Gluing idioms back together: A phraseo-combinatorial analysis

Manfred Sailer
Goethe-University Frankfurt a.M.

Sascha Bargmann
Goethe-University Frankfurt a.M.

Proceedings of the 28th International Conference on
Head-Driven Phrase Structure Grammar

Online (Frankfurt/Main)

Stefan Müller, Nurit Melnik (Editors)

2021

Frankfurt/Main: University Library

pages 186–204

Keywords: decomposability, idiom, lexical rule, syntactic flexibility, phrasal lexical entry

Sailer, Manfred & Sascha Bargmann. 2021. Gluing idioms back together: A phraseo-combinatorial analysis. In Stefan Müller & Nurit Melnik (eds.), *Proceedings of the 28th International Conference on Head-Driven Phrase Structure Grammar, Online (Frankfurt/Main)*, 186–204. Frankfurt/Main: University Library. DOI: 10.21248/hpsg.2021.10.

Abstract

The formal analysis of idioms has been oscillating between approaches that emphasize the unit-like character of idioms and approaches that focus on the autonomy of the idioms' parts. In this paper, we summarize the main arguments for and against these two positions to then propose an account that tries to capture and combine the insights and advantages of both types of analysis. The resulting theory is heavily influenced by the approach taken in Riehemann (2001).

1 Introduction

Idioms like *kick the bucket* 'die' in (1) or *pull strings* 'use connections' in (2) have mostly been analyzed as either fixed phrases that are coupled with the idiom's meaning as a whole (henceforth *phrasal accounts*) or as two or more separate idiomatic parts that combine according to the conventional rules of combinatorics and that each contribute their own meaning to the meaning of the idiom as a whole (henceforth *combinatorial accounts*).

(1) Our gold fish *kicked the bucket* last night.

(2) My boss *pulled strings* to get his current job.

Whereas phrasal accounts emphasize the unit-like character of idioms, combinatorial accounts focus on the (relative) autonomy of the idioms' parts. In this paper, we summarize the main arguments for and against these two positions to then propose an analysis that tries to capture and combine the insights and advantages of both types of analysis.

2 Phrasal versus combinatorial approaches

Early generative approaches, like Chomsky (1965), consider all idioms lexical units with internal structure; idioms are taken to be part of the lexicon but more complex than single words. Most subsequent approaches to idioms take a (much) more differentiated position.

On the basis of their empirical observation that not all idioms behave like monolithic units, but many of them actually show a certain degree of syntactic flexibility, Wasow et al. (1983) and Nunberg et al. (1994) argue that idioms come in at least two versions: (i) idiomatic phrases (IPs) and (ii) idiomatically combining expressions (ICEs).

IPs are semantically non-decomposable idioms that are analyzed as fixed phrases stored in the lexicon in the form of one single monolithic entry, which is

[†]We thank the reviewers and the participants of HPSG 2021 for their comments, in particular Emily Bender, Jamie Findlay, Paul Kay, Nurit Melnik, and Adam Przepiórkowski. We are grateful to Pascal Hohmann for help with LaTeX. All errors are ours.

directly coupled with the idiomatic meaning. A typical and often used example of an IP is the idiom *kick the bucket*, whose interpretation is 'die'.

ICEs, on the other hand, are semantically decomposable idioms that are analyzed as consisting of two or more separate word-level lexical entries that each contribute only a part of the idiom and of its meaning. Typical and often used examples of ICEs are *pull strings* 'use connections' (where *pull* is interpreted as 'use' and *strings* as 'connections') and *spill beans* 'divulge information' (where *spill* is interpreted as 'divulge' and *beans* as 'information').

The basic insight in Wasow et al. (1983) and Nunberg et al. (1994) is that semantic decomposability correlates with syntactic flexibility: Semantically decomposable idioms can undergo syntactic processes such as passivization, topicalization, or the insertion of adjuncts, whereas semantically non-decomposable idioms cannot – see (3), where "$" indicates the unavailability of an idiomatic reading.

(3) a. The beans were spilled by Pat. (Nunberg et al., 1994, 510)
b. $ The bucket was kicked by Pat. (Nunberg et al., 1994, 508)

Semantically decomposable idioms motivated a combinatorial analysis in GPSG (Gazdar et al., 1985). This has been carried over to HPSG in Krenn & Erbach (1994), Sailer (2003), and Soehn (2009).

Kay et al. (2015) and Bargmann & Sailer (2018) then point to empirical evidence against the syntactic fixedness of non-decomposable idioms. For example, passivization of non-decomposable idioms is not blocked in principle but interacts in a predictable way with the discourse constraints on passive in a given language, as in (4).

(4) When you are dead, you don't have to worry about death anymore. ... The bucket will be kicked. (Bargmann & Sailer, 2018, 21)

Bargmann & Sailer (2018) account for the passivizability of *kick the bucket* in (4) in the following way: First, the subject of an English passive clause must be given or inferable from the preceding context (Kuno & Takami, 2004, 127). Second, the idiomatic noun phrase *the bucket* refers to a dying event. Since dying is given in the context, passivization is possible.

In the light of such observations, Kay et al. (2015) and Bargmann & Sailer (2018) analyze all idioms with a regular syntactic structure – decomposable or not – in a combinatorial way. This leads to an ICE-style analysis for idioms such as *kick the bucket* and restricts a phrasal analysis to expressions such as *kingdom come* 'paradise'.

Findlay (2019) points out two big challenges for combinatorial analyses: (i) The idiomatic versions of the words need to be prevented from occurring independently of the idiom (henceforth *collocational challenge*), and (ii) there is a new lexical entry for each (idiomatic) word in each idiom (henceforth *lexical*

explosion challenge). Findlay (2019) then suggests a phrasal analysis of idioms in a tree-grammar-based version of LFG.

A hallmark of phrasal analyses is that an idiom's parts are licensed directly and exclusively through the idiom's phrasal entry. Compared to combinatorial analyses, this has three advantages: (i) An idiom's parts are automatically prevented from occurring independently of the idiom, which leads to a confinement of idiom parts and avoids the collocational challenge. (ii) There is no need for individual lexical entries for individual idiomatic words, which avoids a lexical proliferation or "explosion" and hence ensures a leaner lexicon. (iii) It captures the intuition that idioms are lexical units.

