Phraseological Clauses as Constructions in HPSG

Manfred Sailer
Seminar für Englische Philologie
Universität Göttingen
manfred.sailer@phil.uni-goettingen.de

Frank Richter
IMS, Universität Stuttgart &
SfS, Universität Tübingen
fr@sfs.uni-tuebingen.de

1 Introduction

The literature on idioms often focuses on VP idioms such as *kick the bucket* or *spill the beans*, where a particular verbal lexeme combines with a particular NP or PP complement. These combinations show different degrees of flexibility. Hardly any attention has been paid to idioms which comprise complete clauses. Idioms with phraseological clauses are mentioned in passim in phraseological studies such as Fleischer (1997) but have never been in the focus of empirical studies, or detailed theoretical discussions. As clausal parts of idioms are structurally more complex than NPs or PPs, they are ideally suited for investigating a greater range of structural and semantic variation in idiomatic expressions.

In this paper we will look at phraseologically fixed clauses (PCl) in German. The discussion of PCls is particularly interesting in light of attempts to combine aspects of Construction Grammar with HPSG. One of the important insights of Construction Grammar is that constructions may span more than a local tree. This contrasts with the lexical nature of HPSG and its historical ties to context-free phrase structure grammars. There have been two types of proposals to analyze constructions in HPSG. The first one models complex constructions in terms of specialized phrases that are specified in phrasal lexical entries. Since phrases in HPSG include the syntactic structure that they dominate, a description of these phrases may refer to embedded structural properties. Proposals of this sort have been made, among others, by Riehemann (2001) and Soehn (2004). The second type revives the program of Gazdar et al. (1985), in which the description of a phrase may only span a local tree and all aspects of nonlocality are expressed with special feature percolation mechanisms. Such an approach has been favored in *Sign-Based Construction Grammar* (SBCG, Sag (2007a)).

In Section 2 we introduce properties of German phraseological clauses. We show how they can be captured with *Phrasal Lexical Entries* in Section 3 and discuss what an SBCG account of the data may look like in Section 4. There is a brief conclusion in Section 5.

2 Data

In (1) and (2) we list idioms with phraseological clauses (PCl). In (1) the PCl combines with a particular verb or a small group of verbs. These PCls behave similarly to *headway* in *make headway*, i.e. they act as a complement in a VP idiom where both the verb and the complement are part of the idiom. The PCls are declarative clauses ((1-c), (1-g)), interrogative clauses ((1-a), (1-b), (1-d), (1-f)), and a free relative clause in (1-e). The PCls in (2) are adjunct clauses.

(1) PCl is a complement clause to one or a small group of verbs

a. wissen, wo Barthel den Most holt
know where Barthel the young wine gets ('know every trick in the book')

b. (nicht) wissen, wo X_dat der Kopf steht
not know where X the head stands ('have a lot of stress')

c. glauben, X_acc tritt ein Pferd
believe X kicks a horse ('be very surprised')

d. wissen, wo (X_acc/dat) der Schuh drückt
know where X the shoe hurts ('know what is worrying X')

e. hingehen/ bleiben (sollen), wo der Pfeffer wächst
go/ stay (should) where the pepper grows ('go/ stay away')

f. jdm zeigen, wo der Zimmermann das Loch in der Wand gelassen hat
s.o. show where the carpenter the hole in the wall left has
('send s.o. away')

g. glauben, X's Schwein pfeift
believe X's pig whistles ('be very surprised')

(2) PCl is an adjunct

a. bis der Arzt kommt
until the doctor arrives ('ad nauseam')

b. wenn Ostern und Pfingsten auf einen/ denselben Tag fallen
when Eastern and Pentecost on one/ the same day fall ('never')

c. aussehen, als hätten X_dat die Hühner das Brot weggefressen
look as if had X the chicken the bread eaten away ('look stupified')

d. wie Gott X_acc geschaffen hat
as god X created has ('naked')

Apart from their idiomatic semantics, the PCls are regular sentences of German. They display an interesting continuum of grammatical and lexical fixedness and flexibility.

