Cranberry Words in Formal Grammar

Frank Richter
Seminar für Sprachwissenschaft
Wilhelmstr. 113
72074 Tübingen

Manfred Sailer
Sonderforschungsbereich 441
Nauklerstr. 35
72074 Tübingen

April 10, 2002

Abstract

We will provide a formal account of cranberry words in German within the framework of HPSG. Cranberry words only occur in a very restricted set of linguistic environments but behave as any other word would in these environments. We will show that the distribution of these words can neither be reduced to constructions nor to selection. Therefore, a formal theory of distribution and collocations will be necessary. We will begin work on developing such a theory.

1 Introduction

In this paper[1] we will motivate the integration of a collocation module into the architecture of formal grammar. *Cranberry words* (CW, also called *unique lexemes* or *hapax legomena*), appear to be particularly well-suited for this purpose. In Aronoff (1976:15) they are characterized as follows:

(1) "There are words which, like cranberry morphs, concatenate only with specific words and not with syntactic classes. For example the noun *headway* occurs only as the direct object of the verb *make*, just as *cran* occurs only in *cranberry*."

We will show that CWs have all the properties of "normal" words except for their strict distributional requirements. In particular, we will show that it is inadequate to treat expressions which contain CWs as idiomatic or as fixed constructions.

Reflections on CWs allow us to contrast three different architectures of grammar: First generative theories, including Head-Driven Phrase Structure Grammar (HPSG, Pollard and Sag (1994), henceforth PS94); second constructional theories such as Construction Grammar (see Fillmore et al. (1988); Kay (1997); Kay and Fillmore (1997)), Tree Adjoining Grammar (Joshi (1987); Abeillé (1995)) or constructional HPSG (see Sag (1997); Riehemann (2001)); and third collocation theory as originated in the work of J.R. Firth and exemplified in Butler (1985) or Sinclair (1991).

In the Chomskian paradigm and in HPSG there are two basic assumptions that we will address: First, the structure of a complex phrase is built out of smaller phrases or words

[1]The research to the paper was funded by the *Deutsche Forschungsgemeinschaft* in the *Special Research Program (SFB) 441*. We thank Lothar Lemnitzer, Jan-Philipp Söhn, Ton van der Wouden, and the audience of CSSP '01 for comments, Janina Radó and Carmella Payne for help with English, and Olivier Bonami for advise and formatting our paper.

according to general structure formation rules. We call this assumption the *structural regularity assumption* (SRA). Second, the distribution of words is fully determined by their syntactic category, their semantics and their selectional properties. We call this the *distributional regularity assumption* (DRA). These two assumptions are important components of HPSG in PS94.

Each of the other two frameworks questions a different assumption: The constructional theories argue against the SRA with the observation that there is a continuum ranging from highly idiosyncratic syntactic and semantic combinations to the very general abstract rules assumed for example as IMMEDIATE DOMINANCE SCHEMATA in PS94. In the literature on idiomatic expressions (for example Nunberg et al. (1994)), we find many instances of highly idiosyncratic syntactically complex phrases, such as the expression *trip the light fantastic* (*dance nimbly*). These show that the SRA cannot be correct.

Collocational approaches challenge the second of the basic assumptions of generative theories, the DRA. To take a representative quotation, Butler (1985:130) states:

(2) "The defining feature of a lexical item, by which such an item is recognized, is its pattern of co-occurrence with other items, that is its *collocational* behaviour. A lexical item is recognized as different from other lexical items because its total pattern of collocation is unique."

This collocational behavior is different from semantic or syntactic selection; if it weren't, it would not be possible to particularize lexical elements. In the collocation literature, a distinction between *grammar* and *lexis* is made, where the latter is supposed to encode the collocational behavior of lexical elements (Halliday (1966), Sinclair (1966)).

The architecture of grammar that we will argue for, combines insights from these three systems, starting with HPSG as formalized in Relational Speciate Re-entrant Language (RSRL, Richter (2000)). We will assume that lexical elements need not be words, but can be signs of any linguistic level, i.e., morphemes, words, phrases or entire utterances. Thus, the distinction between words and phrases is not sufficient to identify whether or not a sign is lexical. As the core of our analysis, we will introduce a new attribute, COLL, which differentiates between lexical and non-lexical signs and whose value determines the distributional (collocational) restrictions in the case of a lexical sign.

We will first illustrate how idiosyncratic constructions which are counterexamples to the SRA, can be accounted for in HPSG (Section 2). In Section 3 we will show that it would be inadequate to treat many expressions that contain CWs as constructions. Moreover, these expressions cannot be captured by selection (Section 4). In Section 5 we will present a distributional analysis of CWs. In the concluding section we will discuss the overall architecture of grammar resulting from our study.

