License to Coll: How to Bind Bound Words and Readings to their Contexts

Jan-Philipp Soehn

University of Jena

Abstract

In this contribution we propose a new module for handling idioms and distributional idiosyncrasies. Based on the concept by Richter/Sailer (1999) the new feature COLL (context of lexical licensing) plays the central role in our approach. We provide a way to handle decomposable and nondecomposable idioms and idioms containing bound words. Our module guarantees the co-occurrence of all idiom parts and of bound word and licensing context, respectively.

A prerequisite for our analysis is a means to select for particular elements in the lexicon. We introduce another feature, LISTEME, which gives each lexical item its unique identifier and makes it possible to select for a particular lexical word or phrase.

Finally, we compare our proposal with alternative approaches and give some ideas regarding further applications beyond idiomaticity.

1   Motivation

Idioms are omnipresent in everyday language, enriching our communication with metaphoric imagery and fulfilling various communicative goals.

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[1] and idioms containing bound words. We concentrate on technical aspects of the analysis and refrain from presenting detailed linguistic corpus data due to space limitations. 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.*

(1) *make waves* ("cause trouble")

(2) *spill the beans* ("divulge a secret")

The expressions in (1) and (2) 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 (1), 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 (3) and (4):

---

[1]Cf. Gibbs et al. (1989) or Nunberg et al. (1994) for this distinction.

(3) *saw logs* ("snore")

(4) *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".

Finally, we want to draw the attention to idioms comprising bound words or "cranberry words" (Aronoff, 1976). These are expressions which are highly collocationally restricted. Dobrovol'skij (1988) compiled quite a lot of examples for German, Dutch and English.

(5) *to learn/do sth. by <u>rote</u>* (automatically, by heart)

(6) *to cock a <u>snook</u>* (to thumb the nose)

The underlined words are restricted to the given contexts. Sometimes there is some variation, as in *to lie/go/lay <u>doggo</u>* (Brit. slang; "to hide oneself"), but a free distribution is not possible. Such idioms can be either decomposable or non-decomposable.

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.

A mechanism for selection of lexical elements has to meet three requirements:

1. The information has to be locally available (below SYNSEM).

2. The information has to be available along the syntactic projection line.

3. The information must be identical for a pronoun and its antecendent.

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): the locality principle for selection (Pollard and Sag, 1994, p. 23) was not implemented and there was not means for the LEXEME value to percolate. We therefore propose that the LEXEME approach has to be discarded.

A different concept that helps 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 morphosyntactic 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 (7).

(7) 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 Sciullo and Williams, 1988, p. 1) 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.

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 to restrict the range of influence of the COLL feature.

barrier objects have an attribute LOCAL-LICENSER (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: *complete-clause*, *utterance*, *np*, *vp* and *pp*. The subsorts of barrier correspond to nodes in the syntactic tree with particular properties. The relations depicted in figure 2 identify the nodes which relate to the subsorts of *barrier*.[4]

$$\begin{bmatrix} barrier \\ \text{LOC-LIC } local \end{bmatrix}$$

complete-clause   utterance   xp
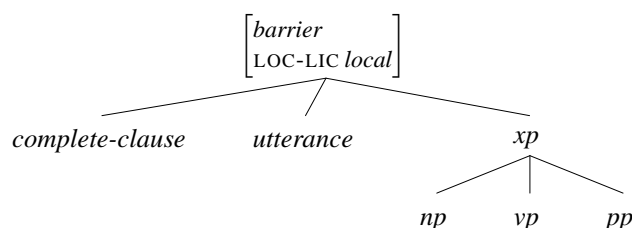
                                    np   vp   pp

Figure 1: Sort hierarchy for *barrier*

The LICENSING-PRINCIPLE (informally in 8) 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 barrier.

(8) 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 and z dominates
   no sign y which in turn dominates x and forms an equivalent barrier.

In this principle there are three conditions to be satisfied. The first one is simply that there must be a domination relation between the phrase z and the idiomatic element x. The second condition, that a particular barrier can be identified, means

---

[4]Cf. (Richter, 1997, pp. 68f) for the STATUS feature.