In (1-b), (1-c), (1-g), (2-c), and (2-d) the constituent marked with X is coreferential with the matrix subject. In (1-d) the constituent marked with X is optional and need not refer to the matrix subject.

(3) Ich möchte wissen, wo (dich) der Schuh drückt.
(lit.: I want to know where the shoe hurts you)

PCls permit a certain degree of grammatical variation. In German, speakers of some dialects prefer to use proper nouns with definite articles. This is reflected in (1-a), which comes in a variant with *der Barthel* (*the Barthel*). Similarly, *until*-clauses in German may optionally contain an overt complementizer *dass* (*that*). Indeed, a variant of (2-a) with an overt complementizer is attested, i.e. *bis dass der Arzt kommt* (*until that the doctor arrives*).

However, not just any grammatical variation is permitted. Let us consider the idiom in (1-b). Outside of idiomatic phrases a combination of a possessive dative NP and a definite NP can be freely replaced with a construction with the same dative NP and a definite NP that contains a possessive determiner. The possessor is then coreferential with the dative NP. The pattern is illustrated in (4-a). This otherwise systematic variation is not possible with the idiom. We use "#" to indicate the non-availability of an idiomatic interpretation. The same alternation is also excluded for (2-c).

(4) a. Ich habe Peter den/seinen Kopf verbunden. (lit: 'I bandaged Peter the/his head')
b. Peter weiß nicht, wo ihm der/#sein Kopf steht.

Another systematic variation is the active-passive alternation. None of the PCls with a transitive verb in (1) allow a passive in their idiomatic meaning.

(5) a. #wissen, wo vom Barthel der Most geholt wird. (passive of (1-a))
b. #wissen, wo X vom Schuh gedrückt wird. (passive of (1-d))
c. #glauben, X wird von einem Pferd getreten (passive of (1-c))

d. #jdm zeigen, wo vom Zimmermann das Loch in der Wand gelassen wurde. (passive of (1-f))

Finally, the PCl in (1-c) is a verb-second clause. In free uses, we can find two kinds of alternation. First, verb-second complement clauses alternate with verb-final complement clauses. Second, any constituent of the clause can occur as the first constituent in verb-second clauses, the so-called *Vorfeld*, without a change in meaning. Both types of grammatical alternation are excluded in (1-c).

(6) a. #Ich glaube, dass mich ein Pferd tritt. (*dass*-clause)
b. #Ich glaube, ein Pferd tritt mich. (different first constituent)

There is also some lexical variation. In (2-b) the holidays can be changed, i.e. the subject may be any combination of Easter, Pentecost, and Christmas. Some form of the verb *(zusammen-) fallen* is obligatory.

(7) #wenn Ostern und Pfingsten auf demselben Tag liegen/ zu liegen kommen/ am selben Tag sind.

In addition to the variation of obligatory material, some PCls may host more lexical and/or semantic material. For example, there is variation in the tense form of some but not all PCls.

(8) temporally flexible idioms
a. Ich hab damals Tetris gespielt, bis der Arzt gekommen ist.
(pres. perfect in (2-a); from www)
b. Er wusste nicht, wo ihm der Kopf stand. (simple past in (1-b))

(9) temporally fixed idioms
a. #Sie hat nicht gewusst, wo Barthel den Most geholt hat. (pres. perf. in (1-a))
b. #Ich glaube, mein Schwein hat gepfiffen. (pres. perf. in (1-g))

Similarly, modals are allowed in some but not all of the PCls.