2 Internally Irregular Phrases

The German idiomatic expression in (3) can be considered an instance of what we call an *internally irregular phrase* (IIP). We give three criteria of regularity which similar expressions fail to meet.[2]

(3) den Löffel abgeben
the spoon away.give
"to die"

First, the expression cannot undergo passivization (4) or similar syntactic operations. The "$" sign indicates that the idiomatic reading is not available for the sentence.

(4) $ Hier wurde der Löffel (von Fritz) abgegeben.
Here was the spoon by Fritz away.given

Second, it does not allow for internal modification (see Ernst (1981)), as shown in (5).

(5) $ Er gab den komischen Löffel ab.
He gave the funny spoon away

Third, the meaning of the VP cannot be derived in a compositional way.

In the light of these observations and similar data from English idiomatic expressions, it is clear that a formal grammar which attempts to account for this type of idiomatic expressions must also allow for *phrasal lexical entries* (PLE). This conclusion was already drawn in early generative approaches such as Weinreich (1969) and Katz (1973), in Wasow et al. (1983), in GPSG (Gazdar et al. (1985)) and in TAG (Abeillé (1995)).

To integrate PLEs into HPSG, we will make the following conservative changes: In addition to the attributes declared for the sort *sign* in PS94, we introduce a list-valued attribute, COLL.[3] We assume that for regular phrases, the COLL value is the empty list.

[2]The data are taken from Krenn and Erbach (1994:370f).

[3]The motivation of list values will be given in Section 5. For the time being, the value *empty-list* can be interpreted as the boolean value *minus* (for non-lexical signs), *non-empty-list* as *plus* (for lexical signs).

$$\begin{bmatrix} \textit{phrase} \\ \text{COLL} \ \textit{elist} \end{bmatrix} \Rightarrow \Big(\ \text{ID}_1 \ \textbf{or} \ \text{ID}_2 \ \textbf{or} \ \text{ID}_3 \ \textbf{or} \ \text{ID}_4 \ \textbf{or} \ \text{ID}_5 \ \textbf{or} \ \text{ID}_6 \ \Big)$$

Figure 1: The IMMEDIATE DOMINANCE PRINCIPLE

$$\begin{bmatrix} \textit{sign} \\ \text{COLL} \ \textit{nelist} \end{bmatrix} \Rightarrow \begin{pmatrix} \text{LE}_1 \ \textbf{or} \ldots \textbf{or} \ \text{LE}_n \\ \textbf{or} \ \text{PLE}_1 \ \textbf{or} \ldots \textbf{or} \ \text{PLE}_m \end{pmatrix}$$

Figure 2: The LEXICON PRINCIPLE:

Consequently, we modify all principles of syntactic and semantic combination in such a way that they only apply to phrases that have an empty COLL value. For illustration, see the new version of the IMMEDIATE DOMINANCE PRINCIPLE in Figure 1. Recall that the consequent is a disjunction of the immediate domminance schemata (ID_i) which determine the possible ways of syntactic combination.

Lexical elements, which for us are words and irregular phrases, have a non-empty COLL value. The LEXICON PRINCIPLE, then, contains a lexical entry for all elements that have a non-trivial COLL value. It is given schematically in Figure 2, where we have written LE_i for lexical entries of words and PLE_i for phrasal lexical entries.

The PLE for the expression in (3) is given in Figure 3. We can take the term $\lambda x.\mathsf{die}'(x)$ to be the CONTENT value of the VP.[4] With its non-empty COLL value, the VP is exempt from the regular principle of semantic combination. Therefore the semantics of its immediate constituents need not contribute to its own semantics.[5]

The PLE in Figure 3 specifies the constituent structure of the VP. In particular it mentions that the direct object *den Löffel* (*the spoon*) must occur as a sister to the verb. This specification excludes the possibility of passive formation.[6] Similarly, by detailed specification of the properties of the direct object NP, internal modification can be excluded.

[4]Note that in contrast to the semantics of PS94, we use a standard semantic representation language in the CONTENT value. See Sailer (2000, Part I) for the necessary definitions and proofs. We assume that the SEMANTICS PRINCIPLE states that for every regular phrase (phrases with an empty COLL value), the CONTENT value of the phrase is the functional application of the CONTENT values of its daughters.

[5]This is different from Constructional HPSG (Riehemann (2001)), where a construction can add idiosyncratic parts to the semantics but cannot "remove" or "overwrite" the semantics of its immediate constituents.