$$\forall \boxed{1} \left(\text{is\_utterance}(\boxed{1}) \leftrightarrow \boxed{1} \begin{bmatrix} \textit{unembedded-phrase} \\ \text{SS} \begin{bmatrix} \text{STATUS } \textit{complete} \\ \text{LOC CAT} \begin{bmatrix} \text{HEAD } \textit{verb} \\ \text{SUBCAT } \textit{elist} \end{bmatrix} \end{bmatrix} \\ \text{ILLOCUTION } \textit{illocution} \end{bmatrix} \right)$$

$$\forall \boxed{1} \left(\text{is\_complete-clause}(\boxed{1}) \leftrightarrow \boxed{1} \begin{bmatrix} \textit{phrase} \\ \text{SS} \begin{bmatrix} \text{STATUS } \textit{complete} \\ \text{LOC CAT} \begin{bmatrix} \text{HEAD } & \textit{verb} \\ \text{SUBCAT } & \textit{elist} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right)$$

$$\forall \boxed{1} \left(\text{is\_vp}(\boxed{1}) \leftrightarrow \boxed{1} \begin{bmatrix} \textit{phrase} \\ \text{SS} \begin{bmatrix} \text{STATUS } \textit{incomplete} \\ \text{LOC CAT} \begin{bmatrix} \text{HEAD } & \textit{verb} \\ \text{SUBCAT } & \textit{nelist} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right)$$

$$\forall \boxed{1} \left(\text{is\_np}(\boxed{1}) \leftrightarrow \boxed{1} \begin{bmatrix} \textit{embedded-phrase} \\ \text{SS} \begin{bmatrix} \text{STATUS } \textit{incomplete} \\ \text{LOC CAT} \begin{bmatrix} \text{HEAD } & \textit{noun} \\ \text{SUBCAT } & \textit{elist} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right)$$

$$\forall \boxed{1} \left(\text{is\_pp}(\boxed{1}) \leftrightarrow \boxed{1} \begin{bmatrix} \textit{embedded-phrase} \\ \text{SS} \begin{bmatrix} \text{STATUS } \textit{incomplete} \\ \text{LOC CAT} \begin{bmatrix} \text{HEAD } & \textit{prep} \\ \text{SUBCAT } & \textit{elist} \end{bmatrix} \end{bmatrix} \end{bmatrix} \right)$$

Figure 2: Relations for *barrier*-subsorts

that one of the relations `is_vp`, `is_complete-clause`, etc. must hold. By the third condition we exclude the case where there is another phrase y between z and x being also a possible barrier. Thus, z is always the minimal one.

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 quite a small number of new sorts, relations and attributes to be included in the signature. All idiosyncratic information comes from the lexicon, as we will see in the next section.

3.1 Decomposable Idioms

Let us show how a decomposable idiom can be analysed with our proposal. Take for instance the idiom in (1) *make waves*[5]. 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.[6]

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

$$(9) \quad \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{LISTEME } make_3 \end{bmatrix} \\ \text{SUBCAT} \left\langle \text{NP}, \begin{bmatrix} \text{LOC} \begin{bmatrix} \text{CAT HEAD} \begin{bmatrix} noun \\ \text{LISTEME } wave_2 \end{bmatrix} \\ \text{CONT INDEX NUM } plural \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

$wave_2$ for its part bears a non-empty COLL list which looks as follows:

$$(10) \quad \begin{bmatrix} \text{COLL} \left\langle \begin{bmatrix} vp \\ \text{LOC-LIC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} verb \\ \text{LISTEME } make_3 \end{bmatrix} \\ \text{SUBCAT} \left\langle \text{NP} \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

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 *make* and *waves*. That would have the following semantics: $\lambda x.[\mathsf{waves}''(y)](\mathsf{make}''(x,y))$[7]. Defining the barrier as a VP correctly implies that passivization of this idiom is not possible.[8]

---

[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)

[6]Other paraphrases of the idiom are "call attention" or "attract interest". We leave open the exact definition of the meaning and take "cause trouble" as example.

[7]In this contribution we do not go into details of semantics. Under CONTENT LF we give the logical form of the expression, using a double apostrophe to indicate an idiomatic meaning. Our approach is compatible with any semantic representation as e. g. MRS (cf. Copestake et al., 1998) or LRS (Richter and Sailer, 2003).

[8]Riehemann (ibid.) 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*[9] 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:

(11) *Eventually she spilled all the beans. But it took her a few days to spill them all.*[10]

Here the pronoun *them* refers back to the idiomatic *beans*. As described in section 2 a pronoun has the same LISTEME value as its antecendent, 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 COLL value stating that the idiomatic *beans* can only occur together with the verb *spill* in its idiomatic use. The barrier is a *complete-clause* which allows e. g. passive or relative constructions. Thus, our proposal can handle pronominalization data, too.

