9.1 Motivation

Idioms are omnipresent in everyday language. Nonetheless, they have been widely neglected by linguists developing grammar fragments. And even where an account for idioms has been given, most approaches have their shortcomings (cf. Riehemann, 2001, ch. 4).

In this contribution we want to focus on decomposable and non-decomposable idioms and discuss technical aspects of an HPSG analysis. For reasons of space we will have to neglect detailed linguistic corpus data. By “idiom” we mean idiomatic expressions that do not form complete sentences as would be the case for e.g. His bark is worse than his bite.

(33) make waves (“cause trouble”)

(34) spill the beans (“divulge a secret”)

The expressions in (33) and (34) are instances of decomposable idioms, i.e. their meaning can be derived from the idiom parts. Note that idiom
parts are not necessarily to be understood literally. In (33), e. g., we can attribute the meaning “cause” to make and “trouble” to waves. The idiomatic meaning of the whole idiom consists of the idiomatic meanings of its parts.

Where this is not the case, an idiom is non-decomposable: the meaning of the whole phrase has nothing to do with the meaning of the words the idiom consists of. Consider (35) and (36):

(35) saw logs (“snore”)
(36) shoot the breeze (“chat”)

It is not clear how to assign the meaning “snore” to the words saw and logs, the same holds for “chat”.

After providing the prerequisites for a revised approach to idioms in the next section, we analyse instances of these idiom classes and discuss previous approaches. Then we briefly sketch how our proposal fits into the overall architecture of HPSG and illustrate this by examining how to merge it with a fronting analysis of German idioms.

9.2 Lexemes and Listemes

Before we present our analysis, we point out a way that enables us to select a specific word. This forms a prerequisite of our approach.

Idioms often consist of particular words which cannot be substituted by semantically equivalent terms. It seems in general that each word has a unique “identity” with an idiosyncratic behavior. The possibility to select a particular word would, thus, be a useful feature. Up to now, there has been a discussion about the necessity of having such kind of selection. One could argue that any data in question are to be handled as Constructions or collocations. But why impose such a “heavy thing” on an expression like to furrow one’s brow? Would it not be plausible that the verb furrow simply selects a word of the form brow? For perfect tense in German a main verb has to be combined with the right auxiliary (haben/sein; in HPSG with the attribute auxf, cf. Heinz and Matiasek, 1994, p. 222). Here one does nothing other than to select a particular lexeme.

Krenn and Erbach (1994) made an important contribution to idiom analysis within the HPSG framework. They suggested selecting particular lexemes via their feature LEXEME below CONTENT INDEX. This idea of having lexeme information in the CONTENT is questionable. A lexeme combines phonetic, morphological, syntactic and semantic properties all together, not only semantic information. Besides, their approach had several technical shortcomings (cf. Soehn and Sailer, 2003). We therefore propose that the LEXEME approach has to be discarded.
A different concept that can help to distinguish between individual words is that of a listeme\(^2\). As the concept holds the characteristic of listedness in a lexicon, we use it in our grammar to identify a particular word or phrase. Thus, we insert LISTEME into the feature geometry below CATEGORY, emphasizing the morpho-syntactic character of information. More precisely, we put it below HEAD. This has two consequences: firstly, it is available for selection, as a HEAD value is below SYNSEM. Secondly, the LISTEME value of a projection is the same as the one of the head, as all HEAD features “percolate” according to the HEAD-FEATURE-PRINCIPLE. For our furrow-example that means that a modified direct object his heavy brow still has the same LISTEME value as brow alone.

A third question to address is the handling of pronominalization. It is necessary that pronouns have the same LISTEME value as their antecedent.\(^3\) In Krenn and Erbach’s approach this was the major motivation of putting the LEXEME feature in the INDEX. To emulate this quality, we propose a constraint ensuring that each pronoun which is co-indexed with an antecedent takes over its LISTEME value. In the lexical entries of pronouns that value would be left underspecified in that way, that it consists of a disjunction of an identifying value (she, her, etc.) and a wildcard. In case of co-indexation the wildcard is identical to the LISTEME value of the antecedent and – by virtue of the constraint – becomes the actual and concrete LISTEME value of the pronoun. An informal description of such a pronoun constraint is illustrated in (37).