However, phrasal analyses do not seem to be the appropriate analytic tool for some syntactic and/or textual constellations. We will consider three relevant cases here. The first problem is posed by the occurrence of idiom parts in relative clauses, see (5).

(5) a. The strings$_k$ [$_{RC}$ that Pat pulled ___$_k$] got Chris the job. (Nunberg et al., 1994, 510)

b. John never pulled the strings$_k$ [$_{RC}$ that his mother told him should be pulled ___$_k$]. (Henk v. Riemsdijk's example)

If the relative clause contains only the idiomatic verb, here *pull*, it is unclear how a phrasal account can connect the verb with the idiomatic noun, here *strings*. Example (5b) is particularly challenging, as there is only one occurrence of the idiomatic noun *strings* but two occurrences of the idiomatic verb *pull*. According to the combinatorial analysis in Webelhuth et al. (2018, 257), *pull* is licensed in the relative clauses in (5) via the semantics of *strings*$_k$, which is present via the gap ___$_k$.

The second problematic constellation for phrasal accounts is exemplified in (6). Just as in (5b), there is one occurrence of the idiomatic noun *beans* that is related to two distinct occurrences of the idiomatic verb *spill*. It is not clear how this could be reconciled with a phrasal account, as the noun *beans* would have to be part of two distinct instances of the idiomatic phrase *spill beans* simultaneously.

(6) The beans [$_{VP}$ [$_{VP}$ have not been spilled yet], but [$_{VP}$ will be spilled very soon]].

In the combinatorial approach in Webelhuth et al. (2018), an occurrence of the idiomatic verb *spill* is licensed via the semantics of the idiomatic noun *beans*. Since *beans* is the head of the subject of the conjoined verb phrases, this single occurrence of the noun is sufficient to license two occurrences of the idiomatic verb *spill*.

The third challenge for phrasal accounts that we would like to discuss involves the pronominalization of idiom parts, as in (7).

(7) Eventually she spilled all the beans$_k$. But it took her a few days to spill them$_k$ all. (Riehemann, 2001, 207)

In the *but*-clause in (7), the idiomatic verb *spill* combines with the pronoun *them* rather than an overt realization of the noun *beans*. This is only possible if the antecedent of the pronoun is a noun phrase whose head is idiomatic *beans*. This condition is hard to integrate into a phrasal analysis.

A combinatorial approach does not necessarily face this problem. According to Webelhuth et al. (2018, 251–252, 256), a pronoun shares relevant parts of its semantics with its antecedent. In our case, this means that the pronoun *them* in the *but*-clause has the semantics of idiomatic *beans*. Consequently, idiomatic *spill* is licensed in the *but*-clause.

This brief discussion shows that phrasal accounts naturally capture the confinement of idiom parts and allow for a leaner lexicon. Both of these are rather conceptual arguments. A combinatorial account, on the other hand, seems to be the better fit when it comes to the actual phenomena, like idioms in relative clauses, parts of idioms occurring twice, and idiom parts being pronominalized.[1]

3 Riehemann's approach

In Riehemann (2001), all idioms are phrases that consist of two or more words of which at least one is an idiomatic word. Such an idiomatic word differs from its non-idiomatic counterpart in exactly two aspects: (i) It has a different meaning (figurative or empty), and (ii) it does not have an individual entry in the lexicon, as it is obligatorily part of the idiom it belongs to and, as a consequence, has no status of its own. Apart from these two differences, an idiomatic word is identical to its literal counterpart, i.e. the former shares the latter's phonology, morphology, and syntax.

In order to ensure this overlap between idiomatic words and their non-idiomatic counterparts, Riehemann (2001) establishes a relation between them. Using *asymmetric default unification* "$\underset{\sqcap}{<}$" (Lascarides & Copestake, 1999, 69), she has an idiomatic word adopt all the characteristics of its literal counterpart that are not specified within the idiomatic word. The only characteristic that she specifies in idiomatic words is their semantics. See Fig. 1 for a sketch of Riehemann's encoding of *spill beans*.

Riehemann (2001) keeps the words that occur in a phrase (any phrase, not just idiomatic ones) in an unordered repository that she tellingly calls WORDS.

[1] In addition to these empirical problems of phrasal approaches, the underlying formalism of HPSG makes it impossible to express a genuinely phrasal analysis. The reason for this lies in HPSG's notion of *locality*. Every linguistic object needs to satisfy all constraints of the grammar (Richter, 2019). For idioms, this means that every idiomatic word must be licensed by the grammar all by itself. In other words, if an idiom such as *kick the bucket* is assigned an internal structure, every node in this structure needs to be licensed by the grammar as well.

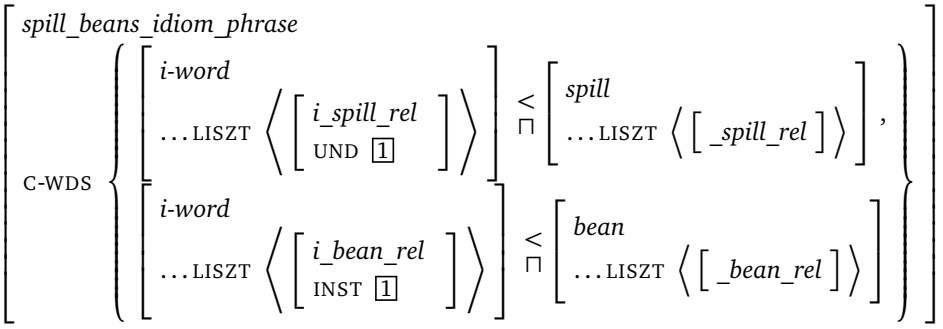

Figure 1: Description of the idiom *spill beans* in Riehemann (2001, 192)

The words in the WORDS repository are identical (including their subcategorization requirements) to the terminal nodes of the syntactic tree of the phrase. In a phrase that consists of or contains an idiom, the words are not only stored in WORDS but also divided up into two different sub-repositories of WORDS: CONSTRUCTIONAL-WORDS (C-WORDS) and OTHER-WORDS (O-WORDS). C-WORDS contains all and only the words that are part of the idiom.