[6]We leave aside the issue of fronting in verb second clauses.

$$\begin{bmatrix} \textit{phrase} & \\ \text{PHON} & \boxed{2} \oplus \boxed{3} \\ \text{SYNSEM} & \begin{bmatrix}\text{LOC} & \begin{bmatrix}\text{CAT ARG-ST} & \langle \boxed{1} \rangle \\ \text{CONT} & \lambda x.\mathsf{die}'(x)\end{bmatrix}\end{bmatrix} \\ \text{DTRS} & \begin{bmatrix} \text{H-DTR} & \begin{bmatrix}\text{PHON} & \boxed{3} \\ \text{SYNS} & \begin{bmatrix}\text{LOC} & \begin{bmatrix}\text{CAT ARG-ST} & \langle\boxed{1}\rangle \\ \text{CONT} & \lambda y \lambda x.\mathsf{pass}'(x,y)\end{bmatrix}\end{bmatrix}\end{bmatrix} \\ \text{N-DTR} & \begin{bmatrix}\text{PHON} & \boxed{2}\ \langle \textit{den, Löffel}\rangle\end{bmatrix}\end{bmatrix} \\ \text{COLL} & \textit{non-empty-list}\end{bmatrix}$$

Figure 3: Sketch of the PLE for the expression *den Löffel abgeben*

3 Expressions with Cranberry Words

In (6) we give some German expressions that contain CWs. For more examples see Dobrovol'skij (1988) and Fleischer (1989, 1997). We will underline CWs throughout the paper

(6) a. die Nase rümpfen
the nose wrinkle
"wrinkle one's nose"

b. jemandem Angst einjagen
someone fear in.chase
"frighten someone"

c. kein Hehl aus etwas machen
no secret of something make
"not to make a secret of sth."

The expressions in (6) contrast sharply with the IIP in (3) as they pass all the regularity tests. Passivization is possible (7) as well as internal modification (8).

(7) a. Bei dem Gestank wird schon mal die Nase gerümpft.
At this stench is PARTICLE the nose wrinkled
"People wrinkle their noses with this stench."

b. Ihm wurde gründlich Angst eingejagt.
him was deeply fear in.chased
"He was deeply frightened."

c. Daraus wurde bei der Sitzung kein Hehl gemacht.
thereof was at the meeting no secret made
"No secret was made of this at the meeting."

(8) a. Er rümpfte seine große Nase.
he wrinkled his big nose

b. Das musste ihm ja große Angst einjagen.
this had to him big fear in.chase
"This had to cause him much fear."

c. Er machte kein großes Hehl aus seiner Abneigung.
he made no big secret of his dislike

If we assign the CWs a specific meaning, the overall expressions can be interpreted in a compositional way. For the verb *rümpfen* in (6a) we assume the meaning *wrinkle*. The verb *einjagen* in (6b) is a light verb. The noun *Hehl* in (6c) means *secret*. If we substitute the CWs in (6) by free words with the corresponding meaning, we get synonymous VPs.

(9) a. die Nase hochziehen
the nose pull.up

b. jemandem Angst machen
somebody fear make

c. kein Geheimnis aus etwas machen
no secret of something make

From these observations we conclude that the expressions with CWs given in (6) behave just like free combinations if we assign the CWs the proper syntactic category and meaning. In this respect, they clearly differ from IIPs. This shows that a constructional analysis in terms of a PLE is not adequate for these expressions.[7]

4 Beyond Selection

We argued that the expressions in (6) should not be analyzed as IIPs. If, on the other hand, we consider the VPs in (6) to be the result of free combination, we would incorrectly predict the occurrences of the CWs in the examples in (10).

(10) a. den Gesichtserker/ das Riechorgan hochziehen/ *rümpfen
the face.bay-window/ the olfactory organ up.pull/ wrinkle
[*Gesichtserker* is a conventionalized metaphor for *nose* in German]

[7]Notice, though, that there are expressions with CWs that require a constructional analysis: the verb *balbieren* only occurs in the expression *jemanden über den Löffel balbieren* (*somebody over the spoon V*), which means *to swindle somebody*. Semantically this expression is fully opaque, just as is the case with (3).

b. jemandem Bammel/ Muffe machen/ *<u>einjagen</u>
someone fear/ fear make/ chase.in
[*Bammel* and *Muffe* are colloquial synonyms of *Angst*]

c. kein Geheimnis/ *<u>Hehl</u> für sich behalten können
no secret/ secret for oneself keep can

The question then is how the distributional restrictions of the CWs can be expressed. The natural place to state distributional properties of a lexical element are its selectional requirements. For our data they cannot be stated as such.