(37) **PRONOUN-LISTEME-CONSTRAINT:**

If a pronoun is co-indexed with an antecedent, it takes over the LISTEME value of that antecedent. Otherwise the LISTEME value of this pronoun is that of the other disjunct.

The value of LISTEME is an atomic sort as brow, heavy, furrow, take, she etc. In order to identify listemes for the same words having different meanings, we use numeric indices just as in a dictionary.

In summary, discarding the LEXEME approach, we propose a more adequate solution for the problem of selecting particular words, at least with respect to terminology, technical feasibility and the feature geometry. We introduce a feature LISTEME which is appropriate for the sort head taking atomic sorts as its value.

---

\(^2\)This term has been introduced by Di Sculpo and Williams (1988) for a sign that is listed in the lexicon.

\(^3\)E. g. in the phrase He furrowed it. the pronoun has the same LISTEME value as its antecedent, satisfying the subcategorizational requirement of the verb.
9.3 Licensing Contexts

Getting to the analysis, we have to define a second attribute in the feature geometry. We declare objects of sort sign to bear a list-valued feature COLL (context of lexical licensing), first introduced by Richter and Sailer (1999). The COLL list may contain objects of sort barrier. These barriers are particular nodes in the syntactic configuration, like XPs, complete clauses or utterances (a complete clause with an illocutionary force). The concept of barriers is borrowed from the tradition of generative grammar, where these form boundaries for government and binding principles. We avail ourselves of this concept and use similar barriers for the restriction of distributional phenomena.

Barrier objects have an attribute LOCAL-LICENSEE (LOC-LIC) which has a value of sort local. In the lexical entry of an idiomatic word one can thus specify a barrier on its COLL list with a specific local configuration. Subsorts of barrier are illustrated in figure 1. The subsorts of barrier correspond to nodes in the syntactic tree with particular properties.

The following relations identify the nodes which relate to the barriers complete-clause and vp, respectively.4

\[ \forall x \left( \text{is\_complete\_clause}(x) \leftrightarrow \left[ \begin{array}{l} \text{phrase} \\ \text{status complete} \\ \text{loc cat} \{\text{head verb}\} \\ \text{subcat elist} \end{array} \right] \right) \]

\[ \forall x \left( \text{is\_vp}(x) \leftrightarrow \left[ \begin{array}{l} \text{phrase} \\ \text{status incomplete} \\ \text{loc cat} \{\text{head verb}\} \\ \text{subcat nlist} \end{array} \right] \right) \]

The Licensing-Principle (informally in 38) makes sure that if there is a barrier specified on a word’s COLL list, there is an actual barrier in the phrase our word occurs in. This barrier must fulfill the local requirements and it has to be minimal, i.e., there is no other potential barrier of the same kind between the word and the actual

---

4Cf. (Richter, 1997, pp. 68f) for the status feature.
barrier.

(38) Licensing-Principle (LIP):
For each barrier object on the coll list of a sign $x$ and for each phrase $z$:
- the local value of $z$ is identical with the loc-lic value,
  iff $z$ dominates $x$, $z$ can be identified as the barrier specified\(^5\) and
  $z$ dominates no sign $y$ which in turn dominates $x$ and forms an
  equivalent barrier.

Hence, a word for which a barrier is defined cannot occur elsewhere; its
distribution is already specified in the lexical entry.

This concludes the description of technical requirements for our approach
to idioms. Note that we have defined a very small number of
new sorts and attributes to be included in the signature. All idiosyn-
 cratic information comes from the lexicon, as we will see in the next
section.

9.3.1 Decomposable Idioms
Let us show how a decomposable idiom can be analysed with our proposal. Take for instance the idiom in (33) *make waves*\(^6\). We can assign
the meanings “cause” and “trouble” to *make* and *waves* and assume
that there are two lexical entries for the idiomatic usage of these words.\(^7\)
The meaning of the whole idiom can be calculated in a regular composi-
tional way.