Due to the fact that Riehemann (2001) defines idioms as phrases that consist of two or more words of which at least one is an idiomatic word, there must be two or more words in C-WORDS and at least one of them must be idiomatic. O-WORDS is the complementary repository to C-WORDS and, therefore, contains all and only the words that are *not* part of the idiom, which are always non-idiomatic. The reason why Riehemann (2001) allows for non-idiomatic words in her idiomatic phrases (and hence in C-WORDS) are idioms in which at least one of the words has its literal meaning, as in *miss the boat* 'miss out'. Since C-WORDS and O-WORDS are the only sub-repositories of WORDS, the union of their members results in the members of WORDS again.

At the level of the complete utterance, a 'head count' is carried out to ensure that all and only those idiomatic words are present that originated in C-WORDS. This guarantees that idiomatic words only appear when licensed by an idiom phrase, so that no idiom is incomplete. If a part of an idiom is present at the level of the complete utterance, the other parts have to be present as well and stand in the appropriate semantic relationship. The way the mechanism is built, it requires a one-to-one correspondence of idiomatic words in the structure and idiomatic words on phrasal C-WORDS lists.

Compared to combinatorial accounts, Riehemann's phrasal account offers the general advantages of phrasal analyses mentioned above: There is no need for individual lexical entries of the idiomatic uses of words, and there is a central place in which the idiom is defined as a whole. At the same time, the approach is subject to two empirical problems of such accounts. The pronominalization of idiom parts as in (7), is one of these cases. For example, in (7), the WORDS list of the sentence contains the idiomatic word *spill* and the pronoun *them*. However, it is plausible that the pronoun *them* differs from the literal word *beans*

in more than just the semantics. Consequently, the pronoun cannot asymmetrically default-unify with the literal word *beans*. This means that there cannot be the required two idiomatic words on the C-WORDS list.

The second problem is exemplified by data such as (6): There are two occurrences of the word *spill*, but only one of the word *beans*. The mechanism for checking the occurrence of the correct C-WORDS does not allow this.

In addition to these empirical problems, the technical realization of the underlying idea is not fully satisfactory. First, the WORDS mechanism is not used for anything other than the licensing of idioms. Second, the mechanism of asymmetric default unification is equally not part of the core machinery of HPSG.

To summarize, Riehemann's approach tries to capture the flexibility of combinatorial approaches with the conceptual advantages of phrasal accounts. For this reason, we will take her analysis as our basis and propose modifications to solve its problems.

4 A new phraseo-combinatorial analysis

The proposal that we will present in this section conserves the basic ideas of Riehemann (2001) but expresses them in a different way. First, we will encode Riehemann's "$\leq_{\sqcap}$" as a lexical rule. Second, we will replace the WORDS mechanism with a constraint on idiomatic phrases and a collocational restriction on idiomatic words. Third, we follow Webelhuth et al. (2018) in assuming that the completeness requirement on idioms is semantic rather than syntactic, let alone phrasal. We will present the ingredients of our analysis step by step – in a simplified version in Section 4.1, and in a refined version in Section 4.2.

Throughout the paper, we are largely agnostic with regard to the type of semantic approach to be assumed. We only need to assume that there is a semantic constant that is associated with a particular reading of a word. This constant would be the value of the RELN attribute in Pollard & Sag (1994), of MAIN in Richter & Sailer (2004), or of LID in Sag (2012). In this paper, we will simply call this attribute RELN.

4.1 Basic version of the analysis

Riehemann establishes a relation between a literal and an idiomatic version of words that occur in idioms in terms of her asymmetric default unification. We will express Riehemann's idea as an *object-level lexical rule* à la Meurers (2001). This is a well-defined and commonly recognized mechanism. It is a natural choice for us, as there is a clear connection between the two mechanisms: A lexical rule expresses the differences between its input and its output, with the assumption that anything not specified is taken over. Similarly, Riehemann's operator is intended as saying that the literal use of a word and its idiomatic use share all properties except of those explicitly specified by the idiom.

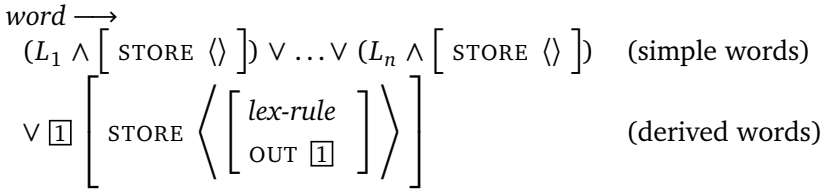

Figure 2: The Word Principle from Meurers (2001, 176)

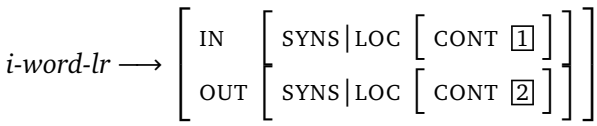

Figure 3: Constraint on the sort *i-word-lexical-rule* (*i-word-lr*; 1st version)

Meurers (2001) introduces a sort *lexical-rule* with two attributes, IN and OUT, both of which take a *word* object as their value. Such *lexical-rule* objects occur inside words. Meurers (2001) defines an attribute STORE on the sort *word*. The value of this attribute is a list, which is empty in the case of a simple word. For derived words, the STORE value contains the *lexical-rule* object which licenses the derived word. This is expressed by identifying the derived word with the output of the lexical rule, i.e., with the OUT value. Meurer's version of the Word Principle is given in Fig. 2.

