A phraseo-lexical analysis of idioms

Manfred Sailer & Sascha Bargmann
(Goethe-University Frankfurt a.M.)

The formal analysis of idiomatic expressions (IE) has been oscillating between phrasal and lexical analyses, i.e., analyses that emphasize the unit-like character of IEs and those that exploit the autonomy of their component parts. We will summarize the main arguments and then propose an analysis that tries to capture the insights of both positions. The resulting theory is heavily based on Riehemann (2001).

Lexical versus phrasal analyses

Early Generative approaches, Chomsky (1965), consider all IEs as lexical units with internal structure. Later formal approaches take a more differentiated position. Wasow et al. (1983) and Nunberg et al. (1994) argue that idioms should be split into two categories, based on whether the idiomatic meaning needs to be assigned to the idiom as a whole or can be distributed over its component parts. The two types of IEs are usually illustrated with the non-decomposable idiom *kick the bucket*, which has the semantic representation **die**, and the decomposable idiom *spill beans*, whose meaning can be arrived at compositionally if one interprets the verb *spill* as **reveal** and the noun *beans* and **information**. This special interpretation of the words is strictly confined to the IE. The basic insight in Wasow et al. (1983) and Nunberg et al. (1994) is that the semantic decomposability correlates with syntactic flexibility, i.e., decomposable idioms can undergo syntactic processes such as passivization, topicalization, or the insertion of adjuncts, whereas non-decomposable idioms cannot – see (1), where "$" indicates the unavailability of an idiomatic reading.

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

This distinction motivated a lexical analysis of decomposable idioms in GPSG (Gazdar et al., 1985) and, subsequently in HPSG (Krenn & Erbach, 1994; Sailer, 2003; Soehn, 2009). More recently, empirical evidence was put forward against the syntactic fixedness of non-decomposable idioms, see (2). This let to an extension of the lexical analysis to non-decomposable idioms (Kay et al., 2015; Bargmann & Sailer, 2018).

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

Lexical analyses of IEs postulate separate lexical entries for the words that occur in IEs, which have the syntactic and semantic properties required for the idiom. With this assumption, the idiomatic meaning can be derived by the orderinary mechanisms of syntactic and semantic combinatorics.
Findlay (2019, Section 3.3) points out two basic problems of lexical approaches: First, the *collocational challenge* that the idiomatic version of the words need to be prevented from occurring outside the IE. Second, the *lexical explosion problem* that there is a new lexical entry for each word in each IE – but no single lexical entry for the IE as a whole. Of these problems, only the collocational challenge has been addressed in HPSG/SBCG: by an extended notion of selection (Krenn & Erbach, 1994; Kay et al., 2015), or an explicit integration of a collocational component (Sailer, 2003; Soehn, 2009).
Phrasal, or constructional, approaches to IEs are natural in formal frameworks that are based on tree grammars, such as the *Tree Adjoining Grammar* analysis in Abeillé (1995) and the tree grammar-based version of LFG in Findlay (2019): There is a single, phrasal description of an IE. The idiomatic meaning can be assigned to the entire tree (for non-decomposable idioms) or to substructures in the tree (for decomposable idioms). Constructional analyses have neither of the above-mentioned problems of lexical approaches: First, as the parts of the idiom are only connected to an idiomatic meaning inside the IE, they cannot be used with this meaning in other combinations. Second, there is no need for idiom-specific lexical entries for the IE parts as they are licensed directly through the phrasal lexical entry. Furthermore, the IE is introduced into the (phrasal or constructional) lexicon as a unit.
Even in a very recent form, as in Findlay (2019), the constructional approach to idioms inherits problems of its more classical versions: First, it is not fully clear how the syntactic flexibility is permitted and constrained in the appropriate way without recurring to diacritic marking of the possible syntactic configurations – as in Fraser (1970); Abeillé (1995). There are also some syntactic constellations for which constructional approaches just seem not to be the appropriate analytic tool. We will point to three such cases, based on Webelhuth et al. (2018).

First, idiom parts split over a matrix clause and a relative clause (McCawley, 1981), see (3). Attaching a relative clause to an idiomatic word, as in (3-a), is not a problem for Findlay (2019), as decomposable idioms allow for modifiers in general. Findlay (2019, 307) shows that the tree associated with a decomposable idiom can serve as the input for a rule that outputs the form *bean* [*that NP pulled*], which is needed for (3-b). A problem arises, however, in (3-c). Here, the idiomatic word *strings* would have to be part of two IE trees that combine. It is not obvious how this would be captured in the tree grammar developed in Findlay (2019). Webelhuth et al. (2018) show that such data are unproblematic for a lexical account: in all cases in (3), the collocational requirements of each idiomatic word are satisfied.