The idiomatic *make* subcategorizes for a plural noun with the word
form *wave* (the idiomatic version) creating a VP with the meaning
“cause trouble”:

\[
\begin{align*}
\text{cat} & \quad \text{head} \quad \text{listeme make}_3 \\
\text{subcat} & \quad \text{np} \\
\text{loc} & \quad \text{cat} \quad \text{head} \quad \text{noun listeme waves}_2 \\
\text{cont index num plural} & \quad 2 \\
\end{align*}
\]

*waves* for its part bears a non-empty coll list which looks as follows:

\[
\begin{align*}
\text{coll} & \quad \text{vp} \\
\text{loc lic} & \quad \text{cat} \quad \text{head} \quad \text{verb listeme make}_3 \\
\text{subcat (np)} & \quad 3 \\
\end{align*}
\]

The distribution of the idiomatic noun *waves* is restricted in that it
must be the complement of idiomatic *make*. The LIP makes sure that
the specified vp on the coll list is identical to the actual VP containing

\(^5\)as in “Italian film makes waves” from
http://news.bbc.co.uk/1/hi/entertainment/film/3171907.stm (All weblinks were
found by Google on 01-27-2004)

\(^7\)Another meaning of the idiom is “call attention” or “attract interest”. 
make and waves. Defining the barrier as a VP correctly implies that passivization of this idiom is not possible.\footnote{Riehemann found 5 examples out of 243 (2\%) where the idiom parts do not occur within the same VP. If one wants to account for those (including passivization and a relative clause) the barrier is simply to be set accordingly.}

Our example spill the beans\footnote{as in “Tom Cruise has spilled the beans on Nicole Kidman’s relationship with US musician Lenny Kravitz.” from http://www.smh.com.au/articles/2003/11/29/1070081589377.html?from=storyrhs} can be analysed analogously. As we assume regular syntactic composition to be in force, we predict that different specifiers (some beans) or modifications (as some very compromising beans) are grammatical.

A special case of the idiom not occurring in its canonical form is that of pronominal reference. In fact, pronominalization is quite hard to handle in idiom analysis. Cf. the following example:

(39) Eventually she spilled all the beans. But it took her a few days to spill them all.\footnote{Riehemann (2001), p. 207}

Here the pronoun them refers back to the idiomatic beans. As described in section 2 a pronoun has the same \textsc{listeme} value as its antecedent, so them gets its correct meaning. This being the case, the subcategorization requirements of idiomatic spill in both clauses are satisfied. The antecedent of them in turn is licensed by its own \textsc{coll} value stating that the idiomatic beans can only occur together with the verb spill in its idiomatic use. The barrier is a \textit{complete-clause} which allows e.g. passive or relative constructions. Thus, our proposal can handle pronominalization data, too.

\subsection*{9.3.2 Non-decomposable Idioms}

For idioms that have a non-decomposable meaning we define phrasal lexical entries (PLE), according to Sailer (2003) and following the idea of Gazdar et al. (1985). PLEs are lexical entries for syntactically complex expressions. Thus, they have properties of both words and phrases. As words, they are licensed by their lexical entry. As phrases, lexical rules cannot apply to them and syntactic operations like topicalization can be excluded by defining structural requirements in their \textsc{dtrs} attribute.

The semantics of a non-decomposable idiom is defined in its PLE. The parts of such an idiom are licensed by their ordinary lexical entries. In the syntactic structure of a sentence containing such an idiom there is a node where all necessary idiom parts are present. This node is either licensed by regular compositional principles or by a PLE. If a
PLE is applied, it replaces the semantics computed so far with its own meaning.

