Lexical Licensing in Formal Grammar

Jan-Philipp Soehn
jp.soehn@gmail.com

October 2009

Abstract

This paper discusses instances of restricted combinability of lexical items (words and multi-word units) with their contexts. Different subtypes of distributional idiosyncrasies are presented, which occur on the phonological, morpho-syntactical, and semantic levels. Notably, external sandhi, cranberry words, decomposable idioms and (idiosyncratic) polarity items are addressed. These phenomena reveal an interesting interplay with regular language as well as between the different levels themselves. A detailed lexicalist analysis is provided within a formal grammar framework, Head-Driven Phrase Structure Grammar. This approach motivates an architecture of grammar that includes a module to accommodate specific restrictions on the occurrence environment of a lexical unit.

Keywords

Collocation, Sandhi, Cranberry Words, Idioms, Polarity Items, HPSG

1 Introduction

Idiosyncrasy, which is often confused with idiomaticity, is a controversial term in grammar theory. On the one hand, it has often been neglected as a “marginal” phenomenon. On the other hand, whole grammar theories such as Construction Grammar (Fillmore et al., 1988; Sag, 1997) are built on the idea of fixed combinations that form a construction, whereby idiomatic expressions seem to be the rule rather than the exception. In this paper, word cooccurrence phenomena are boiled down to the lexical level. Instead of treating all idiosyncratic cases alike as more or less fixed multi-word-expressions or "long words with spaces", the view is adopted that there are single lexical items that underlie idiosyncratic restrictions of compatibility with their context while nevertheless being perceived as independent units.

These distributional idiosyncrasies can be found on different levels. There are strict lexeme-lexeme combinations: cranberry words such as *to be **on** tenterhooks*[1] (cf. Dobrovol’skij, 1988; Dobrovol’skij and Piirainen, 1994), decomposable idiomatic expressions such as *to spill the beans* (cf. Nunberg et al., 1994; Sailer,

2003), and semantic selectional restrictions (cf. Katz and Postal, 1963; Chomsky, 1965; Androutsopoulos and Dale, 2000) such as *blonde*, which can modify only a very restricted class of real world entities (*hair* and *beer* and metonymically related ones). Certain lexemes such as polarity items (*any* or *rather*, cf. Ladusaw, 1980; Baker, 1970) have to co-occur with specific semantic contexts thus forming instances of a lexeme-semantics combination (see van der Wouden, 1997).

In the following, we will discuss yet another kind of lexical licensing, situated on the syntax-phonology interface: Sandhi forms such as *an* have to adjoin to appropriate phonological contexts, thus creating a lexeme-phonology co-occurrence phenomenon.

The paper is organized as follows: Section 2 presents relevant sandhi data for English, French, and Welsh. In Section 3, an analysis in a formal grammar framework, Head-Driven Phrase Structure Grammar (HPSG, along the lines of Pollard and Sag, 1994), is developed. The centerpiece of the approach is a collocation module, which has been previously discussed in connection with some of the phenomena mentioned above (Sailer and Richter, 2002; Sailer, 2003; Soehn, 2006). It allows the specification of idiosyncrasies in lexical entries and at the same time the combination of lexical units in regular (syntactic and semantic) ways. Section 4 sets the analysis into the broader context of these other phenomena, namely cranberry words, idioms and polarity items. The findings concerning all of these idiosyncrasies are brought together in order to create a synoptic picture and to emphasize the common property of restricted distribution. It turns out that the collocation module itself has to be revised to account for the different kinds of data. In Section 5 the revised module is applied to positive polarity items, which have not been discussed in the context of lexical licensing so far. Finally, a conclusion (Section 6) rounds off this proposal.