According to standard assumptions, selection by a lexical item may refer to semantic properties and the syntactic category, as expressed in *synsem* objects in HPSG, but not to phonological properties or to the constituent structure of the selected element. These *synsem* objects occur as the SYNSEM value of linguistic signs and as elements of the list values of the valence attributes.

To account for the distribution of the verb *rümpfen*, we would have to refer to the phonology of its theme argument, or more specifically to the phonology of the syntactic head of its theme argument. This clearly violates the usual restrictions on selection.

Alternatively, and probably more adequately, one could assume lexeme specific selection. This would imply that the information about the head lexeme of a phrase has to percolate from the lexical head to the SYNSEM value of the saturated phrase that it heads. Given the principles of HPSG this is only possible if the lexeme can be identified either via its HEAD value of via its INDEX value, as only these two values are guaranteed to be shared between a phrase and its lexical head.

Krenn and Erbach (1994) opt for the latter. They introduce an attribute, LEXEME, on the sort *index*, whose value is a unique identifier for each lexeme. This is mainly motivated by data from pronominalization and relative clauses where only the index is shared between the element on the SUBCAT list of some verb that is part of an idiomatic expression (e.g. *sie*$_i$ and *die*$_i$ in (11)) and the overt realization of another part of that expression (here *Nase*$_i$).

(11) a. Peters Nase$_i$ sah ja normalerweise schon ziemlich drollig aus, aber wenn er
Peters nose looked normally already quite funny out but when he
sie$_i$ <u>rümpfte</u>, musste man einfach loslachen.
it wrinkled must one simply start laughing
"Normally Peter's nose already looked quite funny, but when he wrinkled it, one simply had to start laughing."

b. Jetzt hat er eins auf die Nase$_i$ bekommen, die$_i$ er einmal zu oft gerümpft hat.
now has he one on the nose got that he once too often wrinkled has
"Now he got a blow on his nose, which he wrinkled once too often."

In (11a) the direct object of *rümpfen* is a personal pronoun which is co-indexed with an NP whose lexical head is *Nase*. In HPSG such a co-indexation is encoded as identity of INDEX values. In (11b) the relative clause modifies the noun *Nase* and the relative pronoun is co-indexed with this noun. Again, only the INDEX values are shared.

There are at least two problems with this proposal. First, it is conceptually dubious that pronouns should not have a proper LEXEME value, because they have their own lexical entries. Put differently, we would need a theory of lexemes to tell us which lexical elements have their own LEXEME value and which ones do not.

Second, selection of the LEXEME value does not cover all the relevant cases. In PS94 adjectives share their INDEX value with the noun they modify. Thus, in (12) the adjective *bekannte* (*known*) and the noun *Probleme* have the same LEXEME value, *problem-lex*.

(12) Ich will nicht mehr länger über <u>sattsam</u> bekannte$_i$ Probleme$_i$ diskutieren.
I want no more longer over ad nauseam known problems discuss
"I do not want to discuss problems known ad nauseam any longer."

The example contains an adverb, *sattsam*, a (pejorative) intensifier. It is a CW bound to modify the adjective *bekannt*, as evidenced by the ungrammatical combinations in (13).

(13) das sattsam bekannte$_i$/ *besprochene$_i$/ *beschriebene$_i$ Problem$_i$
the ad nauseam known/ discussed/ described problem

In the theory of selection of PS94 a modifier selects the modified head via some special selection feature, MOD. The value of MOD is a *synsem* object. In (12), the adverb *sattsam*, thus, selects the adjective *bekannt* via MOD. As *sattsam* is distributionally restricted to modify only this adjectival lexeme, it should be possible to identify it. But the LEXEME value of the adjective is *problem-lex*. Therefore lexeme selection cannot work here.[8]

To conclude, the stipulation of a LEXEME attribute as part of the INDEX value is not a general solution to the problem of selection of a particular lexeme. No local parts of a sign can be used to account for the data in (11) and (12). Instead, a larger structure needs to be available to check whether the distributional requirements of *rümpfen* are met.

[8]Kasper (1997) shows that the analysis of recursive modification in PS94 yields the wrong interpretation. While his proposal overcomes that problem, the identity of INDEX values is still as described above.

In the cases of *rümpfen* and *sattsam*, the CW is the syntactic selector. Matters get even more complicated if we consider CWs in selected positions. This is the case for *Hehl* in (6c). As we saw in (10c) the noun *Hehl* requires that it appear as the theme argument of a particular verb, the light verb *machen* (*make*). There are also examples in which a non-light verb is required:

(14) a. Tacheles reden/ *sprechen/ *erzählen
goal talk/ speak/ tell
"talk straight"

b. kein Sterbenswort sagen/ *flüstern
no dying word say/ whisper
"not to say a word"

In such cases, the head of the VP has its usual meaning, but the CW has to be selected by the verb. In HPSG there is no information of the head available on a complement. Thus, these cases cannot be expressed in terms of the selectional properties of the CW.