According to standard HPSG assumptions we adopt Immediate Dominance Schemas that license ordinary phrasal signs. In order to exclude the application of ID-Schemas to a phrase licensed by a PLE we can redefine the ID-PRINCIPLE in the following way:

\[
\text{[phrase}\ 
\text{coll}\ e\text{-list]} \rightarrow (\text{Head-Complement-Schema} \lor \text{Head-Adjunct-Schema} \lor \\
\text{Head-Marker-Schema} \lor \text{Head-Filler-Schema})
\]

Accordingly, we have to change all principles of grammar that are concerned with regular combination of signs (like the HEAD-FEATURE-PRINCIPLE or the SEMANTICS-PRINCIPLE) in such a way that they only apply to phrases bearing an empty COLL list. This can simply be done by adding a line in the antecedent (remember that the principles consist of an implication) stating [\text{coll}\ e\text{-list}].

In order to specify which lexical entries must have an empty COLL list, we introduce subsorts of listeme, namely the sorts \text{coll-listeme} and \text{no_coll-listeme}, and make the following constraint:

\[
\text{sign}
\rightarrow
\text{[collelist]}
\]

Note that all lexical entries have different values of LISTEME and, conversely, the set of all LISTEME values covers the entirety of lexical entries.

We have now made a distinction between regular phrasal signs which have an empty COLL list and non-regular or idiomatic phrases having a non-empty COLL list.\textsuperscript{11} Thus, in a PLE of an idiom like (35) saw logs\textsuperscript{12} we define its COLL list as non-empty. Besides, this idiom cannot be passivized without losing its idiomatic reading. Passivization is already excluded by the nature of the PLE itself: an object in accusative case is required and thus, logs can never occur as the subject.

\textsuperscript{11}The distribution of COLL values could be easily constrained by another principle which we omit here for reasons of space.

\textsuperscript{12}as in “Two young boys stand by their mother’s bed while she saws logs in her sleep.” from http://www.collegestories.com/filmfrat/igby-goes-down.html
In defining a non-empty COLL value, we provide a unified way to treat decomposable and non-decomposable idioms, marking their quality of being idiomatic. Parts of decomposable idioms bear a non-empty COLL list, which restricts their occurrence to certain contexts. Nondecomposable idioms also have a non-empty COLL list, exempting them from regular syntactic and semantic principles.

In addition, the occurrence of nondecomposable idioms can be restricted to certain contexts via the same feature. This is important for idiomatic intensifiers, among others, like as a sandboy in to be happy as a sandboy or as a kite in to be high as a kite.

9.4 How does it fit in?
As we have introduced a (rather manageable) number of new sorts, principles and features, and explained how to adapt existing grammar principles to our approach, it is not difficult to imagine how our proposal fits into the overall HPSG Grammar. Non-idiomatic signs bear an empty COLL list and thus are exempt from any consequence of our analysis. In addition, the fact that each word has now a value of its LISTEME feature is only important where an access to LISTEME is needed. At the same time our proposal can be combined smoothly with other modules of grammar concerning idioms.

In order not to settle for this mere claim, we will illustrate it by means of an idiom-independent part of grammar – topicalization and verb movement, for instance.

In a declarative sentence in German the finite verb occupies the second position. Such sentences can be derived from verb-last-clauses by movement of the verb. The position in front of the verb is called
Vorfeld and can contain an argument or an adjunct. The positioning of an element in the Vorfeld is usually analyzed as a nonlocal dependency, which accounts for the vast majority of declarative clauses. A verb movement analysis by Borsley and Kiss, which is discussed in Müller (2005), can account for this phenomenon.

First, a special lexical entry for a verbal trace (40) is introduced, where verbal movement is treated as a local phenomenon. Thus, for a verbal trace there is no application of the standard unbounded dependency analysis (cf. Pollard and Sag (1994)). The local-valued feature dsl allows the valence structure of the moved verb to be available along the head projection line. The lexical entry of the verbal trace (slightly adapted from Müller, 2005, p. 22) looks as follows:

\[
\begin{align*}
\text{(40)} & \quad \left[ \begin{array}{c}
\text{word} \\
\text{ Phon } \\
\text{ ss Loc } \\
\text{ Cat } \\
\text{ subcat } \\
\text{ cont }
\end{array} \right] \\
& \quad \left[ \begin{array}{c}
\text{Hea d} \\
\text{ listeme} \\
\text{ dsl} \\
\text{ subcat} \\
\text{ cont}
\end{array} \right]
\end{align*}
\]