2 The Data

2.1 English Determiners

One of the earliest discussions on lexical-specific allomorphy influenced by the phonological environment such as the *a*/*an* alternation in English can be found in Bloomfield's description of external sandhi:

> Features of modulation and of phonetic modification play a great part in many syntactic constructions; they are known as sandhi. The form of a word or phrase as it is spoken alone is known as its absolute form; the forms which appear in included positions are its sandhi-forms. Thus, in English, the absolute form of the indefinite article is a [ˈej]. . . . If the next word begins with a vowel, there is a sandhi-form instead, an [ˈɛn], as in "not an uncle, but her uncle." (Bloomfield, 1935, p. 186).

This phenomenon is also known as 'shape conditions' or external allomorphy because allomorph selection has to take place after syntactic rules have applied (cf. e. g. Zwicky, 1985; Pullum and Zwicky, 1988). Zwicky proposes to treat all kinds of external allomorphy, such as the *a*/*an* alternation, in a separate component of grammar – the shape component. This has been addressed by Spencer (1991, Section 4.6) who also regards such data as "rather troublesome" for conventional grammar architectures "if we wish to maintain that lexically or morphologically conditioned alternations are limited to the lexicon, for this alternation is certainly lexically conditioned (it only happens to one word!), yet it seems to take place in the syntax." (Spencer, 1991, p. 128)

For a monostratal grammar framework such as HPSG this does not pose a problem. On the one hand, external sandhi can be treated as separate to other grammatical phenomena by dint of an extra part of the feature geometry (the collocation module for which there is independent evidence, see below). On the other hand, complex interactions are no problem to account for as there a no different stages or levels of derivation: The well-formedness of a sign is the consequence of all feature values being in accord with the signature, the lexicon, and the principles of grammar at the same time.

Despite the question about a separate grammar component there is the question of where to store the information. For Pullum and Zwicky (1988) this is a dilemma: "It is not part of the lexical entry for the word, because it refers to the following syntactic context. It is not a phonological rule of English, for it applies only to the indefinite article and has no general applicability to phonological domains." (p. 262). However, together with Spencer I want to stress the idiosyncrasy ("it only happens to one word!") of this alternation and simply drop "Pullum and Zwicky's assumption that lexical entries cannot refer to syntactic context; after all, reference to local syntactic domains is one of the tasks that lexicalized theories of grammar were originally designed to accomplish." (Asudeh and Klein, 2002, p. 1)

Asudeh and Klein offer a discussion within the framework of HPSG. In their analysis, the concept of *phonological context* plays a crucial role. However, in this approach the phonological context of a sign is described as its inherent lexical property. This means that a lexical entry of a word comprises not only phonological, syntactic, semantic, and pragmatic information – this kind of relationship is standardly assumed for syntactic subcategorization – but also a part of the phonology of the following or preceding word. However, as the phonology of an adjacent word does not "belong" to the sign in question but only exerts influence on it (e. g. triggers sandhi effects), one does not need to go that far for phonological interactions. A collocation module that stores the interaction lexically but whose mechanism does not work analogous to subcategorization seems to be the happy medium.

Thus, the allomorphy of English *a* and *an* (*a paper* vs. *an article*) is to be specified lexically because the "insertion" of an *n* before a vowel is not due to a general rule for English and occurs only for this one lexical item, i. e. the indefinite article. However, the phenomenon that an English determiner changes in front of

vowels is true for the definite article (*the*) as well. The form [ðiː] is used before a sandhi-triggering sound and before a speech pause.[2] Comparing, for example, the way that *the paper* [ðə ˈpeipəɹ] and *the article* [ðiː ˈɑrtɪkəl] are pronounced, illustrates this change. Similar phenomena can be found in other languages which are discussed in turn.

2.2 French Adjectives

Certain prenominal adjectives in French have an irregular masculine singular form in front of vowels. In French, a vowel clash has to be prevented in most syntactic environments (Obligatory Contour Principle) and therefore, certain so-called *liaison forms* exist for adjectives ending in a vowel sound. One way of "generating" a liaison form is to borrow the sound form of the lexeme in the opposite gender, such as [pətit] in *un petit article* ('a small article'), which sounds just like the feminine form *petite*. Interestingly, for French possessives it works in the opposite direction. The feminine forms *ma*, *ta*, and *sa* become *mon*, *ton*, and *son* in front of vowels (*mon amie* instead of **ma amie*, 'my friend'), forms identical to their masculine counterpart. For the adjectives in (1), it would have been possible to adopt the same strategy and take the feminine form as well but this is not the case: Only the sound form of the deviant masculine forms is identical to the respective feminine forms but not the way they are written.