Finally, there are cases with no selectional connection at all between the CW and its required context. There are two frequent environments for this: intensifying comparative adverbs (15), and binomials (16).

(15) a. essen/ *futtern wie ein Scheunendrescher
eat/ tuck in like a barn.thresher
"eat like a horse"

b. aufpassen/ *hinschauen wie ein Schießhund
be alert/ watch like a shoot dog
"be on one's toes"

(16)	a. klipp und klar	b. kreuz und quer	c. frank und frei
	?? and clear	cross? and crosswise	frank and free
	"very clear"	"higgledy-piggledy"	"frankly"

In (15) the preposition *wie* (*like*) connects the CW and the required verb. In (16), one of the conjuncts is a CW. Semantically it can be considered a (near) synonym of the other conjunct, which explains the intensifying function of the binomials.[9] There is no reason to assume a mutual selection of the conjuncts; only the conjunction links them.

[9]The binominal in (16a) has a counterpart with no CW: *klar und deutlich* (*clear and clear*). There, the two words are near synonyms and the use of the binominal instead of one of its conjuncts has the same intensifying effect as (16a).

We have shown that there is no general way to account for the distributional restrictions of the CWs in terms of selection: (i) Even when the required context is selected by the CW, its characterization would need to be a lot more precise than what can be expressed within the boundaries of standard theories of selection; (ii) in many cases the CW is selected rather than being itself the selector; (iii) there are even cases in which there is no selectional relation between the CW and a required element of its distributional context.

5 Analysis of Cranberry Words

The expressions in (6) should not be analyzed as IIPs, nor can the distributional particularities of the CWs be expressed in terms of selection. This shows that neither a constructional approach nor the theoretical assumptions that PS94 shares with generative theories of grammar are sufficient to handle the phenomenon. This leads us to consider the third concept mentioned in Section 1, collocations.

The basic observation in collocation theory is that the word distribution in texts deviates from what we would expect to find if we only considered their syntactic category and their meaning. Such deviation from random distribution can be caused by personal preferences, stylistics and many other factors. All of these are explored within the collocational paradigm (Sinclair (1991), Dodd (2000)). Due to the diversity of the phenomena, collocation theory has not yet been integrated with formal theories of grammar.

In the case of CWs, it is clear that the collocational restrictions should be considered a lexical property of the CWs themselves. Furthermore, as their violation leads to ungrammaticality rather than to stylistic dispreferences, we have identified an empirical area which gives collocation theory and formal grammar the oportunity to merge.

In Section 2 we introduced a list-valued feature, COLL, which is appropriate for the sort *sign* and distinguishes lexical elements (words and IIPs). It will be crucial for our analysis of distributional idiosyncrasies at two levels: First, in the **lexicon**, the COLL specification in a lexical entry will indicate the occurrence restrictions of every sign that is licensed by the lexical entry. Second, we introduce a COLL PRINCIPLE that guarantees that these restrictions are actually met for each lexical element in a given **utterance**.

In Figure 4 we give a sketch of the lexical entry of the noun *Hehl*. We assume that the CONTENT value contains the semantic constant secret'. This indicates that the meaning

$$
\boxed{0}\left[\begin{array}{ll}
\textsc{phon}\ \langle \textit{Hehl} \rangle & \\
\textsc{syns loc} & \left[\begin{array}{ll} \textsc{cat} & \left[\textsc{head}\ \textit{noun}\right] \\ \textsc{cont} & \lambda x.\mathsf{secret}'(x) \end{array}\right] \\
\textsc{coll} & \left\langle \boxed{1} \right\rangle
\end{array}\right]
$$

and sign $\boxed{1}$ dominates $\boxed{0}$ and a sign $\boxed{2}$ such that

sign $\boxed{2}$ dominates $\boxed{0}$, and

the CONT value of $\boxed{2}$ is:

$$\ldots\neg[\ldots\exists x[\mathsf{secret}'(x) \wedge \ldots \text{‘}v \text{ turns } w \text{ into } x\text{’}\ldots]\ldots$$

Figure 4: Sketch of the lexical entry of the cranberry word *Hehl*

of *Hehl* is roughly the same as that of the English word *secret*.