Secondly, Müller outlines a lexical rule (41) for a special version of the finite verb which is moved. The verb being licensed by this rule takes the projection of the verbal trace as its argument. A phrase is only grammatical if the valence properties of the verb are identical to those of the verbal trace. The rule takes a verb in non-initial position as input and outputs a verb in initial position (marked by the verb feature initial). The lexical rule for initial verb position (V1-LR, slightly adapted from Müller, 2005, p. 23) is defined in the following way:

\[
\begin{align*}
\text{(41)} & \quad \left[ \begin{array}{c}
\text{word} \\
\text{ Phon } \\
\text{ ss Loc } \\
\text{ Cat } \\
\text{ subcat } \\
\text{ loc cat}
\end{array} \right] \\
& \quad \left[ \begin{array}{c}
\text{Hea d} \\
\text{ listeme} \\
\text{ dsl}
\end{array} \right]
\end{align*}\]

To illustrate this analysis together with our approach, we take the German idiom jemandem den Garaus machen (to cook so.‘s goose, ‘to kill someone’). We can roughly state its decomposable meaning as “to
put an end to”.

(42) Zucker, Salz, chemische Aromen und Geschmacksverstärker
    make the natural taste the end
    ‘Sugar, salt, chemical flavorings and flavor enhancers destroy our
    natural taste.’

(43) Den Garaus macht den Seglern die Langleinenfischerei.
    ‘Long line fishing encroaches upon sailors.’

The sentence in (42) shows that the NP can appear detached from
the verb. In (43) the NP is fronted and has been extracted from the
original VP. The structure of (43) is the following:

(44) [Den Garaus]i machtj den Seglern die Langleinenfischerei

We can analyse this sentence by means of the standard unbounded
dependency analysis, the approach in Müller (2005) and our proposal.
Figure 2 shows the syntax tree of (44) including all relevant information.

Let us begin to explain this figure with the extraction trace. The un-
bounded dependency analysis involves a lexical entry for a trace where
the local value (1) of the extracted element is identical to an element
in the inner slash list. The Nonlocal-Feature-Principle guarantees that this value “percolates” up the tree. By the Head-Filler-
Schema this slash value is bound via TO-BIND SLASH because the
LOCAL value of the filler is the same as the SLASH value.

Further, the verbal trace attracts the arguments of the moved verb
and puts it onto both its SUBCAT list and its DSL CAT SUBCAT list. The Head-Feature-Principle enforces the presence of the DSL value
along the head projection line. The verb machen for its part is input to
the V1-LR (41). The output subcategorizes for a sign (6) whose
DSL value is identical to the LOCAL value of the input (7).

This concludes the pure fronting analysis. So far as our proposal is
concerned, the important parts are the values of COLL and LISTEME respectively.

The filler den Garaus has a COLL value defining a barrier15 bearing

13http://www.erichlutz.de/publikationen/globus.html
14Salzburger Nachrichten, 15.11.2000; Found with COSMAS II by IDS Mannheim
15We do not specify the barrier to be a VP deliberately. If den Garaus occurs in
   situs, a vp would be correct, but in this case we have a complete-clause (if not even
   an utterance) as barrier. In order to not exclude such fronting cases, the barrier has
   thus, on the one hand, to be general enough. On the other hand, we can exclude
   ungrammatical cases of extraction beyond clausal boundaries.
Figure 2: Analysis of Den Garaus macht den Seglern die Langleinenfischerei.
the listeme value machen3. In the structure we have a node (S), which is the minimal barrier above the filler and whose head is exactly of the required listeme value. Thus, our Licensing-Principle is satisfied.