(1)	*du beau temps* 'nice weather'	*un bel homme* 'a handsome man'	*une belle femme* 'a beautiful woman'
	un nouveau riche 'a new-rich'	*le nouvel an* 'new-year'	*la nouvelle république* 'the new republic'
	un fou rire 'roaring laughter'	*un fol hasard* [3] 'crazy coincidence'	*une folle journée* 'a mad day'
	mou [4]	*un mol oreiller* [3] 'a soft pillow'	*une molle intonation* 'a soft intonation'
	un vieux soldat 'an old soldier'	*un vieil homme* 'an old man'	*une vieille dame* 'an old lady'
	ce soldat 'this soldier'	*cet homme* 'this man'	*cette dame* 'this lady'

It is noteworthy that the feminine forms in the third column are not regular either with regard to their masculine forms. However, the feminine form is derived from the respective masculine sandhi-form in a regular fashion. In addition, one of these adjectives has an idiosyncratic masculine plural form: *vieux*. The others behave regularly in this respect, i. e. their plurals are derived from (independent) standard rules (*beau-x*, *nouveau-x*, *fou-s*, and *ce-s*). A different behavior can be seen with *gros*/*gros*/*grosse* ('fat') which doesn't take the feminine form either but,

differently to the items above, has an idiosyncratic sound form in the masculine liason case ([-z]) while the written form remains unaltered. This shows that the adjectives under discussion are really idiosyncratic. Traditional grammars of French (cf. Frontier, 1997, p. 184) just list these items as exceptions.
At the end of this subsection, the four different strategies in order to avoid a vowel clash are summarized.

1st strategy: borrow from the feminine form ([pətit] in *un petit article* 'a small paper')

2nd strategy: borrow from the masculine form (*ma*, *ta*, and *sa* become *mon*, *ton*, and *son*; *mon amie* instead of **ma amie* 'my friend')

3rd strategy: idiosyncratic sound for liaison cases (*gros* [-z])

4th strategy: idiosyncratic form for liaison cases (*bel, vieil*)

2.3 Welsh Function Words

The third kind of data to be discussed are some function words in Welsh. They exhibit word-final alternation (cf. Lapointe, 2001, and references therein), which is of exactly the same sort as the English *a*/*an* pair. However, the Welsh conjunction *a*/*ac* 'and', for example, occurs with the C-final form (which ends with a consonant) also before a set of C-initial function words (which begin with a consonant), among them *fel* 'like', *mewn* 'in', *myach* 'henceforth', *maddaf* 'I say', *na*/*nac* 'neither, nor', and *sydd* '3rd sg. form of *be*' (cf. Lapointe, 2001, p. 275 for the complete list containing prepositions, conjunctions, adverbs, polarity items, a quotative form, and two forms of *be*). Thus, it has to be e. g. *ac myach* 'and henceforth' and not *a myach*.

Lapointe notes (ibid.) that these sandhi-triggering function words "do not constitute an otherwise coherent semantic, grammatical, or phonological class in the language. [...] However, Welsh contains many other words in these categories which do not induce the exceptional behavior in the form of *a*/*ac*." Moreover, other function words that also begin with *f, m, n,* or *s* like the examples above do not trigger the C-final form *ac*. Finally, the *ac*-triggering function words do not themselves induce the presence of C-final forms in general; the pair *na*/*nac*, which should behave analogously to *a*/*ac*, is not affected, cf. *na(*c) mewn* NP 'nor in NP' vs. *a*(c) mewn* NP 'and in NP'. In conclusion, the conjunction *a*/*ac* behaves idiosyncratically – Thorne 1993, p. 425 just lists these cases in his grammar – not only with respect to the following segment but also with respect to the following lexeme, thus revealing an interplay between different kinds of lexical licensing.