Our sketch of the lexical entry comprises an informal description of the sign that appears in the COLL list of *Hehl*.[10] The lexical entry expresses the collocational restriction that the sign $\boxed{1}$ dominates the word *Hehl* (i.e., the sign $\boxed{0}$) and another sign, $\boxed{2}$. The sign $\boxed{2}$ furthermore has to dominate the word *Hehl* and have a CONTENT value of a particular shape, namely the term that corresponds to the logical form of the expression *not to make a secret of something*.

Consider the sentence in (17), whose structure is given in Figure 5.

(17) dass Hans kein <u>Hehl</u> daraus macht
that Hans no secret thereof makes
"that Hans does not make a secret out of it"

Figure 5 shows that the word *Hehl* as it occurs in this sentence is described by the lexical entry in Figure 4. To check this, let us consider the role of the sign $\boxed{0}$, i.e., of the word itself, and of $\boxed{1}$ and $\boxed{2}$. The sign $\boxed{1}$ appears in the COLL value of *Hehl*. As required in the lexical entry, it dominates the word *Hehl*. Under the assumption that the signs $\boxed{1}$ and $\boxed{2}$ are identical, sign $\boxed{1}$ also (reflexively) dominates sign $\boxed{2}$. Just as required, sign $\boxed{2}$ dominates the CW and has the required logical term as its CONTENT value. Thus, all the requirements of the lexical entry of the word *Hehl* are met.

What is still missing at this point is a principle that guarantees that the sign in the COLL list of a lexical elements actually dominates the lexical element in a given structure. In (18) we close that gap and state the COLL PRINCIPLE informally.[11]

[10] Sailer (2000, Section 8.2) gives examples of fully formalized collocation statements.

[11] See Sailer (2000, Section 8.3) for an RSRL formalization.

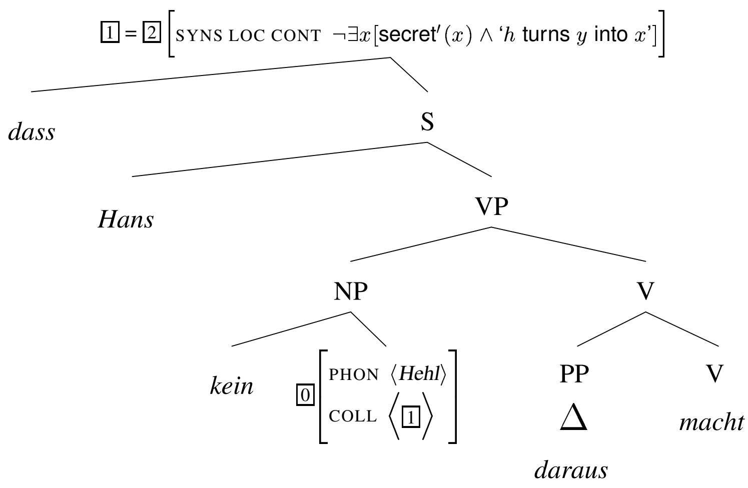

Figure 5: The structure of sentence (17)

(18) The COLL PRINCIPLE (COLLP):

If a sign $\boxed{1}$ dominates a sign $\boxed{2}$ which has a non-empty COLL value, then the element in $\boxed{2}$'s COLL value is a sign $\boxed{3}$ such that $\boxed{3}$ dominates $\boxed{1}$.

To see how the COLLP ensures that the distributional requirements of a lexical sign are enforced in every structure in which this sign occurs, consider again the tree in Figure 5. Assume that the VP node is the sign $\boxed{1}$ of the COLLP. It dominates a sign with a non-empty COLL value, namely the CW *Hehl* ($\boxed{2}$ in the COLLP). The element on the COLL list of *Hehl* is the highest node in the sentence. Thus, this node dominates the VP as well, just as required by the principle in (18). The same reasoning applies to all the other nodes in the tree. As an effect, the COLLP enforces that all lexical elements in a sentence have exactly the highest node of the sentence on their COLL list.

The COLLP is a single global principle which has the effect that the idiosyncratic distributional requirements of the lexical signs are respected. As the COLL element is the overall utterance, it contains enough structure to account for the data in (11) as well.