For our cases, we introduce a special lexical rule for idiom components. We assume a sort *idiomatic-word-lexical-rule* (*i-word-lr*), which is a subsort of *lexical-rule*. A first version of this lexical rule is given in Fig. 3. The input of the lexical rule specifies the literal version of an idiom component. The output of the rule specifies the properties of the idiom-specific use of the same word. The way the rule is stated, it only requires that the input and the output differ with respect to their CONT, i.e., that the idiomatic word differs in meaning from its non-idiomatic base. In the example, we use *spill-id* and *bean-id* for the meaning of *spill* and *bean* as they occur in this idiom.

The next ingredient of our theory is a phrasal constraint that actually defines an idiom. Again, we follow the basic ideas from Riehemann (2001). We assume that each phrase has an additional attribute C(ONSTRUCTION)-W(OR)DS. The value of this attribute is a list of *i-word-lexical-rule* and *word* objects. It contains the specification of an *i-word-lr* for each idiomatic word, and of a *word* for each literal word that is an obligatory component of the idiom.[2]

Ordinary phrases have an empty C-WDS list. We provide the phrasal specification for the idiom *spill beans* in Fig. 4. Note its striking similarity to Riehemann's analysis, as given above in Fig. 1.

We introduce a *Lexicon of idiomatic expressions*, given in Fig. 5. This is a constraint on phrases with a non-empty C-WDS list. The consequent of the con-

[2]We will not discuss idioms containing words in their literal meaning, such as *miss the boat*. Nonetheless, we will formulate our constraints in a way compatible with these cases.

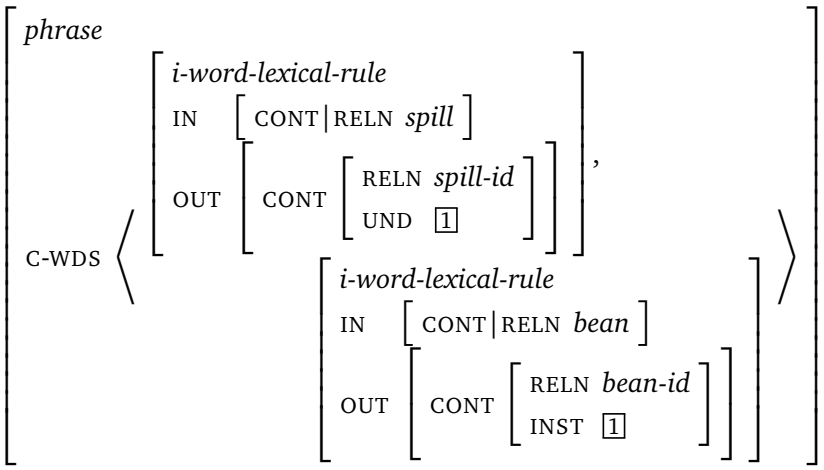

Figure 4: Sketch of the description of the idiomatic phrase *spill beans*

straint is a disjunction of descriptions of the idiomatic expression of the language. The individual entries of the idioms are heavily simplified in the figure. Each of them will look like the example in Fig. 4. Any phrase with a non-empty C-WDS list must satisfy one of the disjuncts in Fig. 5, i.e., must be an instantiation of an idiom.

So far, we have seen how to write the specification of the properties of an idiom. What is missing is how this will be put to work in a sentence. To do this, we need two constraints: First, a constraint that guarantees that whenever a particular idiomatic phrase is used, it dominates all words that constitute this idiom. Second, whenever an idiomatic word is used, it must be dominated by a corresponding idiomatic phrase.

The first of these constraints can be expressed straightforwardly. We provide a first version of it in (8). It enforces that an idiomatic phrase dominates the words that constitute the idiom. These can be words in their literal meaning or in their idiom-specific use – i.e. the words in the OUT value of an *i-word-lr* object.

(8) Idiom Completeness Constraint (first version):
A phrase with a non-empty C-WORDS list must dominate words identical with the elements on its C-WORDS list or their OUT values.

We need some additional mechanism for the second constraint. As this constraint is concerned with the distribution of lexical elements, we adopt a version of the HPSG collocation theory, presented in Richter & Sailer (2003) for bound words, in Richter & Soehn (2006) for negative polarity items, and in Soehn (2009) additionally for external allomorphy (such as the *a/an* alternation). This theory is assumed as a prerequisite in a number of existing HPSG analyses of idioms, such as Sailer (2003), Soehn (2006), or Webelhuth et al.

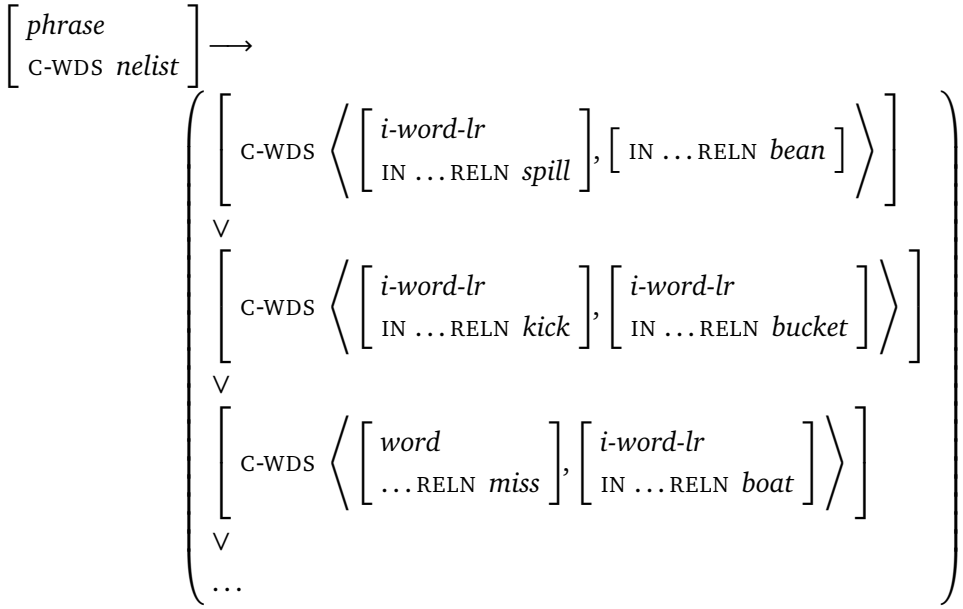

Figure 5: Lexicon of idiomatic expressions

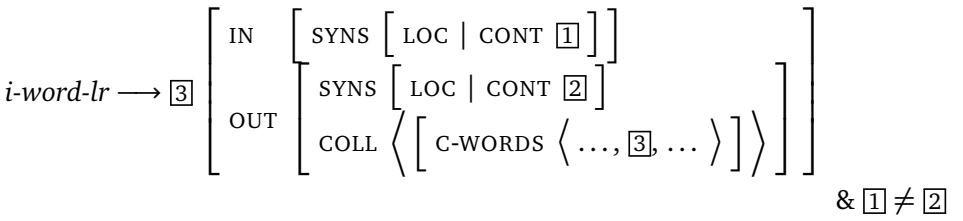

Figure 6: Constraint on the sort *i-word-lr* (final version, with collocation)

