specifiers:

(7) a. bark up the wrong tree
b. give someone some skin

And even PPs that are modifiers:

(8) a. to put it mildly
b. skate on thin ice

Information about adjuncts is not available in word-level approaches, unless one thinks all adverbs and other adjuncts that occur in idioms

[3] This data is actually part of a paradox for transformational approaches to idioms discovered by McCawley 1981. The paradox arises under the assumption that idiomatic elements have to be adjacent in D-structure. It might be possible to explain (4a), if it is assumed that *the strings* are adjacent to *pull* at D-structure and raised out of the relative clause. But then sentences like *Pat pulled the strings that got Chris the job* should not have an idiomatic interpretation, since *the strings* would not be in the main clause at D-structure. There is no one single set of assumptions about movement that can accommodate the acceptability of both these examples.

can be analyzed as complements or are otherwise available in the valence of verbs.

At the far end of the spectrum, in proverbs, basically anything can be fixed. While some proverbs seem almost completely fixed, others are more variable and idiom-like. Even relatively fixed proverbs like (9a) can preserve their proverbial interpretations in spite of some variations, as in (9b).

(9) a. When the cat's away the mice will play.

b. When the cat's away the mice tend to play.

1.2 Previous Approaches

The alternative approaches are classified in two dimensions: whether they represent idioms at the word level or view them as phrasal, and whether the kind of information that gets specified is syntactic or semantic. Approaches that view idioms as specifying a particular phonological value run into problems even with simple inflectional variation, so these approaches are not discussed here.

In this paper the semantic information of signs is expressed in MRS (Minimal Recursion Semantics), as developed in CSLI's English Resource Grammar Online (ERGO) project and described in Copestake et al. 1995 and Copestake et al. 1997. Most semantic information in MRS is contained under the feature LISZT, which takes a list of *rel*s (relations) as its value. Verb *rel*s have a feature EVENT, which takes a Davidsonian event variable, and *index*-valued features such as ACT(or) and UND(ergoer). Common nouns have *rel*s with the feature INST, which takes an *index* as its value. Each *rel* also has a feature HANDEL, which is used to simulate embedding using ARG(ument) values and represent scope information. The Semantics Principle ensures that the LISZT of a phrase is formed by appending the LISZTs of its daughters. The feature KEY points to the main relation in a *liszt*. For example, for an NP that would be the *rel* of the head noun, and for a PP the *rel* of the preposition. Idiomatic senses of a word are indicated by *i_*, for example, the *i_bean_rel* of the word *bean* in *spill the beans* is neither the same as *_bean_rel* nor *_secret_rel*.

1.2.1 Word-Level Approaches

Subcategorizing for the Syntax

(10)
$$\left[\begin{array}{l} \textit{spill\_beans\_verb} \\ \textsc{phon} \left\langle \textit{spill} \right\rangle \\ \textsc{synsem}\,|\,\textsc{loc} \left[\begin{array}{l} \textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps} \left\langle \textbf{NP}\left[\textsc{head-dtr}\ \textit{i\_bean}\right] \right\rangle \\ \textsc{cont}\,|\,\textsc{liszt} \left\langle \textit{i\_spill\_rel} \right\rangle \end{array}\right] \end{array}\right]$$

Subcategorizing for the Semantics

(11)
$$\left[\begin{array}{l} \textit{spill\_beans\_verb} \\ \textsc{phon} \left\langle \textit{spill} \right\rangle \\ \textsc{synsem}\,|\,\textsc{loc}\,|\,\textsc{cat}\,|\,\textsc{val}\,|\,\textsc{comps} \left\langle \textbf{NP}\left[\textsc{loc}\,|\,\textsc{cont}\,|\,\textsc{key}\ \textit{i\_bean\_rel}\right] \right\rangle \end{array}\right]$$

One problem with word-level approaches is that they violate the locality principle, because they require subcategorizing for whole signs, to allow access to other information via the DTRs, e.g. a fixed PP complement of a noun (*scrape the bottom of the barrel*). There are also cases where more than just the head noun is fixed, but variation is still possible, e.g. *bark up the same wrong tree*, which makes locating the head within the DTRs impossible.[4]

These approaches also require an idiomatic *i_bucket_rel*. This is not appropriate for non-decomposable idioms like *kick the bucket* which have an unanalyzed *i_kick_bucket_rel* in their semantics and do not distribute this meaning over their syntactic parts. Furthermore, word-level approaches need an additional mechanism to make sure that parts of idioms cannot occur by themselves. Without such a mechanism sentences like (12) are predicted to occur with the idiomatic meaning *secret_rel* for *beans*.