6 The Global Picture

Section 5 outlined a formal theory of collocational restrictions. The attraction of the present proposal is its simplicity: We only introduce one new attribute, COLL, and a single universal principle, the COLLP. Inspired by collocation theory, the attribute that identifies lexical elements is the same attribute that encodes the distributional restrictions

$$
\begin{bmatrix} \textit{word} \\ \text{PHON } \langle\rangle \\ \text{SYNS } \boxed{1} \begin{bmatrix} \text{LOC} & \boxed{2} \\ \text{NONL} & \begin{bmatrix} \text{INHER} & \begin{bmatrix} \text{SL} & \left\{ \boxed{2} \right\} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{COLL} \left\langle \boxed{3} \right\rangle \end{bmatrix}
\begin{array}{l} \textbf{and } \texttt{dominate}(\boxed{3},\boxed{4}) \\ \textbf{and } \boxed{4} \begin{bmatrix} \textit{word} \\ \text{SYNS LOC CAT} \begin{bmatrix} \text{HEAD} & \textit{subst} \\ \text{SBC} & \left\langle \begin{bmatrix} \textit{synsem} \end{bmatrix}, \ldots, \boxed{1}, \ldots \right\rangle \end{bmatrix} \end{bmatrix} \end{array}
$$

Figure 6: Sketch of the lexical entry of the trace, including the TRACE PRINCIPLE

of a lexical element and that determines its collocational behavior. In this sense COLL is the link between constructions and collocations. At the same time, the approach is fully lexical, as the lexical entry is the only place where idiosyncratic properties are stated.

The proposed architecture predicts that every lexical signs and only lexical signs may have distributional idiosyncrasies. While this claim would be subscribed to by researchers within the collocation tradition, it appears too strong and too unrestricted for linguists working in a Chomskian framework or in HPSG. On the other hand, the collocational approach may help in understanding the theoretical status of some grammatical principles. We are thinking of the principles of Binding Theory and the TRACE PRINCIPLE of PS94. As observed in Richter et al. (1999) these principles have a special status: they are stated as general principles of the grammar, but their purpose is to restrict the occurrence of single lexical elements, namely anaphora, pronouns and traces.

Let us consider the TRACE PRINCIPLE (TP) for illustration. It is the HPSG analogue of the *Empty Category Principle* of Chomsky (1981). It guarantees that a trace be a complement of some lexical head. In (19) we quote the principle from PS94.

(19) The TRACE PRINCIPLE (TP) of Pollard and Sag (1994:400):
The SYNSEM value of any trace must be a (noninitial) member of the SUBCAT list of a substantive word.

We can incorporate the TP into the lexical entry of the trace. The result is shown in Figure 6. The AVM on the left side is the description of the trace given in PS94, augmented by the COLL attribute and omitting the TO-BIND QUE and REL attributes. The two conjuncts on the right reformulate the TP as a lexical distributional restriction.

In a very similar way, the Binding Principles can be encoded in the lexical entries of pronouns and anaphora as done in Sailer (2000:435f). In van der Wouden and Zwarts (1993) and van der Wouden (1997) it is argued that the distribution of polarity items and of elements that participate in Negative Concord can be described in terms of collocational restrictions. In work on Negative Concord in French and Polish in Richter and Sailer (1999a,b) we provide such a collocational analysis.

The theoretical significance of collocational restrictions may reach far into areas of grammar that while being well-studied, could not yet be integrated coherently into the overall architecture of a formal grammar such as HPSG. Our study of the distribution of CWs highlights the necessity for a collocational module within formal grammar and opens a door for such an integration.

Even though the formal machinery introduced in Section 5 is small, the danger exists of opening a Pandora's box. As the COLL value makes the overall utterance available at the level of lexical elements, any sort of restriction can be stated as a collocational restriction. The question arises, whether there remains a place for selection. At the moment there is no definitive answer. As a working hypothesis, we accept the generative theory of selection and adopt a collocational analysis for those and only those distributional phenomena that cannot be handled within traditional assumptions of formal grammar. We hope that further research can motivate an architecture of linguistic signs that clearly delineates a separation of tasks between selection and collocation.

References

ABEILLÉ, A. (1995). The flexibility of French idioms: A representation with Lexical Tree Adjoining Grammar. In M. Everaert, E.-J. v. d. Linden, A. Schenk, and R. Schreuder (Eds.), *Idioms. Structural and Psychological Perspectives*, pp. 15–42. Lawrence Erlbaum Associates.

ARONOFF, M. (1976). *Word Formation in Generative Grammar*. MIT Press, Cambridge, Massachusetts and London, England.

BUTLER, C. S. (1985). *Systemic Lingistics: Theory and Application.* Batsford, London.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.

DOBROVOL'SKIJ, D. (1988). *Phraseologie als Objekt der Universallinguistik.* Verlag Enzyklopädie, Leipzig.

DODD, B. (2000). Introduction. The relevance of corpora to German studies. In B. Dodd (Ed.), *Working with German Corpora*. University of Birmingham Press, Birmingham.

ERNST, T. (1981). Grist for the linguistic mill: Idioms and 'extra' adjectives. *Journal of Linguistic Research 1*, 51–68.

FILLMORE, C., P. KAY, and M. O'CONNOR (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language 64*, 501–538.