3.2 Bound Words

Now we come to bound words: The idiom in (5) *to learn sth. by rote*[11] contains a word that never occurs in other contexts than as a complement of a PP with head *by*. The idiom is decomposable: *rote* means something like "routine". The relevant parts of its lexical entry can thus be stated as follows:

$$
(12) \begin{bmatrix} word \\ \text{PHON} \langle rote \rangle \\ \text{SS LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} noun \\ \text{LISTEME } rote \end{bmatrix} \\ \text{SUBCAT } e\text{-}list \end{bmatrix} \\ \text{CONT LF } \lambda x \lambda Q. \exists x (\text{rote}'(x) \wedge Q(x)) \end{bmatrix} \\ \text{COLL} \left\langle \begin{bmatrix} pp \\ \text{LOC-LIC} \begin{bmatrix} \text{CAT HEAD LISTEME } by \\ \text{CONT LF } ...\exists x(\text{rote}'(x) \wedge \text{by}'(x,e))... \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}
$$

By defining the CONTENT value of the barrier *pp* we prevent a modification of *rote*, which would be ungrammatical. The PP can modify any verb, allowing the occurrence of *(know, learn, sing, do,...) sth. by rote*.

To account for the example in (6), the lexical entry of *snook* requires a *vp* barrier with an appropriate LISTEME value of the head, as seen for the idiom *make waves*. We can restrict the distribution of these bound words in the same way as we handle idiomatic words contained in a decomposable idiom.

---

[9]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

[10](Riehemann, 2001, p. 207)

[11]as in "Students forced to learn history by rote" from
http://www.shanland.org/Political/News_2002/students_forced_to_learn_history.htm

3.3 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 DTRS attribute. 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:

$$(13) \quad \begin{bmatrix} phrase \\ \text{COLL } e\text{-}list \end{bmatrix} \rightarrow \begin{pmatrix} \text{HEAD-COMPLEMENT-SCHEMA} \lor \\ \text{HEAD-ADJUNCT-SCHEMA} \lor \\ \text{HEAD-MARKER-SCHEMA} \lor \\ \text{HEAD-FILLER-SCHEMA} \end{pmatrix}$$

Accordingly, we have to change all other principles of grammar that are concerned with regular combination of signs 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 all principles are formulated as implications) stating [COLL *e-list*].

In order to specify which lexical entries must have an empty COLL list, we introduce subsorts of *listeme*, namely *coll_listeme* and *no_coll_listeme*, and make the following constraint:

$$(14) \quad \begin{bmatrix} sign \\ \text{SS LOC CAT HEAD LISTEME } no\_coll\_listeme \end{bmatrix} \rightarrow \begin{bmatrix} \text{COLL } elist \end{bmatrix}$$

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. Thus, in a PLE of an idiom like (3) *saw logs*[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.

---

[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

$$(15) \quad \begin{bmatrix} \textit{phrase} \\ \text{PHON} \quad \boxed{3} \oplus \boxed{4} \\[4pt] \text{SS LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \textit{verb} \\ \text{LISTEME } \textit{saw-logs} \end{bmatrix} \\ \text{SUBCAT} \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \text{CONT LF } \lambda x(\mathsf{snore}'(x)) \end{bmatrix} \\[4pt] \text{DTRS} \begin{bmatrix} \textit{head-comp-struc} \\ \text{H-DTR} \begin{bmatrix} \textit{word} \\ \text{PHON } \boxed{3} \left\langle \textit{saw} \right\rangle \\ \text{SS LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \textit{verb} \\ \text{LISTEME } \textit{saw} \end{bmatrix} \\ \text{SUBCAT} \left\langle \boxed{2} \text{ NP, } \boxed{5} \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{N-DTR} \begin{bmatrix} \text{PHON } \boxed{4} \left\langle \textit{logs} \right\rangle \\ \text{SS } \boxed{5} \text{ LOC} \begin{bmatrix} \text{CAT} \begin{bmatrix} \text{HEAD} \begin{bmatrix} \text{CASE} \quad \textit{acc} \\ \text{LISTEME } \textit{log} \end{bmatrix} \\ \text{SUBCAT} \left\langle \right\rangle \end{bmatrix} \\ \text{CONT INDEX NUM } \textit{plural} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\[4pt] \text{COLL} \quad \textit{ne-list} \end{bmatrix}$$

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*.

4 Alternative Analyses

4.1 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 phenom-

ena 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 unrestrictedness, that version of COLL is to be met with criticism.

4.2   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 different constructions (e. g. a *head-subject-phrase* or a *head-filler-structure*). Even if the existence of sorts for different constructions themselves 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 (11), because idiomatic *spill* is not licensed as it seems to appear by itself and not within a *spill_beans_idiom_phrase*. In addition, Riehemann is unable to account for bound words, as she cannot constrain their distribution once she assumes lexical entries for them.

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.

5   Prospects for a Modular Approach

We have proposed one way of analyzing idioms and similar phenomena of distributional idiosyncrasies. It can handle distributional characteristics of idiomatic words and even difficult cases like pronominalization.

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 as one can specify objects that are on different levels in the sign-hierarchy (*sign*, *phonstring*, *local* etc.).