(12) I heard some very interesting beans yesterday.

Word-level approaches also require that all adverbs and adjuncts are made available on the complements list, and they are not powerful enough to handle the examples not related by lexical rules, those not involving verbs, or the McCawley data.

[4] Some sort of functional uncertainty would be needed but would require further constraints, since presence of the relevant items somewhere in the daughters is not sufficient—*... barked up the cat that was sitting in the wrong tree*.

1.2.2 Phrasal Approaches

We have seen that there is no word-level approach that can adequately handle all the data, and that there are various other problems associated with them. So a phrasal approach of some sort is needed, and there must be a different (functional or cognitive) explanation of the fact that many idioms involve only verbs and their immediate complements. We will now examine how phrasal approaches fare in comparison.

Partially Fixed Syntax

(13)
$$\left[\begin{array}{l} \textit{spill\_beans\_phrase} \\ \textsc{head-dtr}\ \textit{i\_spill} \\ \textsc{comp-dtrs}\ \left\langle \left[\textsc{head-dtr}\ \textit{i\_bean}\right] \right\rangle \end{array}\right]$$

This approach works only for idioms that do not passivize, because in a passive sentence the *beans* would not be among the COMP-DTRS. It also does not allow modification, because in that case the HEAD-DTR of the COMP-DTRS would not be the head noun *i_bean*, but the N' that includes the adjective. Again, some sort of functional uncertainty would help, but would need to be restricted semantically because occurrence somewhere in the phrase is not sufficient to prevent overgeneration. The approach also cannot handle the McCawley data.

Of course the UPS approach I am proposing is not inconsistent with syntactic specification. In fact, for idioms that are not syntactically flexible, both syntactic and semantic information needs to be specified.

Partially Fixed Semantics

(14)
$$\left[\begin{array}{l} \textit{spill\_beans\_phrase} \\ \textsc{synsem}\,|\,\textsc{loc}\,|\,\textsc{cont}\,|\,\textsc{liszt}\ \left\langle \ldots \left[\begin{array}{l} \textit{i\_spill\_rel} \\ \textsc{und}\ \boxed{1} \end{array}\right], \left[\begin{array}{l} \textit{i\_bean\_rel} \\ \textsc{inst}\ \boxed{1} \end{array}\right] \ldots \right\rangle \end{array}\right]$$

In this approach, which is similar to the one in Copestake 1994, only the semantic relationship between the parts of the idiom is specified. The UPS approach is of this general kind, but there are several differences, which have the effect of enabling the approach to handle non-decomposable idioms and do not require additional lexical entries for the parts of idioms, thereby avoiding the problem of how to ensure that they do not occur outside the idiom.

1.2.3 Inference

A related approach has been suggested in Pulman 1993. In this system, idioms are first parsed and assigned their literal compositional semantics. If this results in a logical form that entails the antecedent of one of the idiom rules, that rule can be applied. Here is an example of an idiom rule:

(15) $\forall x, y\ [cat(x) \wedge bag(y) \wedge out\_of(x, y)] \approx \exists a, z\ [secret(z) \wedge revealed(a, z)]$

This approach can deal with a broad range of variation data, including idiom variations not related by lexical rules, and probably the McCawley data. Unfortunately it overgenerates significantly. One problem is the fact that a complex indexing scheme is required to make sure that only particular lexical items and not others that might be connected via meaning postulates can trigger the idiom rules—this has to involve more than merely checking for the presence of those words somewhere in the sentence.

Also, it seems that it is not necessary or desirable to have the full power of inference available—(16) does not have an idiomatic meaning:

(16) The cat and the dog got out of the laundry bag.

Furthermore, there is no way in this approach to limit syntactic variation, and it can only handle idioms that have a literal parse.

Most of the psycholinguistic evidence about the processing of idioms is inconsistent with an approach that requires literal meanings to be computed first. It is hard to see how idioms can be understood faster if they require additional processing during the inference stage. It is also hard to see what would explain why the canonical forms of idioms are understood faster than other variants in an approach where there is no canonical form and no way of specifying anything about the form of an idiom.

1.3 The Proposed UPS Approach

In the proposed UPS approach *phrases* have a set-valued feature WORDS, which contains all the words in the phrase.[5] More specifically, the items in this set are structure-shared with the yield of the syntactic tree, which means that the valence of verbs is complete. Parts of the idiom are listed as members of this set of WORDS:

[5] This attribute might also be useful for Linearization approaches to syntax, although it is distinct from the DOM attribute used in the approach developed at OSU and elsewhere, which does not contain all words individually.