(3) a. Parky pulled the strings that got me the job. (McCawley, 1981, 137)
b. The strings that Pat pulled got Chris the job. (Nunberg et al., 1994, 510)
c. John never pulled the strings that his mother told him should be pulled. (Webelhuth et al., 2018)

The second problem for which there has not been a proposal within phrasal accounts of IEs concerns the pronominalization of idiom parts, as in (4). In (4-a), the idiomatic verb occurs in the second clause but its direct object is a pronoun rather than the required idiomatic word. In (4-b), the second clause contains a pronoun refering to the idiomatic strings, though not the rest of the idiom. Again, a paradox arises: if part of an idiom can be substituted with a pronoun in a phrasal lexical entry – as needed for (4-a), we would expect (4-b) to be ungrammatical, as the pronoun *they* would require an occurrence of the idiomatic use of *pull*. Webelhuth et al. (2018) provide an elegant lexical account of the data, in which the pronoun in (4-a) is interpreted as a definite description which contains the necessary information to satisfy the collocational restriction of the idiomatic verb *spill*. As for (4-b), the authors argue that nothing needs to be said as there is no occurrence of an idiomatic word in the second clause and only overtly present idiomatic words impose collocational constraints.

(4) a. Eventually she spilled all the beans. But it took her a few day to spill them all. (Riehemann, 2001, 207)
b. Kim's family pulled some strings on her behalf, but they weren't enough to get her the job. (Nunberg et al., 1994, 502)

The third fundamental empirical problem for phrasal approaches concerns apparently free occurrences of idiom parts, as in (5). Webelhuth et al. (2018) show that such cases are severely restricted contextually. In particular, the full idiom must have been mentioned explicitly in the previous context and must still be salient. Webelhuth et al. argue that idiomatic words can be used if they have been licensed previously and their semantics is still salient. This is, essentially, an anaphoric analysis of such idiom parts.

(5) Pat and Chris graduated from law school together with roughly equal records. Pat's uncle is a state senator, and he pulled strings to get Pat a clerkship with a state supreme court justice. Chris, in contrast, didn't have access to any strings, and ended up hanging out a shingle. (Wasow et al., 1983, 113)

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 IEs, 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.

Riehemann's approach

Riehemann (2001) seems to be the only approach in HPSG that acknowledges the locality dilemma of any HPSG theory of IEs but that still attempts to provide a holistic account of idioms. I will briefly sketch her approach and the difficulties it faces.

Riehemann assumes that there is a single, holistic, constructional constraint for each idiom – as shown for *spill beans* in Figure 1. She introduces a set-valued feature C(ONSTRUCTIONAL)-WORDS on idiomatic phrases. The C-WORDS value specifies exactly which words constitute the idiom. The IE *spill beans* can be attested whenever there is a phrase that dominates the idiomatic forms of the words *spill* and *beans* in the right constellation, i.e., where the referent of *beans* has the semantic role of an undergoer of the verb *spill*. Riehemann (2001, 189) also assumes idiomatic words, of type *i-word*. While there is a subtype for each idiomatic word, there is no constraint on them. Instead, she introduces the default unification operator "$\overset{\leq}{\sqcap}$". This operator specifies that the idiomatic words that occur in the constraint

in Figure 1 satisfy all specifications given in the lexical constraint on the words *spill* and *bean* except for those mentioned in the phrasal constraint on the IE. This makes it possible to avoid separate, idiom-specific lexical entries, i.e., to avoid the lexicon explosion problem.
Riehemann (2001) also addresses the collocational challenge. In addition to the feature C-WORDS there are two more set-valued features: First, the feature WORDS is defined on all phrases. Its value contains all words dominated by a phrase. Second, the feature OTHER-WORDS – just like C-WORDS – is only defined on idiomatic phrases. It may not contain words of the type *i-word*. In an idiomatic phrase, the WORDS value must consist exactly of the elements in the sets C-WORDS and OTHER-WORDS. With this constraint, Riehemann (2001) attempts to exclude free uses of parts of an idiom.
The basic ideas underlying Riehemann's analysis point to a possibility to actually have a combination of a lexical and a phrasal account of IEs within HPSG. However, there are problems with the way in which she formalizes her insights. First, the default operator "$\stackrel{\leq}{\sqcap}$" lacks a clear definition – in particular within a default-free version of HPSG as the one proposed in Pollard & Sag (1994). Second, the WORDS-mechanism is problematic. It is rather complex and not connected to other parts of the grammar. It is not fully worked out and Riehemann seems to only scratch at the surface of various problematic constellations that might arise.