(10) modally flexible idioms
a. Hudezeck versteht sich auf die Kunst, die Lachmuskeln so zu strapazieren, bis der Arzt kommen muss. (additional *must* in (2-a), from www)
b. Als Reiseleiter ist Terje ein Mann der Praxis und weiß, wann und wo auf Reisen der Schuh drücken könnte. (additional *could* in (1-d), from www)

(11) modally fixed idioms
a. #Peter soll bleiben, wo der Pfeffer wachsen kann. (additional *can* in (1-e))
b. #Ich glaube, mein Schwein könnte pfeifen. (additional *could* in (1-g))

The PCls do not permit negation (see (12)), but non-truth-conditional modifiers such as *eigentlich* (*actually*), *sprichwörtlich* (*proverbial*) can usually be added (see (13)).

(12) a. #Peter weiß, wo ihn der Schuh nicht drückt. (negation in (1-d))
b. #Peter weiß, wo Barthel den Most nicht holt. (negation in (1-a))
c. #wenn Ostern und Pfingsten nicht auf einen Tag fallen (negation in (2-b))

(13) a. Peter weiß nicht mehr, wo ihm eigentlich der Kopf steht. (*actually* in (1-b))
b. Martha weiß, wo Barthel den sprichwörtlichen Most holt. (*proverbial* in (1-a))

The properties of PCls show that they cannot be treated as big "words with spaces". Instead, they are inherently complex syntactic units with different degrees of flexibility. This is parallel to what was observed for other idioms in Wasow et al. (1983) and elsewhere, and clearly sets PCls apart from fully fixed forms such as proverbs.

3 Modeling Phraseological Clauses as Phrasal Lexical Entries

In this section we will sketch an analysis of PCls in terms of *phrasal lexical entries* (PLE), as used in Sailer (2003) and Soehn (2004). In these approaches, a feature, COLL, is used to mark idiosyncratic phrases. This attribute also plays an important role in other types of idioms in these theories, and it may contain complex structures. In the present paper it is enough to assume a binary distinction: [COLL +] stands for idiomatic phrases and [COLL −] for regular, non-idiomatic phrases. The mechanisms of regular syntactic and semantic combinatorics apply to phrases marked as [COLL −], whereas those marked as [COLL +] are exempt from them.

In (14) we sketch the PLE for the idiom in (2-a). We use the head feature LISTEME to identify individual lexical elements. This feature is taken from Soehn (2004) — it directly corresponds to the feature LEXICAL-ID (LID) in Sag (2007b). The PLE specifies that the overall clause is a modifier with the semantics *ad nauseam*. The phrase is a head-complement combination, where the head daughter is the preposition *bis* (*until*). The nonhead daughter is a finite clause. Inside the complement, there must be a verbal word with the LISTEME value *kommen* whose subject is a definite singular NP with the word *Arzt* as its lexical head.

(14) Sketch of the phrasal lexical entry of *bis der Arzt kommt*:

$$\left[\begin{array}{l}\textit{phrase}\\ \textsc{syns loc}\ \left[\begin{array}{l}\textsc{cat head mod}\ \left[\textsc{loc cont}\ \boxed{1}\right]\\ \textsc{cont}\ \left[\begin{array}{l}\textit{ad-nauseam-rel}\\ \textsc{pos-arg}\ \boxed{1}\end{array}\right]\end{array}\right]\\ \textsc{dtrs}\ \left[\begin{array}{l}\textit{head-complement-struc}\\ \textsc{head-dtr}\ \left[\begin{array}{l}\textsc{syns loc cat}\ \left[\begin{array}{l}\textsc{head}\ \left[\begin{array}{l}\textit{prep}\\ \textsc{listeme}\ \textit{bis}\end{array}\right]\\ \textsc{val comps}\ \left\langle \boxed{2}\left[\textsc{loc cat head}\ \left[\begin{array}{l}\textit{verb}\\ \textsc{vform}\ \textit{fin}\end{array}\right]\right]\right\rangle\end{array}\right]\end{array}\right]\\ \textsc{nonh-dtr}\ \left[\begin{array}{l}\textsc{syns}\ \boxed{2}\\ \textsc{dtrs}\ \left[\ldots\textsc{head-dtr}\ \boxed{3}\left[\begin{array}{l}\textit{word}\\ \textsc{syns loc cat}\ \left[\begin{array}{l}\textsc{head listeme}\ \textit{kommen}\\ \textsc{val subj}\ \left\langle \textsc{np}\left[\textsc{listeme}\ \textit{arzt},\ \textsc{def}\ +,\ \textit{sg}\right]\right\rangle\end{array}\right]\end{array}\right]\end{array}\right]\end{array}\right]\\ \textsc{coll}\ +\end{array}\right]$$