(2018). Consequently, we do not introduce additional machinery by using it.

In its simplest version, the collocation theory consists of a list-values attribute COLL, which is defined on lexical items. The elements of this list are *sign* objects. Finally, there is a constraint that a lexical item can only occur in a structure in which it is dominated by each of the elements on its COLL list.

In collocational approaches to idioms such as the ones just mentioned, an idiomatic word is collocationally restricted to co-occur with the other words that belong to the idiom. In our approach, the output of the *i-word-lr* needs to be collocationally restricted to occur within a phrase that licenses the idiom. In other words, the output of the *i-word-lr* is collocationally restricted to a phrase that has this instantiation of the lexical rule on its C-WORDS list. This is expressed in the revised version of the constraint on the sort *i-word-lr* in Fig. 6.

The Idiom Completeness Constaint in (8) and the revised version of the *i-word-lr* in Fig. 6 have the desired effect: First, if there is an idiomatic phrase, it must dominate words that correspond to the OUT values of the elements in

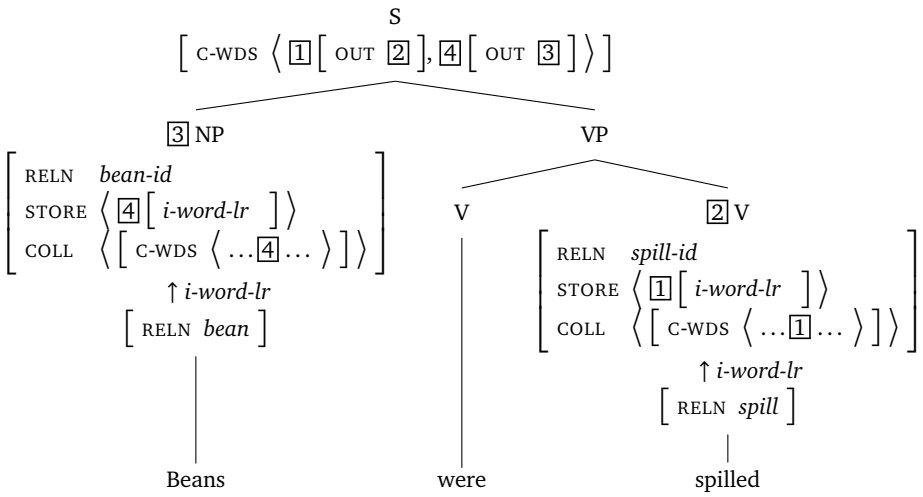

Figure 7: Sketch of the analysis of sentence (9b)

the phrase's C-WDS list. Second, if there is an idiomatic word, it requires the presence of an idiom on whose C-WDS list it is.

Even in this preliminary form, our approach allows us to capture some basic properties of idioms. The idiom *spill beans* is syntactically relatively flexible. As shown in (9), the idiom can occur in a active as well as in a passive form. We sketch our analysis of the passive example in Fig. 7.

(9) a. Alex spilled the beans.
b. Beans were spilled.

The sentence in (9b) contains instances of the two idiomatic words *beans* and *spilled*. As indicated in the structure in Fig. 7 by "↑" these words are the output of an application of the *i-word-lr*. Consequently, they have a collocational requirement that they must be dominated by some phrase on whose C-WDS list these idiomatic words are found. This triggers the occurrence of the phrase, specified in Fig. 4. It is, therefore, guaranteed that whenever there is an idiomatic word, it enforces the presence of a disjunct from the lexicon of idiomatic expressions in Fig. 5. On the other hand, once there is a phrase with a non-empty C-WDS list in a structure, the Idiom Completeness Constraint from has the effect that this phrase must dominate all words that are relevant for a particular idiom.

The ungrammatical example (10) contains only one part of the idiom. Our theory correctly excludes this sentence under an idiomatic reading.

(10) \$ Alex told me the beans. ≠ 'Alex told me the secrets.'

If the word *beans* is used idiomatically, it must be dominated by a phrase which has on its C-WDS list two *i-words-lr* objects, one for the idiomatic version

of *beans* and one for the idiomatic version of *spill*. This phrase, then, must dominate words identical to the OUT values of its C-WDS elements. In this particular sentence, however, the idiomatic use of the word *spill* is missing.

The basic version of our phraseo-combinatorial approach has all the basic components: It is combinatorial in that idiomatic words have exactly the properties stipulated for them in corresponding purely combinatorial approaches. We have chosen a collocational variant of a combinatorial approach rather than the selection-based one of Kay et al. (2015) and Michaelis (2019). Our approach is phrasal in that there are no lexical entries for those idiomatic words but a single specification of the idiom as a whole in terms of a phrasal specification. This specification, however, does not constrain the type of phrase but only the idiomatic words that it must dominate. It is at the phrasal level that the mapping from the literal to the idiomatic word is constrained.

4.2 Final version of the analysis

The basic version of our account presented in Section 4.1 does not yet capture the full syntactic and semantic flexibility that we discussed in Section 2. In particular, the insight that the completeness requirement of an idiom is semantic in nature is not yet encoded. To implement this insight, we will loosen the co-occurrence constraints in the Idiom Completeness Constraint to the relevant parts of the semantic representation.