A phraseo-lexical analysis

We will present a Riehemann-style analysis of IEs that is fully compatible with the formal foundations of HPSG as summarized in Richter (2019). We will treat the "$\stackrel{\leq}{\sqcap}$"-operator as a *lexical rule* and use the *collocation module* of earlier, lexical approaches to IEs as summarized in Soehn (2009).

Phrasal idiom constraint We follow Riehemann (2001) in assuming that phrases may have a C-WORDS list in which they specify the words that they require to co-occur for a particular IE. We provide the specification for the IE *spill beans* in Figure 2. The C-WORDS value is analogous to Riehemann's C-WORDS set: it contains exaclty the words required by the IE. Instead of using the operator "$\stackrel{\leq}{\sqcap}$", we assume that the literal use of the idiomatic word appears as the input to a lexical rule and the idiom-internal version of this word appears as its output.
The lexicon is conceptualized in HPSG as a constraint on objects of type *word* in which all lexical entries appear as disjuncts. Similarly, the descriptions of IEs appear as disjuncts of a constraint on phrases with a non-empty C-WORDS lists, see Figure 3. What is special is that these disjuncts merely mention what elements occur in the C-WORDS list and do not specify any other syntactic or semantic properties of the phrase.

Idiomatic word lexical rule We follow lexical approaches in assuming that the words *spill* or *kick* have an idiom-specific meaning when used inside an IE – and, probably, also an idiom-specific lexical identification label (LID value). However, the idiom-specific words are the output of a lexical rule which takes the non-idiomatic words as its input. We adopt the encoding of lexical rules from Meurers (2001). I.e., there are objects of type *lexical-rule* that have an attribute IN and OUT whose values are the input and the output of the lexical rule respectively. The derived word has a STORE attribute which contains the lexical rule object. The word itself is, then, constrained to be identical with the OUT value of the lexical rule (Meurers, 2001). Furthermore, we use Meurer's short hand notation which means that the values of all features of the input and the output are identical unless explicitly mentioned in the lexical rule specification. The lexical rule needed for the IEs is given in Figure 4. In it, the input and the output only differ with respect to their LID value and their semantics. In addition, there is a collocational requirement.
The output of the lexical rule has a non-empty COLL value. This feature specifies the *context of lexical licensing* or *collocation requirement*. While there are various formulations of the collocation module in the literature, they all agree that if an element occurs in the COLL list of a word, that element must dominate the word in some larger structure. The COLL specification is one of the standard answers to the above-mentioned collocational challenge of lexical analysis of IEs in HPSG.
The idiomatic word requires that it be dominated by a phrase that contains the word's lexical rule on its C-WORDS list. This excludes any free occurrence of idiomatic words. For example the word *beans* cannot be used with its idiomatic meaning in the sentence *\$Alex was shocked by the beans*, because there is no sign dominating it whose C-WORDS value satisfies the idiomatic word's COLL requirement.
The COLL specification on idiomatic words makes sure that they do not occur outside their IE. We also need a principle to guarantee that whenever a particular IE is used, its components need to be present in the structure. In analogy to Riehemann (2001), we could propose that each element of a phrase's C-WORDS must be dominated by that phrase. However, as will become clear in the next section, a somewhat looser restriction is sufficient: the phrase only needs to dominate an element that

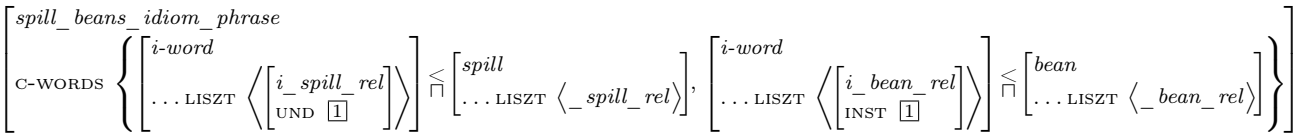

Figure 1: Analysis of *spill beans* in Riehemann (2001, 192)

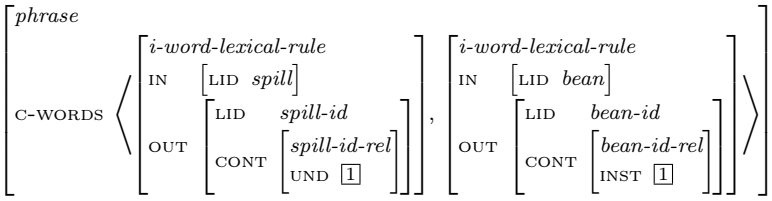

Figure 2: Phraseo-lexical lexical entry of *spill beans*

contributes the semantics specified for the idiomatic word on the C-WORDS list. This constraint is stated in (6).