In this section, idiosyncrasies of increasing level of complexity have been introduced. We have started with English determiners whose allomorphs are selected according to the phonological context on the right. Then, French adjectives show

a similar kind of lexical alternation. However, not only the phonological context is relevant here but also gender as morpho-syntactic feature. Lastly, we have seen Welsh function words which, in addition, are sensitive to certain lexical units on their right hand context.

3 Analyses

An analysis for the data just discussed requires some preliminary considerations. First, there is the issue of how to integrate phonological information into the grammar. Substantial work on phonology in HPSG began with Bird and Klein (1994) and also Höhle (1999), who have spelled out the value of the feature PHON from Pollard and Sag (1994) more precisely. For our discussion, the architecture below is sufficient in order to characterize a sign's PHONOLOGY.

$$\begin{bmatrix} \textit{sign} & & \\ \text{PHON} & \begin{bmatrix} \textit{phonology} & \\ \text{SEGMENTS} & \textit{list(segments)} \\ \text{PROSODY} & \textit{list(prosodic\_signs)} \\ \text{SANDHI-TRIGGER} & \textit{boolean} \\ \text{LONG-FORM} & \textit{boolean} \\ \text{APPENDIX} & \textit{segment} \end{bmatrix} \end{bmatrix}$$

The features LONG-FORM and APPENDIX are taken from Bonami et al. (2004) to model liaison in French. Moreover, their feature LIAISON-TRIGGER is adopted and dubbed SANDHI-TRIGGER (S-TRIG) to use a more general term. One might have two questions about this feature. First, why is it an attribute of *phonology* and not of *segment*? The answer is that the ability to trigger sandhi phenomena is a property of the left edge of a sign and not a property of each segment. If SANDHI-TRIGGER were appropriate to *segment*, it would have to be guaranteed that only the respective value of an initial segment has an effect whereas all other segments would have to be silent about their SANDHI-TRIGGER status or alternatively would have to be [SANDHI-TRIGGER −], which is conceptually unsatisfying.

A second question might be, why such a feature should be present at all, because it should follow from the initial segment's inherent properties whether it triggers sandhi effects or not. This might be true in some languages. To account for the regularity in English and German, the following principle which constrains the value of SANDHI-TRIGGER is assumed:

SANDHI-PLUS-PRINCIPLE (for English and German):

$$\begin{bmatrix} \textit{sign} \\ \text{PHON} \left[\text{SEGS} \left[\text{FIRST} \left[\text{SL MANNER CONSONANT} -\right]\right]\right] \end{bmatrix} \rightarrow \left[\text{PHON} \left[\text{SANDHI-TRIGGER} +\right]\right]$$

However, as can be seen in French, a given segment doesn't necessarily trigger liaison. The sound [w] in [wazo] (*oiseau*, 'bird') triggers liason in this case, but

not in [wikɛnd] (*weekend*), cf. *l'oiseau* vs. *le weekend*, *l'* being the elided form of *le* used for liaison. At first glance, this difference is reflected in the orthography of the given examples and one could envisage an additional principle for French in with the [w] is specified as sandhi trigger unless the first grapheme of the respective sign is "w". At a closer look however, some words starting with [w] and *o-* do not trigger liaison, see *ouate* 'cotton wool', *ouistiti* 'marmoset' or *ouiste* 'supporter'.

Another, well known issue in French liaison are words starting with the letter *h*: *la halle* ('the market hall') vs. *l'homme* ('the man'). In French grammars, the distinct sounds are called *h aspiré* (aspirated *h*) and *h muet* (mute *h*). Instead of stipulating two different segments for *h*, one with S-TRIG + and another with S-TRIG −, one could – as some dictionaries do – lexically specify the phonology of the words starting with an *h aspiré* with a glottal stop [ʔ] in front.