(17) $$\begin{bmatrix} \textit{spill\_beans\_idiom\_phrase} & \\ \text{WORDS} & \left\{ \begin{array}{l} \begin{bmatrix} \textit{i\_word} & \\ \ldots\text{KEY} & \begin{bmatrix} \textit{i\_spill\_rel} \\ \text{UND } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{spill} \end{bmatrix}, \\ \begin{bmatrix} \textit{i\_word} & \\ \ldots\text{KEY} & \begin{bmatrix} \textit{i\_bean\_rel} \\ \text{INST } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{bean} \end{bmatrix}, \ldots \end{array} \right\} \end{bmatrix}$$

The included signs are the ordinary literal lexical entries, which have only their semantic relations overwritten during compilation by skeptical default unification of the kind proposed in Carpenter 1993 and extended to typed feature structures in Lascarides and Copestake 1995. The description on the left side of the $\overset{<}{\sqcap}$ symbol contains the strict information that will be augmented with all the non-conflicting information from the description on the right of the symbol.[6]

Only the semantic relationships between the parts of the idiom are specified—the *beans* are the UNDERGOER of the *spill*ing. This indirectly fixes the syntax to some extent, since the UNDERGOER is usually the head noun of the first NP on the COMPS list. But it is compatible with modification (*he spilled every single bean*) and with syntactic variations like passives and topicalizations, which leave this semantic relationship unaffected. The McCawley data, i.e. idioms distributed over several clauses, can be handled since the parts again stand in the specified semantic relationship.

This approach has the advantage that there is no need to specify lexical entries for idiomatic senses of words, and that the only place where these meanings are listed is in the representation of the whole idiom. Therefore there is no need for an additional mechanism to ensure that these meanings do not occur by themselves. Cases where the same idiomatic sense of a word occurs in more than one idiom can be handled by relating them in the phrasal hierarchy of idioms.[7]

The approach predicts that words in idioms have the same morphology and syntax as their literal counterparts by default, and can occur in all inflected variants unless specified otherwise. It also has the virtue of

[6] Since all this can be precompiled there is no need for default unification at runtime, and the literal meanings are not present anywhere in the parse.

[7] It is possible that some of these idiomatic meaning components become so strongly associated with their words that they become separate lexical entries.

coming closer to explaining how some further semantic properties seem to be inherited from literal meanings.[8]

Variants not related by lexical rules can be handled because it is possible to specify only an underspecified relationship between the parts:

(18)

$$\begin{bmatrix} \textit{tie\_up\_loose\_ends\_idiom\_phrase} & \\ \text{WORDS} & \left\{ \begin{matrix} \begin{bmatrix} \textit{i\_word} & \\ \ldots\text{KEY} & \begin{bmatrix} \textit{i\_tie\_up\_rel} \\ \text{UND } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{tie\_up} \end{bmatrix}, \\ \begin{bmatrix} \textit{i\_word} & \\ \ldots\text{KEY} & \begin{bmatrix} \textit{i\_loose\_rel} \\ \text{ARG } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{loose} \end{bmatrix}, \\ \begin{bmatrix} \textit{i\_word} & \\ \ldots\text{KEY} & \begin{bmatrix} \textit{i\_end\_rel} \\ \text{INST } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{end} \end{bmatrix}, \ldots \end{matrix} \right\} \end{bmatrix}$$

It is not a problem that some idioms do not involve a fixed verb. In *up the creek without a paddle* or *butterflies in one's stomach*, the relevant relationship between the parts, e.g. the location of an underspecified event, can be expressed without stating to which verbal *rel* they belong. Or, if a preposition is involved, it can establish a link between its two arguments as in (19).

[8] Furthermore, in this approach there is one place that contains the metaphorical mapping—both the literal meanings of the words and their idiomatic meanings are present in the same representation. It is possible to imagine that there could be hierarchies of such mappings, with common metaphors making idioms which instantiate them easier to learn and remember.