An application of such a PHON-LIC feature would be the modelling of occurrence restrictions of the English indefinite article *an*. This phenomenon is discussed by Asudeh and Klein (2002) together with other cases of sandhi. The authors integrate phonological shape conditions of the context in the element's lexical entry. Instead of their new feature PHONOLOGICAL-CONTEXT, we can use our COLL approach and define the lexical entry of an as follows:

$$(16) \quad \left[\text{COLL} \left\langle \begin{bmatrix} np \\ \text{PHON-LIC} \left\langle \text{ən} \right\rangle \oplus (\text{æ} \vee \Lambda \vee \text{e} \vee \text{ə} \vee ...) \end{bmatrix} \right\rangle \right]$$

The PHON-LIC value of the barrier *np* on the COLL list is the phonetic string *an* plus a phonetically realized vowel.

Thus, with a quite general approach to idioms using the COLL feature, we can handle very particular phenomena, too. The COLL module is described more extensively in Soehn (In prep.).

To explore the possiblities that our approach holds may be a matter of further research.

References

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge MA, third edition, linguistic Inquiry Monographs.

Asudeh, Ash and Klein, Ewan. 2002. Shape Conditions and Phonological Context. In Frank van Eynde, Lars Hellan and Dorothee Beermann (eds.), *Proceedings of the 8th International HPSG Conference*, pages 20–30, CSLI Publications, http://cslipublications.stanford.edu/HPSG/2/.

Copestake, Ann, Flickinger, Daniel, Pollard, Carl and Sag, Ivan A. 1998. *Minimal Recursion Semantics: An Introduction*. Stanford, ftp://ftp-csli.stanford.edu/linguistics/sag/mrs.ps.gz.

Di Sciullo, Anna-Maria and Williams, Edwin. 1988. *On the Definition of Word*. Linguistic Inquiry Monographs, MIT Press, Cambridge, Mass, second edition.

Dobrovol'skij, Dmitrij. 1988. *Phraseologie als Objekt der Universalienlinguistik*. VEB Verlag Enzyklopädie Leipzig, linguistische Studien.

Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey and Sag, Ivan. 1985. *Generalized Phrase Structure Grammar*. Cambridge, Mass.: Harvard University Press.

Heinz, W. and Matiasek, J. 1994. Argument Structure and Case Assignment in German. In John Nerbonne, Klaus Netter and Carl Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*, pages 199–236, CSLI Publications, lecture Notes 46.

Krenn, Brigitte and Erbach, Gregor. 1994. Idioms and Support Verb Constructions. In John Nerbonne, Klaus Netter and Carl Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*, pages 365–396, CSLI Publications, lecture Notes 46.

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

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Standford University: CSLI/The University of Chicago Press.

Raymond W. Gibbs, JR., Nayak, Nandini P., Bolton, John L. and Keppel, Melissa E. 1989. Speakers' assumptions about the lexical flexibility of idioms. *Memory & Cognition* 17(1), 58–68.

Richter, Frank. 1997. Die Satzstruktur des Deutschen und die Behandlung langer Abhängigkeiten in einer Linearisierungsgrammatik. Formale Grundlagen und Implementierung in einem HPSG-Fragment. In Erhard Hinrichs, Detmar Meurers, Frank Richter, Manfred Sailer and Heike Winhart (eds.), *Ein HPSG-Fragment des Deutschen, Teil 1: Theorie*, Arbeitspapiere des SFB 340, No. 95, pages 13–187, Universität Tübingen.

Richter, Frank and Sailer, Manfred. 1999. LF Conditions on Expressions of Ty2: An HPSG Analysis of Negative Concord in Polish. In Robert D. Borsley and Adam Przepiórkowski (eds.), *Slavic in Head-Driven Phrase Structure Grammar*, pages 247–282, Stanford: CSLI Publications.

Richter, Frank and Sailer, Manfred. 2003. Basic Concepts of Lexical Resource Semantics., lecture notes for the course on Constraint-based Combinatorial Semantics at ESSLLI 2003, Vienna. To appear in the series of the Kurt Gödel Society, Vienna.

Riehemann, Susanne Z. 2001. *A Constructional Approach to Idioms and Word Formation*. PhD Thesis, Stanford University, Stanford, CA.

Sailer, Manfred. 2003. Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar. Phil. Dissertation (2000). Arbeitspapiere des SFB 340. 161, Eberhard-Karls-Universität Tübingen.

Soehn, Jan-Philipp (In prep.). *"Über Bärendienste und erstaunte Bauklötze". Idiome ohne freie Lesart in der HPSG*. PhD Thesis, Friedrich-Schiller-Universität Jena.

Soehn, Jan-Philipp and Sailer, Manfred. 2003. At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In Gerhard Jäger, Paola Monachesi, Gerald Penn and Shuly Wintner (eds.), *Proceedings of Formal Grammar 2003*, pages 149–161.