(6) For each phrase p and for each object o, if o is on p's C-WORDS list,
then p dominates a sign whose CONT value is identical with o's OUT|...|CONT value.

The overall analysis is sketched in the left tree of Figure 5. The tree is just as under a lexical analysis, but the idiomatic words are derived by the *i-word-lexical-rule*. In addition, there is a phrase with a non-empty C-W(OR)DS list. The connection between the idiomatic words and this phrase is constrained in two ways: (i) The idiomatic words enforce the presence of such a phrase through their COLL specification. (ii) The phrase must dominate expressions with the idiomatic content (2 and 3).

Solving the problems of phrasal approaches

In this section, we will briefly show how the phraseo-lexical approach avoids the problems of phrasal approaches. Our theory differs crucially from tree-grammar accounts as it constrains which words can occur together rather than introducing these words into the structure. This immediately allows for cases like (3-c), where the idiomatic word *strings* only occurs once but there are two occurrences of the idiomatic word *pull*.

The pronominalization cases can be captured just as in Webelhuth et al. (2018): While an occurring idiomatic word must be on some phrase's C-WORDS list, the idiomatic phrases only require to dominate words that make the appropriate semantic contribution. If we assume with Webelhuth et al. (2018) that the relevant pronouns are interpreted as definite descriptions copying the core semantics of their antecedent, this constraint is satisified in (4-a). A simplified example of this constellation is shown in the right tree of Figure 5. The collocational requirement of the idiomatic word *spill* is satisfied just as in the tree on the left. In the case of (4-b), there is no occurrence of the IE, and nothing needs to be accounted for.

Finally, the case of apparently free occurrences of idiom parts as in (5) needs to be discussed. Again, the treatment of Webelhuth et al. (2018) can be adopted directly. We can consider this occurrence of the word *strings* as an *anaphoric*, *quote-like* use of the noun rather than as a use of the idiomatic word. Such a use is possible because the idiomatic word is introduced into the preceding discourse as an ordinary word and is, therefore, accessible for anaphoric processes just as other words.

Conclusion

This paper proposes a reconciliation of phrasal and lexical approaches to idioms. It integrates the insights of recent lexical analyses of idioms with the appeal of phrasal approaches. In particular, there is no need for having separate lexical entries for idiomatic words, yet the idioms are not bound to a particular syntactic structure.

The examples discussed in the paper are restricted to prototypical cases of idiomatic expression. If time permits, we will show how other types of idioms can be modelled – including idioms with only some idiomatic words (*miss the boat*) and idioms with bound words (*take umbrage*).

$$\left[\begin{array}{l}\textit{phrase}\\ \textsc{c-words}\ \ \textit{nelist}\end{array}\right] \Rightarrow \left(\left[\textsc{c-words}\ \left\langle\left[\textsc{in lid}\ \textit{spill}\right], \left[\textsc{in lid}\ \textit{bean}\right]\right\rangle\right] \textbf{ or } \left[\textsc{c-words}\ \left\langle\left[\textsc{in lid}\ \textit{kick}\right], \left[\textsc{in lid}\ \textit{bucket}\right]\right\rangle\right] \textbf{ or } \ldots\right)$$

Figure 3: Constraint on idiomatic phrases

$$\boxed{5}\left[\begin{array}{ll}\textit{i-word-lexical-rule} & \\ \textsc{in} & \left[\textsc{syns|loc}\ \left[\begin{array}{ll}\textsc{cat} & \left[\textsc{head|lid}\ \boxed{1}\right]\\ \textsc{cont} & \boxed{2}\end{array}\right]\right]\\ \textsc{out} & \left[\textsc{syns|loc}\ \left[\begin{array}{ll}\textsc{cat} & \left[\textsc{head|lid}\ \boxed{3}\right]\\ \textsc{cont} & \boxed{4}\end{array}\right]\right]\\ & \textsc{coll}\ \left\langle\left[\textsc{c-words}\ \left\langle \ldots, \boxed{5}, \ldots\right\rangle\right]\right\rangle\end{array}\right] \begin{array}{l}\& \boxed{1} \neq \boxed{3}\\ \& \boxed{2} \neq \boxed{4}\end{array}$$