(19)

$$\begin{bmatrix} \textit{cat\_among\_pigeons\_idiom\_phrase} \\ \text{WORDS} \left\{ \begin{array}{l} \begin{bmatrix} \textit{i\_word} \\ \ldots\text{KEY} \begin{bmatrix} \textit{i\_cat\_rel} \\ \text{INST } \boxed{1} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{cat} \end{bmatrix}, \\ \begin{bmatrix} \ldots\text{KEY} \begin{bmatrix} \text{ARG1 } \boxed{1} \\ \text{ARG2 } \boxed{2} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{among} \end{bmatrix}, \\ \begin{bmatrix} \textit{i\_word} \\ \ldots\text{KEY} \begin{bmatrix} \textit{i\_pigeon\_rel} \\ \text{INST } \boxed{2} \end{bmatrix} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{pigeon} \end{bmatrix}, \ldots \end{array} \right\} \end{bmatrix}$$

For non-decomposable idioms the words *kick*, *the*, and *bucket* do not contribute to the meaning of the idiom. Instead, the whole idiomatic construction contributes an *i_kick_bucket_rel*.[9] So no *i_bucket_rel* is required, which is appropriate for these idioms, which do not distribute their meaning over their syntactic parts.

(20)

$$\begin{bmatrix} \textit{kick\_bucket\_idiom\_phrase} \\ \text{WORDS} \left\{ \begin{array}{l} \begin{bmatrix} \textit{i\_word} \\ \ldots\text{SUBJ} \left\langle \begin{bmatrix} \ldots\text{KEY } \boxed{1} \end{bmatrix} \right\rangle \\ \ldots\text{COMPS} \left\langle \begin{bmatrix} \ldots\text{KEY } \boxed{2} \end{bmatrix} \right\rangle \\ \ldots\text{KEY } \textit{empty\_rel} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{kick} \end{bmatrix} \\ \begin{bmatrix} \textit{i\_word} \\ \ldots\text{KEY } \boxed{3} \textit{ empty\_rel} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{the} \end{bmatrix}, \\ \begin{bmatrix} \textit{i\_word} \\ \ldots\text{SPR} \left\langle \begin{bmatrix} \ldots\text{KEY } \boxed{3} \end{bmatrix} \right\rangle \\ \ldots\text{KEY } \boxed{2} \textit{ empty\_rel} \end{bmatrix} \overset{<}{\sqcap} \begin{bmatrix} \textit{bucket} \end{bmatrix}, \ldots \end{array} \right\} \\ \text{CXCONT} \mid \text{LISZT} \left\langle \begin{bmatrix} \textit{i\_kick\_bucket\_rel} \\ \text{ACT } \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

[9] The feature CXCONT is used in the ERGO grammar to encode the semantic contribution of constructions. The meaning of non-decomposable idioms can be seen as a special case of this. This has the advantage that the Semantics Principle applies as usual—the LISZT of a phrase contains the *rels* from all the daughters plus those of the CXCONT.

The necessary information about the literal words is available and can be used to restrict syntactic variation (for example these idioms do not passivize), and the non-idiomatic hierarchies can be exploited. This approach predicts that it is impossible for *proverbial* to modify *bucket* semantically, because there is no *_bucket_rel*. But there is no syntactic problem with a *proverbial bucket* because the exact location of *bucket* in the NP is not specified.[10]

There are no separate lexical entries for the idiomatic words that occur in these phrasal entries. This requires a parser that looks through the set of WORDS of idiomatic phrases as well as through the set of ordinary lexical entries. When a match is found, the entire idiom phrase needs to be present for the parse to succeed.

Idiom families like *throw someone to the lions/wolves/dogs* can be handled because the literal meanings are accessible, and it possible to express which variant with a particular choice of words or syntactic structure is the canonical form of an idiom, by making it a subtype of the more general representation of the idiom while allowing for further non-lexicalized variations.

The proposed approach can also be used for **constructions** like *what's X doing Y* (Kay and Fillmore 1995), for example (21a). In the UPS approach these can be described phrasally in spite of their syntactic flexibility (21b), because the syntax does not need to be fixed.

(21) a. What are your dirty feet doing on the breakfast table?

b. I don't know what Mary thought her feet were doing on the table.