The need for such a refinement can be illustrated with examples such as (7), repeated for convenience in (11). Here, one of the obligatory parts of the idiom is realized by a pronoun, *them*, rather than by a noun phrase containing the idiomatic word *beans*.

(11) Eventually she spilled all the beans$_k$. But it took her a few days to spill them$_k$ all.

We can adopt the solution in Webelhuth et al. (2018): The idiomatic phrase only requires the occurrence of a word with the relevant idiomatic content, i.e. with the relation *beans-id*. This is expressed in the final version of the Idiom Completeness Constraint in (12).

(12) Idiom Completeness Constraint (final version):
For each phrase *p* and for each object *o* on *p*'s C-WORDS list,
p dominates a sign whose CONT|RELN value is identical with *o*'s CONT|RELN value or its OUT|...|CONT|RELN value.

In the critical example (11), the antecedent of the pronoun *them* is the noun phrase *the beans* from the previous sentence. We assume with Webelhuth et al. (2018) that the basic content is among the things shared between an anaphorically used pronoun and its antecedent. This includes enough information to fulfill the completeness requirement of the idiom. Adapting this to the present

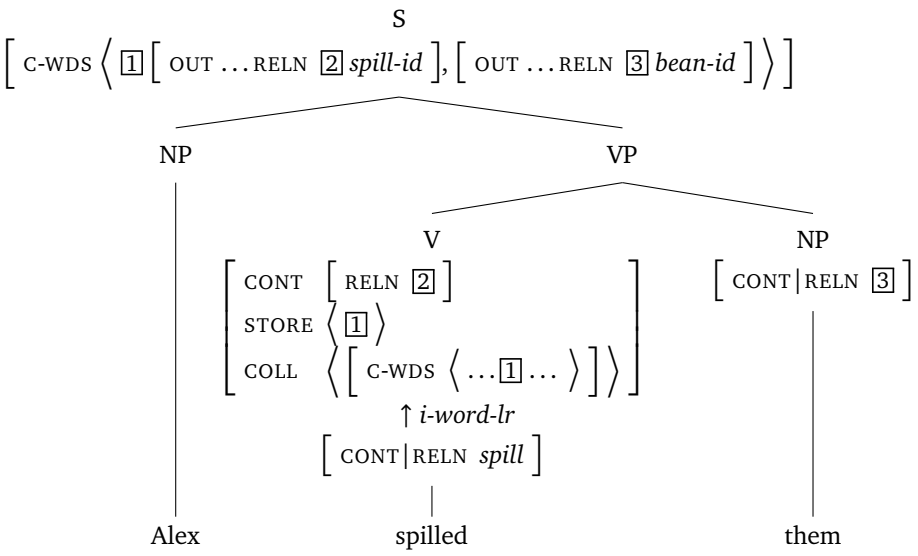

Figure 8: Sketch of the analysis of the sentence *Alex spilled them*.

architecture, we assume that the RELN value of the pronoun is identical with that of its antecedent. This is sufficient to satisfy the Idiom Completeness Constraint in its final version in (12) – though it was not enough to satisfy the earlier form of this constraint in (8).

We provide an analysis of a simple sentence with a pronominalized idiom part, *Alex spilled them*, in Fig. 8. The idiomatic word *spilled* is licensed as it is dominated by a phrase in which the particular mapping from non-idiomatic *spill* to its idiom-specific version is specified. The pronoun *them* is not idiomatic and, as such, has no idiom-specific distributional requirements. It has, however, certain restrictions as a discourse anaphora. These include that parts of its semantics, in particular its RELN value, are identical with those of its antecedent. The top node in the tree is the idiomatic phrase. It satisfies the Idiom Completeness Constraint because it dominates signs with the RELN values required in the outputs of the *i-word-lr* objects of the idiomatic phrase's C-WDS list – namely the idiomatic word *spilled* and the discourse anaphoric pronoun *them*.

The remaining problem for the current version is that a single idiomatic phrase should be able to license several occurrences of an idiomatic word. We saw a relevant example above in (6), repeated as (13).

(13) The beans [$_{VP}$ [$_{VP}$ have not been spilled yet], but [$_{VP}$ will be spilled very soon]].

This sentence contains two occurrences of the word *spill*. They cannot both be identical with the verbal element on the phrase's C-WDS list. For this reason we need to allow for multiple occurrences of idiom parts on the C-WDS

list. To achieve this, we introduce some underspecification in the description of idiomatic phrases. The required change is shown schematically in (14). The phrasal description in Fig. 4 has the form given in (14a), with just two *i-word-lr* objects ρ_1 and ρ_2 on its C-WDS list. We modify this in the way specified in (14b), i.e., by adding a meta-description operator Δ which is defined below the AVM.

(14) a. Phrasal constraint with fully specified C-WDS list:

$$\begin{bmatrix} \textit{phrase} \\ \text{C-WDS} \ \langle \rho_1, \ldots \rho_n \rangle \end{bmatrix}$$

b. Phrasal constraint with underspecified C-WDS list:

$$\begin{bmatrix} \textit{phrase} \\ \text{C-WDS} \ \Delta\big(\langle \rho_1, \ldots \rho_n \rangle\big) \end{bmatrix}$$

where for each list L, $\Delta(\langle \rho_1, \ldots, \rho_n \rangle)$ describes L iff for each element e of L, there is a list L_e of elements of L such that e is on L_e and L_e is described by $\langle \rho_1, \ldots, \rho_n \rangle$.

We implement this change in all definitions of idiomatic phrases such as the one in Fig. 4, i.e., we introduce the operator Δ in the description of the value of the C-WDS list. This underspecified version allows us to account for sentence (13). The corresponding structure of a simplified version of this sentence is given in Fig. 9.