Figure 4: The *i-word-lexical-rule*

References ◇ Abeillé, Anne. 1995. The flexibility of French idioms: A representation with Lexical MTree Adjoining Grammar. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder (eds.), *Idioms: Structural and psychological perspectives*, 15–42. Hillsdale: Lawrence Erlbaum. ◇ Bargmann, Sascha & Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In M. Sailer & S. Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 1–29. Berlin: LSP. ◇ Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press. ◇ Findlay, Jamie Y. 2019. *Multiword expressions and the lexicon*: University of Oxford dissertation. `http://users.ox.ac.uk/~sjoh2787/findlay-thesis.pdf`. ◇ Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language* 6(1). 22–42. ◇ Gazdar, Gerald, Ewan Klein, Geoffrey Pullum & Ivan Sag. 1985. *Generalized phrase structure grammar*. Cambridge: 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 J. Nerbonne, K. Netter & C. Pollard (eds.), *German in HPSG*, 365–396. Stanford: CSLI Publications. ◇ McCawley, James D. 1981. The syntax and semantics of English relative clauses. *Lingua* 53. 99–149. ◇ Meurers, Walt Detmar. 2001. On expressing lexical generalizations in HPSG. *Nordic Journal of Linguistics* 24(2). 161–217. doi:10.1080/033258601753358605. ◇ Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70. 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 S. Müller, A. Abeillé, R. D. Borsley & J.-P. Koenig (eds.), *HPSG: The handbook*, Berlin: LSP. `https://hpsg.hu-berlin.de/Projects/HPSG-handbook/PDFs/formal-background.pdf`. ◇ Riehemann, Susanne Z. 2001. *A constructional approach to idioms and word formation*: Stanford University dissertation. ◇ Sailer, Manfred. 2003. Combinatorial semantics and idiomatic expressions in HPSG. `https://publikationen.uni-tuebingen.de/xmlui/handle/10900/46191`. ◇ 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 M. Krifka, R. Ludwig & M. Schenner (eds.), *Reconstruction effects in relative clauses*, 225–262. Berlin: de Gruyter.

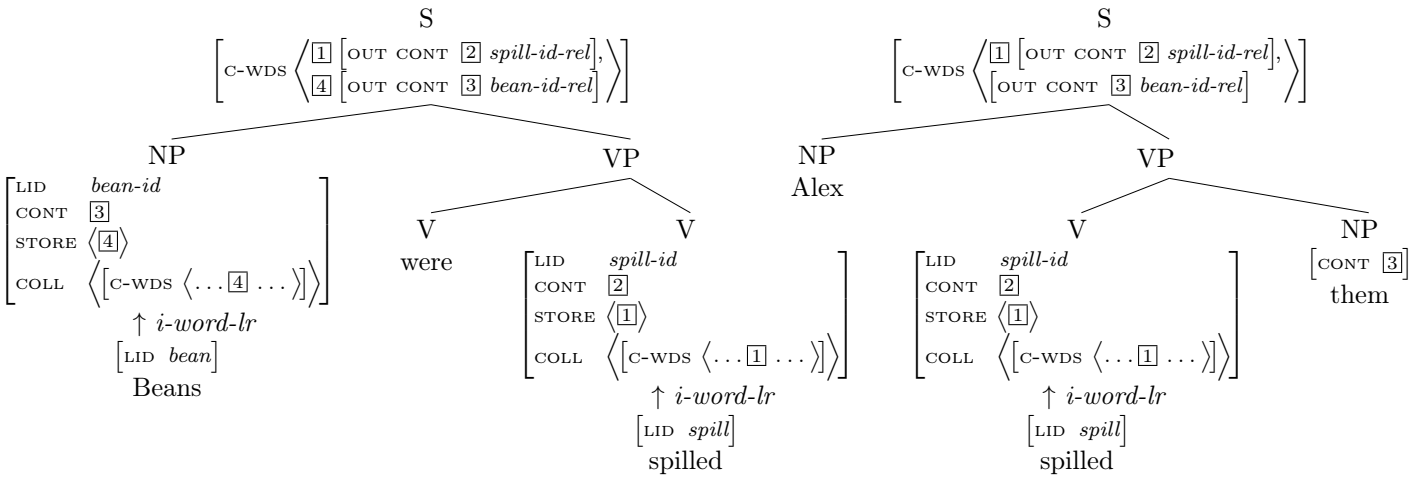

Figure 5: Sketch of the analyses of *Beans were spilled* and *Alex spilled them.*