By the sort machen3 we refer to an idiomatic verb machen with the meaning “to put”. This verb subcategorizes for an NP (with SYNSEM value 5) bearing the listeme value garaus (not depicted in figure 2). As a LOCAL value includes LISTEME, and the LOCAL values of the filler and the gap are identical, the subcategorization requirements of machen3 are satisfied, even though Garaus has been extracted. Note that by the identity of the LOCAL value 6 and the DSL value of 6, the verbal trace has the correct listeme value.

9.5 Alternative Analyses

9.5.1 A Different coll Mechanism

The analysis we suggest here is an enhancement of a proposal by Richter and Sailer (1999). However, in Sailer (2003) the author described a variant of the coll mechanism: In this thesis, the value of coll is a singleton list that may contain a sign. That sign is the overall expression in which the idiomatic word occurs. Take for example the idiom spill the beans: in the lexical entry of the idiomatic word beans its coll value is specified as a sign containing the semantic contributions of a definite article, the idiomatic word spill and beans itself in the right scopal relations. Sailer defines the so-called Coll-Principle ensuring that the sign specified in a coll list dominates the sign bearing that list. As a consequence, information of the overall utterance is available at lexical level and, conversely, local information is available on each node in the structure.

Thus, even though Sailer introduces only one new attribute, this approach is very unrestrictive and if one taps its full potential, nearly all grammatical phenomena can be described, even if they have nothing to do with collocations. Selection, e. g., would only be a special case of a collocation. Because of this power and unrestriction, that version of coll is to be met with criticism.

9.5.2 A Constructional Approach

Riehemann (2001) makes another concrete proposal for the analysis of idioms. She adopts many ideas of Construction Grammar and carries them forward to the HPSG framework. Her approach requires a complex machinery of new sorts and attributes to cover not only the amount of existing idioms but also their occurrences in different syntactic configurations. She has to assume, e. g., distinct subsorts of a spill\_beans\_idiom\_phrase for the idiom occurring in a
head-subject-phrase (Dana spilled the beans.), in a head-filler-structure (Who did Kim claim spilled the beans?) or in a head-specifier-structure (the beans that Dana spilled). Even if the existence of sorts for different constructions is well established in Construction Grammar, it is questionable to assume different subclasses of linguistic signs, only because they contain idiomatic items in different syntactic structures. In other words, why assume different sorts for one single idiom only because it occurs in different constructions?

Moreover, Riehemann herself has to admit that her approach cannot handle cases of pronominal reference like (39), because idiomatic spill is not licensed as it seems to appear by itself and not within a spill Beans_idiom_phrase.

In summary, it seems to us that a lexical approach is to be preferred over a structural one. Nevertheless, her arguments in favor of a constructional analysis of non-decomposable idioms are convincing. Our counterpart to that are phrasal lexical entries which we assume for this kind of idiomatic expressions.

9.6 A Modular Approach: Prospects

We have proposed one way of analyzing idioms and similar lexical idiosyncrasies. It can handle distributional characteristics of idiomatic words and even difficult cases like pronominalization. We have demonstrated by means of topicalization and verb movement that our proposal merges smoothly with other modules of grammar.

We decided to take a word-level collocation-based account using the COLL feature. This approach is modular in two ways. Firstly, the barriers can be adjusted "vertically" according to the range (XP, complete clause or utterance) needed for a particular idiomatic expression. Secondly, by the LOC-LIC feature we can specify any characteristics within the local information. We could now go on and define other attributes of barrier like PHON-LIC to define any requirements of the phonetic string of that barrier. In that way our approach is also horizontally modular.

An application of such a PHON-LIC feature would be the modelling of occurrence restrictions of the English indefinite article an. This phenomenon together with other cases of sandhi is discussed by Asudeh and Klein (2002). With our approach, we define the lexical entry of an as follows: the PHON-LIC value of the barrier np on the COLL list is the phonetic string ⟨an⟩ + a phonetically realized vowel.

Thus, with a quite general approach to idioms using the COLL feature, we can handle very particular phenomena, too. We leave it to further research to explore the possibilities that our approach offers.
References