However, a glottal stop might be inserted in French also before vowels and mute *h* and, conversely, at least in Parisian French, the glottal stop before an *h aspiré* might be absent ([laʔal] or [la͡al] for *la halle*). In addition, it is possible to pronounce the schwa at the end of a feminine adjective before a noun starting with *h aspiré* but never before a noun starting with a vowel, cf. *une nette halte* [ynɛt(ə)alt] 'an abrupt stop' vs. *une vieille amie* [ynvjɛj(*ə)ami] 'an old friend'.[5] It is not a glottal stop, a segment of its own, that prevents liaison but it is an intrinsic property of the *h aspiré*. Thus, since segmental idiosyncrasies are present here, the feature SANDHI-TRIGGER is motivated, which can (and sometimes must) be lexically specified.

Another consideration adresses the question of how to refer to the phonology of a whole utterance, which is a non-trivial issue. In English, word order is in most cases identical to the underlying configurational structure. The relatively free word order in other Germanic languages for instance calls for a more sophisticated theory of linearization to license only the correct sequences of words and constituents. The surface ordering which is empirically observed follows from an interaction of linear precedence rules and lexical properties of some parts-of-speech (e. g. full NPs vs. pronouns). For HPSG, substantial work on word order began with Reape (1994, 1996), introducing a word-order domain (DOM) appropriate to each sign. This approach is further developed in Kathol (2000), who takes topological fields into account as well. Penn (1999) added data from Serbo-Croatian to the discussion, where prosody also has an influence on word order.

For the present analysis, a word order component is needed as well. Restrictions on the phonology of adjacent signs affect the "surface structure" (or phenogrammar, cf. Curry, 1961), not the grammatical structure proper (or tectogrammar), cf. (2). Although the determiner is the first constituent within an NP headed by *apple* in both cases, it is the word to the right that decides on the form *a* or *an*.

(2) **a / an apple* vs. *a / *an big apple*

It is assumed that each sign has a DOM list. Elements of that list are *dom_object*s in the style of Kathol (2000). I follow approaches by Crysmann (2002, 2005) who assumes that *dom_object*s contain information below PHONOLOGY and SYNSEM of the sign. The availability of SYNSEM information will be of particular importance in Section 3.3.

Finally, the collocation module is outlined. The conditions on licensing domains will be expressed in terms of a theory using the attribute COLL (Context of Lexical Licensing, cf. Soehn, 2004b), which builds on Sailer (2003) and provides the foundations of a theory of syntactic licensing domains. The collocational restriction is contained in the value of COLL, see Fig. 1. Each sign with idiosyncratic behavior can specify its requirements in COLL via its lexical entry.[6] Elements in this list are *barrier*-objects that have a PHONOLOGICAL-LICENSER (PHON-LIC) attribute. Barriers are phrases of a certain kind (*utterance*, *complete-clause*, *np*,...) which are identified as nodes in the syntactic configuration above the sign in question. The exact specifications of the relations that identify barriers as phrases in the structure are depicted in the Appendix.

$$\left[\begin{array}{ll} \textit{word} & \\ \text{PHON} & \ldots \\ \text{SYNSEM} & \ldots \\ \text{COLL} & \left\langle \left[\begin{array}{ll} \textit{barrier} & \\ \text{PHONOLOGICAL-LICENSER} & \textit{restriction on phon} \end{array}\right] \right\rangle \end{array}\right]$$

Figure 1: COLL feature

(3) LICENSING-PRINCIPLE:

For each sign *x* with a *barrier* element on its COLL list and for each phrase *z*:

– the value of PHON-LIC is identical to the DOM value of *z*

if and only if

1. *z* dominates *x*,
2. *z* can be identified as the barrier specified and
3. *z* does not dominate any sign *y* which in turn dominates *x* and forms an equivalent barrier.