[10] The mechanism for achieving the unusual wide scope of *proverbial* is not explored here, but it is needed independently for examples like *An occasional sailor walked in.* Further constraints are needed to exclude other adjectives that should be able to modify the whole sentence semantically, such as *he kicked the expected bucket*.

$$\left[\begin{array}{l}
\textit{what\_be\_doing\_phrase} \\
\text{WORDS} \left\{ \begin{array}{l}
\left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{LOC}\ \boxed{6}\ \left[\ldots\text{KEY}\ \textit{empty\_rel}\right] \end{array}\right] \overset{<}{\sqcap} \left[\textit{what}\right], \\
\left[\begin{array}{l} \textit{i\_word} \\ \ldots\text{SUBJ}\ \left\langle \boxed{7}\left[\ldots\text{CONT} \mid \text{HANDEL}\ \boxed{2}\right]\right\rangle \\ \ldots\text{COMPS}\ \left\langle \boxed{5}\left[\ldots\text{SLASH}\ \left\{\boxed{6}\right\}\right], \left[\begin{array}{l} \ldots\text{SUBJ}\ \left\langle\boxed{7}\right\rangle \\ \ldots\text{CONT}\ \boxed{4}\left[\text{HANDEL}\ \boxed{3}\right] \end{array}\right] \right\rangle \\ \ldots\text{CONT}\ \boxed{4} \end{array}\right] \overset{<}{\sqcap} \left[\textit{be}\right], \\
\left[\begin{array}{l} \textit{i\_word} \\ \text{SYNSEM}\ \boxed{5}\left[\ldots\text{KEY}\ \textit{empty\_rel}\right] \end{array}\right] \overset{<}{\sqcap} \left[\begin{array}{l} \textit{do} \\ \ldots\text{VFORM}\ \textit{prp} \end{array}\right], \ldots
\end{array} \right\} \\
\text{SYNSEM} \mid \text{LOC} \mid \text{CONT} \mid \text{HANDEL}\ \boxed{1} \\
\text{CXCONT} \mid \text{LISZT}\ \left\langle \left[\begin{array}{l} \textit{why\_rel} \\ \text{HANDEL}\ \boxed{1} \\ \text{ARG}\ \boxed{2} \vee \boxed{3} \end{array}\right], \left[\begin{array}{l} \textit{non\_canonical\_rel} \\ \text{HANDEL}\ \boxed{1} \\ \text{ARG}\ \boxed{2} \vee \boxed{3} \end{array}\right] \right\rangle
\end{array}\right]$$

In this representation for WXDY the *what* and *doing* do not contribute to the meaning of the construction. Instead, the construction meaning of WXDY is: why is X Y, and it is non_canonical that X is Y. The construction meaning does not have to be localized on the words that make up the construction and there are no problems with scope.[11] There is no need to write separate lexical entries for the non-standard *what* and *doing*, and they will not be able to occur outside of the construction.

As we have seen, the approach accommodates both decomposable and non-decomposable idioms, allows for the variability that idioms exhibit and the variety of types of information that can be fixed, and it is crucially needed for the McCawley data as well as for variations not due to lexical rules.

Acknowledgments

I would like to thank Ann Copestake, Eve Clark, Dan Flickinger, Andreas Kathol, Rob Malouf, Ivan Sag, Tom Wasow, Gert Webelhuth, Ed Zalta, the Berkeley-Stanford-Union City syntax group, and the audiences of my talks at the University of Canterbury and at the 1997 HPSG conference in Ithaca for useful discussions and comments.

[11] For example, if this construction were lexicalized on *be*, the 'non-canonical' aspect of the meaning would have to be located there, and the question meaning would probably take scope over it unless this can somehow be prevented.

Bibliography

Carpenter, Bob. 1993. Skeptical and Credulous Default Unification with Applications to Templates and Inheritance. In *Inheritance, Defaults and the Lexicon*, ed. E. J. Briscoe, A. Copestake, and V. de Paiva. Cambridge University Press.

Copestake, Ann. 1994. Representing Idioms. Presentation at the Copenhagen HPSG Workshop.

Copestake, Ann, Dan Flickinger, Rob Malouf, Susanne Riehemann, and Ivan Sag. 1995. Translation Using Minimal Recursion Semantics. In *Proceedings of The 6th International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven.*

Copestake, Ann, Dan Flickinger, and Ivan Sag. 1997. Minimal Recursion Semantics: An Introduction. Manuscript.

Kay, Paul, and Charles J. Fillmore. 1995. Grammatical Constructions and Linguistic Generalizations: The *What's X doing Y?* Construction. Manuscript.

Lascarides, Alex, and Ann Copestake. 1995. Order Independent Typed Default Unification. ACQUILEX-II working paper 60.

McCawley, James D. 1981. The Syntax and Semantics of English Relative Clauses. *Lingua* 53:99–149.

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

Pulman, Stephen G. 1993. The Recognition and Interpretation of Idioms. In *Idioms—Processing, Structure, and Interpretation*, ed. C. Cacciari et al. 249–270. Lawrence Erlbaum.