The top node in Fig. 9 is the relevant idiomatic phrase. It contains two instances of the lexical rule that licenses the idiomatic word *spill*. This constellation is licensed by the underspecified version of Fig. 4. The C-WDS list of the phrase is the relevant list L from the definition. For each of its elements, we can find the necessary subparts: For the first element, $\boxed{1}$, the list $\langle \boxed{1}, \boxed{5} \rangle$ satisfies the original description. For the second element, $\boxed{3}$, the relevant list is $\langle \boxed{3}, \boxed{5} \rangle$. Finally, either list is a possibility for the third element of the list, $\boxed{5}$. This shows that the C-WDS list of the top node in Fig. 9 satisfies our modified description of the idiomatic phrase.

The resulting structure also satisfies all other constraints introduced in this paper. For each of the idiomatic words, there is an element on the phrase's C-WDS list that is identical to the word's STORE value, i.e., the COLL requirements of the idiomatic words are satisfied. The overall phrase meets the Idiom Completeness Constraint from (12) as well: For each of its elements there is a word in the phrase whose semantics is identical to that specified in the phrase's C-WDS list.

We should briefly turn to the examples of idioms in relative clauses from (5), repeated in (15). In (15a), the idiomatic phrase would be some phrase dominating idiomatic *strings* and the relative clause. This phrase contains the *i-word-lr* object on its C-WDS list that license idiomatic *strings* and idiomatic *pull*. This satisfies the collocational constraint of the idiomatic words. As the gap in the relative clause and the noun *string* have an identical index, the linking requirement of the phrase is satisfied, which is the same as for *spill beans* in

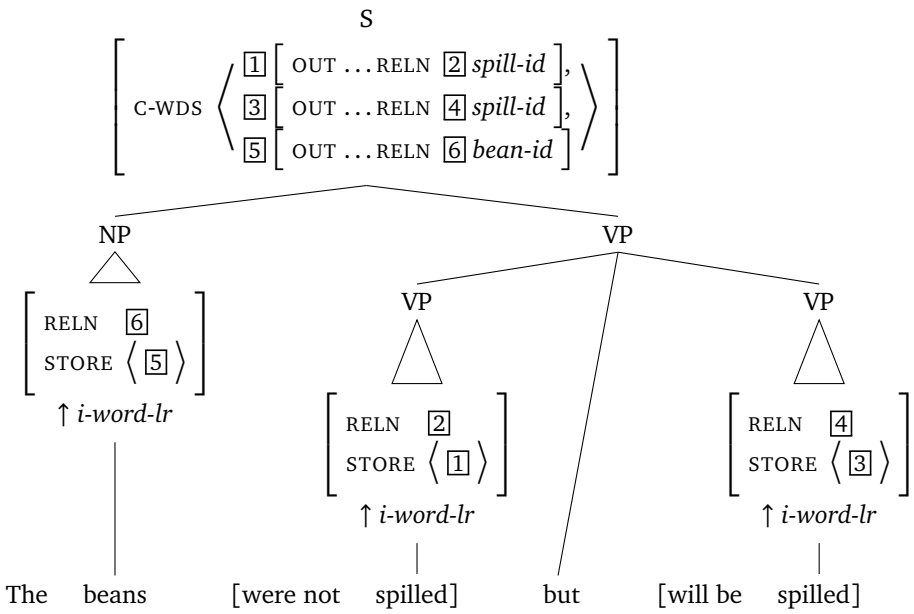

Figure 9: Sketch of the analysis of a simplified version of sentence (13)

Fig. 4. Finally, by dominating the two idiomatic words, the phrase satisfies the Idiom Completeness Constraint.

(15) a. The strings$_k$ [$_{RC}$ that Pat pulled ___$_k$] got Chris the job.
b. John never pulled the strings$_k$ [$_{RC}$ that his mother told him should be pulled ___$_k$].

The case in (15b) is only slightly more complex. The idiomatic phrase can be the matrix verb phrase. It contains two occurrences of idiomatic *pull* and one occurrence of idiomatic *strings*. Consequently, it is a variant of the case illustrated in (13) and in Fig. 9. The matrix occurrence of *pull* and the noun *strings* satisfy the phrasal description just as the two idiomatic words in Fig. 7. The occurrence of idiomatic *spill* inside the relative clause and the noun *strings* satisfy the description in the same way they do in sentence (15a).

We can briefly summarize our implementation of Riehemann's analysis before closing this section. A grammar writer can specify an idiom as a description of a phrase with a non-empty C-WORDS list. In this specification, idiomatic words are related to their non-idiomatic base. There is no need to add lexical entries for those idiomatic words. Consequently, the idiom can be defined in one central spot, as a disjunct in a constraint on phrases with a non-empty C-WORDS list. In the analysis of a sentence, however, the idiomatic words combine just as ordinary words, which gives us the full flexibility of combinatorial accounts of idioms.

In this section, the leanness of our approach for integrating idioms into an HPSG grammar may have been lost in the technical details: embedding lexical rules inside a list-valued feature on a phrase, making use of a collocation mechanism, and, finally, even adding a layer of underspecification to the C-WORDS lists. It is important to take these points as what they are: a technical implementation that is fully defined and that can simply be taken for granted.

5 Conclusion

In this paper, we took a resuming view on the formal research on idioms, in particular within HPSG and related frameworks, focusing mainly on the divide between phrasal and combinatorial approaches. We noted a number of empirical advantages of combinatorial accounts over phrasal accounts, which is reflected in the dominance of combinatorial approaches in recent HPSG and SBCG analyses. On the other hand, such approaches seem conceptually problematic as they disregard the unit-like nature of idioms. Riehemann (2001) had already tried to mediate between these two positions, but her approach could not fully achieve this goal. Taking her insights and her analysis as our starting point, we propose a new phraseo-combinatorial approach that can be seen as a re-implementation of Riehemann's original ideas, extended to cover a wider range of data.

The resulting implementation is admittedly rather technical in its details. If one accepts the proposed constraints and the proposed lexical rule as part of the grammar, our approach allows for a straightforward encoding of idioms. It is combinatorial, but avoids separate lexical entries for uses of words inside idioms ("lexical explosion"). Instead, we can represent each idiom as a single, holistic unit. At the same time, we do not bind the characterization of an idiom to a particular constituent structure, but rather to the co-occurrence of lexical items with a particular meaning.