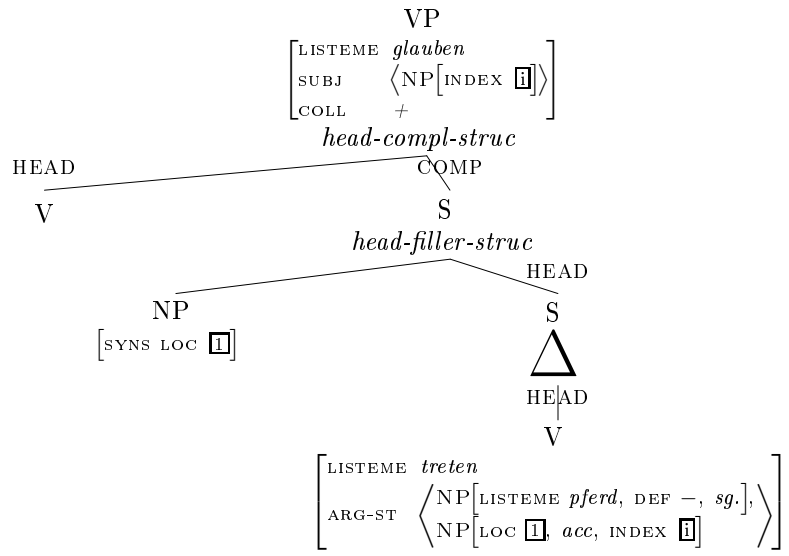
We saw in the data section that there are some restrictions on the structure of the PCl: While tense and modality may vary, negation is not permitted. This can be expressed in (14) by requiring that there is no negation in the content of the PCl. For other PCls we must also ban modal operators from the semantics of the PCl. Since modal content can be contributed by modal verbs or adverbials, the restriction must be imposed on the operators that may occur in the CONTENT value of the PCl.

To exclude valence alternations such as passive and the dative-possessive alternation in (1-a) or (1-b) we impose a syntactic restriction on the ARG-ST or the valence value of the verb *kommen*.

Let us now turn to a more complicated example. The PCl in (1-c) requires coreference between the matrix subject and the accusative argument in the PCl. Simplifying the idiom, let us assume that the matrix verb *glauben* is part of the phrasal construction.[1] In Fig. 1 we sketch the PLE for this idiom in the form of a tree. The tree is intended as an abbreviatory notation for the specifications of the daughter attributes in the PLE. In the PLE we specify that the matrix verb is *glauben*. The complement clause is an embedded verb-second clause. The lexical head of the complement clause

[1] In (1) the combination of the verb and the PCl behaves like in decomposable idioms of the type *spill the beans*. According to the theory in Soehn (2004), this means that the matrix verb selects for a complement with a particular LISTEME value. The complement specifies inside its COLL value that it needs to combine with a particular verb. The COLL value of the complement contains the relevant information about the verb. In Soehn (2004) the verb's LOCAL value is the relevant structure. What is important for our purposes is that information about the matrix verb is available in the formulation of the PLE. This condition is met, both in our simplification and in Soehn's original theory.

Figure 1: Sketch of the PLE for the idiom *glauben, X_ acc tritt ein Pferd*:



is the verb *treten*. This verb must take two arguments. The first one is an indefinite NP headed by *Pferd*. The second one is an accusative NP whose INDEX value is identical with that of the matrix subject, 1. This NP occurs as the first constituent in the complement clause.