FLEISCHER, W. (1989). Deutsche Phraseologismen mit unikaler Komponente — Struktur und Funktion. In G. Gréciano (Ed.), *Europhras 88. Phraséologie Contrastive. Actes du Colloque Internationale, Klingenthal-Strasbourg*, pp. 117–126.

FLEISCHER, W. (1997). *Phraseologie der deutschen Gegenwartssprache* (2nd, revised edition ed.). Niemeyer, Tübingen.

GAZDAR, G., E. KLEIN, G. PULLUM, and I. SAG (1985). *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, Mass.

HALLIDAY, M. (1966). Lexis as a linguistic level. In C. Bazell, J. Catford, M. Halliday, and R. Robin (Eds.), *In Memory of J.R. Firth*, pp. 148–162. Longman, London.

JOSHI, A. K. (1987). An introduction to Tree Adjoining Grammars. In A. Manaster-Ramer (Ed.), *Mathematics of Language*, pp. 87–114. John Benjamins Publishing Company, Amsterdam.

KASPER, R. (1997). Semantics of recursive modification. unpublished manuscript, Ohio State University..

KATZ, J. J. (1973). Compositionality, idiomaticity, and lexical substitution. In S. Anderson and P. Kearns (Eds.), *A Festschrift for Morris Halle*, pp. 357–376. Holt, Rinehart and Winston, New York.

KAY, P. (1997). *Words and the Grammar of Context*. CSLI Publications.

KAY, P. and C. J. FILLMORE (1997). Grammatical constructions and linguistic generalizations: the *what's x doing y?* construction. Manuscript. URL: `http://www.icsi.berkeley.edu/~kay/bcg/wxdy.ps`.

KRENN, B. and G. ERBACH (1994). Idioms and support verb constructions. In J. Nerbonne, K. Netter, and C. Pollard (Eds.), *German in Head-Driven Phrase Structure Grammar*, pp. 365–396. CSLI Publications. Lecture Notes 46.

NUNBERG, G., I. A. SAG, and T. WASOW (1994). Idioms. *Language 70*, 491–538.

POLLARD, C. and I. A. SAG (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

RICHTER, F. (2000). A mathematical formalism for linguistic theories with an application in Head-Driven Phrase Structure Grammar. Dissertation, Universität Tübingen.

RICHTER, F. and M. SAILER (1999a). A lexicalist collocation analysis of sentential negation and negative concord in French. In V. Kordoni (Ed.), *Tübingen Studies in Head-Driven Phrase Structure Grammar*, Arbeitspapiere des SFB 340, Nr. 132, Volume 1, pp. 231–300. Universität Tübingen.

RICHTER, F. and M. SAILER (1999b). LF conditions on expressions of Ty2: An HPSG analysis of Negative Concord in Polish. In R. D. Borsley and A. Przepiórkowski (Eds.), *Slavic in HPSG*, pp. 247–282. CSLI Publications.

RICHTER, F., M. SAILER, and G. PENN (1999). A formal interpretation of relations and quantification in HPSG. In G. Bouma, E. Hinrichs, G.-J. M. Kruijff, and R. T. Oehrle (Eds.), *Constraints and Resources in Natural Language Syntax and Semantics*, pp. 281–298. CSLI Publications.

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

SAG, I. A. (1997). English relative clause constructions. *Journal of Linguistics 33*, 431–483.

SAILER, M. (2000). Combinatorial semantics and idiomatic expressions in Head-Driven Phrase Structure Grammar. Dissertation, Eberhard-Karls-Universität Tübingen, Version of June 29th, 2000.

SINCLAIR, J. (1966). Beginning the study of lexis. In C. Bazell, J. Catford, M. Halliday, and R. Robin (Eds.), *In Memory of J.R. Firth*, pp. 410–430. Longman, London.

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

van der WOUDEN, T. (1997). *Negative Contexts. Collocation, Polarity and Multiple Negation*. Routledge, London and New York.

van der WOUDEN, T. and F. Zwarts (1993). A semantic analysis of negative concord. In U. Lahiri and A. Z. Wyner (Eds.), *SALT III: Proceedings of the Third Conference on Semantics and Linguistic Theory*, pp. 202–219.

WASOW, T., I. A. SAG, and G. NUNBERG (1983). Idioms: An interim report. In S. Hattori and K. Inoue (Eds.), *Proceedings of the XIIIth International Congress of Linguists*, pp. 102–115.

WEINREICH, U. (1969). Problems in the analysis of idioms. In Weinreich (1980), S. 208–264.

WEINREICH, U. (1980). *On Semantics*. University of Pennsylvania Press.