Let us briefly point to a potential extension of our theory. Egan (2008) discusses data as those in (16), which require the simultaneous presence of the idiomatic and the literal reading of the words constituting an idiom.[3]

(16) The strings we've been pulling to keep you out of prison are fraying badly. (Egan, 2008, 391)

To our knowledge, none of the other approaches that we have mentioned in this paper so far, be they phrasal or combinatorial, provide simultaneous access to the literal and the idiomatic reading. Findlay et al. (2019) is a first attempt towards a systematic understanding of the data. We have to leave an analysis

[3]Data on so-called *conjunction modification* such as *He bit his thirst-swollen tongue* (Ernst, 1981, 59) clearly are another case in which the literal and the idiomatic meaning of an expression are simultaneously used in the interpretation. They are discussed in detail in Bargmann et al. (2021).

of these data for future research. However, our approach could be a promising starting point as we assume the presence of both the literal and the idiomatic use of a word in the structure.

References

Bargmann, Sascha, Berit Gehrke & Frank Richter. 2021. Modification of literal meanings in semantically non-decomposable idioms. In Berthold Crysmann & Manfred Sailer (eds.), *One-to-many relations in morphology, syntax, and semantics*, 245–279. Berlin: Language Science Press. doi:10.5281/zenodo.4729808.

Bargmann, Sascha & Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 1–29. Berlin: Language Science Press. doi:10.5281/zenodo.1182587.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

Egan, Andy. 2008. Pretense for the complete idiom. *Noûs* 42(3). 381–409.

Ernst, Thomas. 1981. Grist for the linguistic mill: Idioms and 'extra' adjectives. *Journal of Linguistic Research* 1(3). 51–68.

Findlay, Jamie Y. 2019. *Multiword expressions and the lexicon*: University of Oxford dissertation. http://users.ox.ac.uk/~sjoh2787/findlay-thesis.pdf.

Findlay, Jamie Y., Sascha Bargmann & Manfred Sailer. 2019. Why the butterflies in your stomach can have big wings: combining formal and cognitive theories to explain productive extensions of idioms. Presentation at Europhras 2019, Santiago de Compostella.

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

Kay, Paul, Ivan A. Sag & Dan Flickinger. 2015. A lexical theory of phrasal idioms. Manuscript. www1.icsi.berkeley.edu/~kay/idiom-pdflatex.11-13-15.pdf.

Krenn, Brigitte & Gregor Erbach. 1994. Idioms and support verb constructions. In John Nerbonne, Klaus Netter & Carl Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*, 365–396. Stanford: CSLI Publications.

Kuno, Susumu & Ken-ichi Takami. 2004. *Functional constraints in grammar: On the unergative-unaccusative distinction*. Amsterdam and Philadelphia: Benjamins.

Lascarides, Alex & Ann Copestake. 1999. Default representation in constraint-based frameworks. *Computational Linguistics* 25. 55–105.

Meurers, Walt Detmar. 2001. On expressing lexical generalizations in HPSG. *Nordic Journal of Linguistics* 24(2). 161–217. doi:10.1080/033258601753358605. Special issue on 'The Lexicon in Linguistic Theory'.

Michaelis, Laura A. 2019. Constructions are patterns and so are fixed expressions. In Beatrix Busse & Ruth Moehlig-Falke (eds.), *Patterns in language and linguistics: New perspectives on a ubiquitous concept*, vol. 104 Topics in Linguistics, 193–220. Berlin: Mouton de Gruyter.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago and London: University of Chicago Press.

Richter, Frank. 2019. Formal background. In Stefan Müller, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.), *Head-Driven Phrase Structure Grammar: The handbook*, Berlin: Language Science Press. https://hpsg.hu-berlin.de/Projects/HPSG-handbook/PDFs/formal-background.pdf. Prepublished version.

Richter, Frank & Manfred Sailer. 2003. Cranberry words in formal grammar. In Claire Beyssade, Olivier Bonami, Patricia Cabredo Hofherr & Francis Corblin (eds.), *Empirical issues in formal syntax and semantics*, vol. 4, 155–171. Paris: Presses Universitaires de Paris-Sorbonne.

Richter, Frank & Manfred Sailer. 2004. Basic concepts of Lexical Resource Semantics. In Arne Beckmann & Norbert Preining (eds.), *ESSLLI 2003: Course material I*, vol. 5 Collegium Logicum, 87–143. Vienna: Kurt Gödel Society Wien.

Richter, Frank & Jan-Philipp Soehn. 2006. *Braucht niemanden zu scheren*: A survey of NPI licensing in German. In Stefan Müller (ed.), *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar, Varna 2006*, 421–440. http://cslipublications.stanford.edu/HPSG/2006/richter-soehn.pdf.

Riehemann, Susanne Z. 2001. *A constructional approach to idioms and word formation*: Stanford University dissertation.

Sag, Ivan A. 2012. Sign-Based Construction Grammar: An informal synopsis. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*, 69–202. Stanford: CSLI Publications.

Sailer, Manfred. 2003. Combinatorial semantics and idiomatic expressions in Head-Driven Phrase Structure Grammar. Phil. Dissertation (2000). Arbeitspapiere des SFB 340. 161 Universität Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/46191.

Soehn, Jan-Philipp. 2006. *Über Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG*. Frankfurt am Main: Peter Lang. Ph.D. thesis, Friedrich-Schiller-Universität Jena.

Soehn, Jan-Philipp. 2009. Lexical licensing in formal grammar. http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-42035.

Wasow, Thomas, Ivan A. Sag & Geoffrey Nunberg. 1983. Idioms: An interim report. In S. Hattori & K. Inoue (eds.), *Proceedings of the XIIIth International Congress of Linguistics*, 102–115.

Webelhuth, Gert, Sascha Bargmann & Christopher Götze. 2018. Idioms as evidence for the proper analysis of relative clauses. In Manfred Krifka, Rainer Ludwig & Mathias Schenner (eds.), *Reconstruction effects in relative clauses*, 225–262. Berlin and Boston: de Gruyter.