The PLE in Fig. 1 is such that it excludes passive alternation (see (5-c)) because it specifies that the verb *treten* occurs with a transitive argument structure. It also requires that the PCl be a verb-second clause (see (6-a)) by specifying that it is a head-filler structure, and it determines the first constituent (see (6-b)) by specifying it in the PLE. The coreference between the embedded accusative and the matrix subject is also encoded directly.

4 Modelability under Strict Locality Assumptions?

Our discussion in Section 3 shows that existing accounts in HPSG are capable of capturing the properties of PCls. An important ingredient in this account is the fact that we can refer to deeply embedded parts of a phrase in its PLE. This makes HPSG especially well-suited to integrate a fundamental insight of Construction Grammar: Constructions can span more than a local tree (Fillmore et al., 1988; Jackendoff, 1995).

In this section we turn to the second approach to construction-like phenomena in HPSG and consider a possible alternative to our analysis of PCls. In a recent series of papers (Sag (2007a,b) and others) it was shown that various phenomena of apparent non-locality can be encoded using an extension of HPSG's feature geometry and a restructuring of signs. In the framework proposed there, *Sign-Based Construction Grammar* (SBCG), phrasal signs no longer contain their daughters. Instead, *construct* objects are introduced that correspond to local trees. Signs only occur as nodes in these constructions. The analysis of a sentence is a set of constructions, each of which represents a local tree, but these trees do not form a single joined feature structure. With this change the formulation of PLEs as those in (14) and Fig. 1 is not possible.

To account for non-locality SBCG uses two head features, the listeme attribute LEXICAL-ID and the attribute XARG whose value is the subject of the sentence. These two attributes are sufficient to describe the construction in (2-a), because the obligatory elements in the embedded clause are the lexical head *kommen* and the subject, *Arzt*, i.e. exactly those parts that are locally available for the

overall construction.

(15) A SBCG description of *bis der Arzt kommt*:

$$\begin{bmatrix} \textit{bis-der-arzt-kommt-cxt} & \\ \textsc{mother} & \begin{bmatrix} \textsc{mod} & \boxed{1} \\ \textsc{sem} & \textit{ad-nauseam}(\boxed{1}) \end{bmatrix} \\ \textsc{dtrs} & \left\langle \begin{bmatrix} \textsc{lid} & \textit{bis} \end{bmatrix}, \text{S} \begin{bmatrix} \textsc{xarg} & \begin{bmatrix} \textsc{lid} & \textit{arzt} \end{bmatrix} \\ \textsc{lid} & \textit{kommen} \end{bmatrix} \right\rangle \end{bmatrix}$$

To allow for modal verbs and temporal auxiliaries we can simply assume that the LID value of a verbal complex is identical with that of the lowest lexical verb in the verbal complex. If we want to exclude modal and temporal variation, we can impose the same kind of restrictions as in Section 3, i.e. we can determine which operators may occur in the contents of the daughters.

We saw that semantically neutral, grammatical variation occurs in some but not all PCls. In the PLE account we could refer to the ARG-ST value of an embedded verb to exclude passive or other valence alternations. Since SBCG allows reference to the highest subject in a PCl, active-passive alternations can be excluded by requiring a particular LID value inside the XARG value. Alternations that do not involve the subject are harder to capture. If the dative-possessive alternation is modeled by a lexical rule, we can exclude its occurrence in (1-b) by introducing a special LID value for the verb *stehen* as it occurs in this PCl. Let us call this LID value *kopf-stehen-lid*. We must, then, guarantee that a verb with the LID value *kopf-stehen-lid* cannot serve as the input to the relevant lexical rule.

Once such a special LID value is introduced, we must make sure that a verb with this value only occurs inside the PCl in (1-b). This problem is reminiscent of the restricted occurrence of bound words in idioms, such as *headway*. In unpublished work, Ivan Sag proposes an SBCG account of bound words. He assumes a default specification of LID values. The LID value of the noun *headway* is inconsistent by default, which prevents this noun from occurring freely. The marked, or non-default value is *headway*. If a verb selects an NP with LID value *headway*, the default can be overridden and the noun can occur. It seems natural to apply this analysis to our hypothetical verb with LID value *kopf-stehen-lid* as well. What complicates matters here is that modal verbs may occur inside the PCl. Modal verbs do not explicitly select for the special LID value. Consequently, the default mechanism must extended in such a way as to ignore the modal heads.

A similar problem occurs in (1-e). Here the PCl is a free relative clause. If we analyze it as an AdvP the LID value of the embedded verb, *wachsen* (*grow*), is probably not identical with that of the free relative. This makes it necessary to introduce special LID values for the relative phrase *wo* and for the embedded verb *wachsen*. Again, the default mechanism must be used to prevent these special words from occurring outside the PCl. In a PLE account we have direct access to the embedded verb and do not need a special relative phrase or a special LISTEME value for *wachsen* for this PCl.

Let us finally turn to a property of (1-c) that cannot be captured under SBCG's locality assumptions, the requirement that the accusative NP inside the PCl be bound by the matrix subject. The accusative object is on the ARG-ST list of the embedded verb. The matrix subject is on the ARG-ST list of the matrix verb, the matrix verb has access to the LID value of the embedded verb and to its XARG value. However, neither of these can be used to establish a link between the embedded accusative NP and the matrix subject. The same problem occurs in all other cases where the PCl contains an embedded open slot that must corefer with the matrix subject.

5 Conclusion

We discussed properties of German PCls and showed that they can be expressed in HPSG with phrasal lexical entries. These idioms provide support for the basic claim of construction grammar that constructions can span more than a local tree. We investigated whether a strictly local analysis of PCls is possible, as required in SBCG. We saw that some properties can be handled the same way

as in a PLE analysis, some require certain, possibly undesirable assumptions, and one group cannot be handled at all.

It should be noted that the PLE account is not as nonlocal as it seems at first sight. A PLE does two things: First, an idiosyncratic semantic and/or syntactic combination is licensed in a local tree. Second, properties can be imposed on constituents that are embedded inside this combination. It is crucial that the embedded constituents must be independently well-formed. This means that we can restrict which of the well-formed signs may occur there, but a PLE cannot license embedded, idiosyncratically structured signs. In this respect, the PLE account is more local than a classical Construction Grammar analysis.

References

Fillmore, Charles, Kay, Paul, and O'Connor, M. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*. *Language 64*, 501–538.

Fleischer, Wolfgang (1997). *Phraseologie der deutschen Gegenwartssprache* (2nd, revised edition). Niemeyer, Tübingen.

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

Jackendoff, Ray (1995). The Boundaries of the Lexicon. In M. Everaert, E.-J. v. d. Linden, A. Schenk, and R. Schreuder (Eds.), *Idioms. Structural and Psychological Perspectives*, pp. 133–165. Lawrence Erlbaum Associates, Hillsdale.

Riehemann, Susanne Z. (2001). *A Constructional Approach to Idioms and Word Formation*. Ph. D. thesis, Stanford University.

Sag, Ivan A. (2007a). Remarks on Locality. In S. Müller (Ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, Stanford, 2007*, Stanford, pp. 394–414. CSLI Publications.

Sag, Ivan A. (2007b, August). Sign-Based Construction Grammar An informal synopsis. Manuscript, Stanford.

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

Soehn, Jan-Philipp (2004). License to COLL. In S. Müller (Ed.), *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pp. 261–273. Stanford: CSLI Publications.

Wasow, Thomas, Sag, Ivan A., and Nunberg, Geoffrey (1983). Idioms: An Interim Report. In S. Hattori and K. Inoue (Eds.), *Proceedings of the XIIIth International Congress of Linguists*, pp. 102–115.