The LICENSING-PRINCIPLE (LIP) in (3) guarantees that a specified barrier dominates the sign and meets all the criteria mentioned in the sign's lexical entry within COLL. An illustrative example is given in Fig. 7, discussed below. The restrictions in 1 to 3 in (3) make sure that the minimally dominating barrier is concerned. The conception of barriers provides a "window" in which collocation

restrictions must be satisfied. This is crucial to the restrictiveness of the theory: A sign may impose some restrictions on the smallest possible phrase which contains it. These restrictions in turn may constrain tectogrammatical (syntax, semantics) or phenogrammatical (word order, phonology) properties. The design of the COLL module is illustrated in Figure 2. This figure shows a part of a syntactic structure with an idiosyncratic item at the bottom. It imposes a restriction on a phrasal node (XP 1) which contains the item. The LICENSING-PRINCIPLE spots XP 1 as the minimally dominating node to which the restriction applies and identifies the PHON-LIC value of the item with the respective value of XP 1. If the restriction in COLL is incompatible with XP 1, i. e. the identification cannot be successful, there is a clash and the structure would be considered ungrammatical.

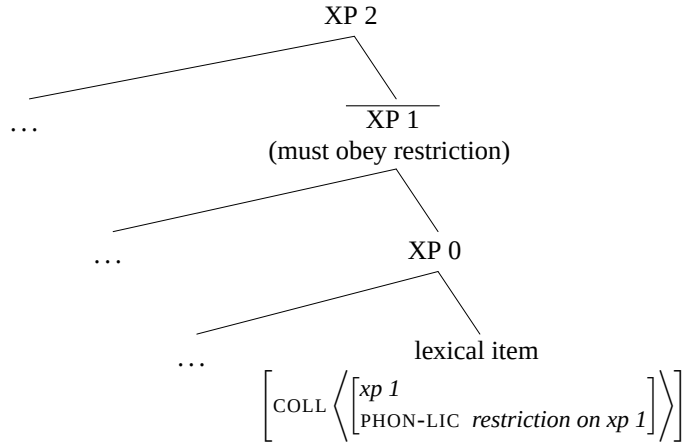

Figure 2: Design of the COLL module

This concludes the discussion of prerequisites for the analyses. In the following, lexical entries will be provided for sandhi phenomena of the sort discussed above.

3.1 English Determiners

Concerning phonological restrictions, the lexical item which imposes the restriction does not have information about adjacent words. Only some phrase which dominates it has sufficient "knowledge" about its parts and their relative position (on DOM). This is why it makes sense to "look upwards" in the tectogrammtical structure in order to access phenogrammatical information. The architecture of the collocation module reflects this and lets operations on the phonology of a phrase take place on the DOM list. The PHON-LIC value of a sign which imposes a restriction on its adjacent word is identical to the barrier's DOM list, which contains the neighbouring domain objects of that sign.

To begin with the analyses, take the lexical entry of *an*, sketched in Fig. 3.[7] The word *an* usually marks the beginning of an English indefinite NP.[8] Thus, the locus

of the licensing-constraint should be of sort *np*[9] and bear the additional constraint that the segment after *an* has to be a sandhi trigger. As *np* will be the minimal NP above *an*, coordination cases such as *an apple and a pear* or elliplis as in *a red _ and an orange fruit* where two NPs are conjoined will not pose any problems for this analysis. Any semantic and syntactic analysis of determiners will be compatible with our approach, thanks to its modular conception. Thus, a specification of the SYNSEM value is omitted in order to highlight only the important ingredients of the analysis.

The lexical entry for [ðiː] looks very similar. Here, only its use in front of vowels is taken into account. The licensing before speech pauses would call for a prosodic licensing which cannot be discussed here.

$$\left[\begin{array}{ll}
\textit{word} & \\
\textsc{phon} & \left[\textsc{segs}\ \boxed{1}\langle \text{ən}\rangle\right] \\
\textsc{orth} & \langle \texttt{an}\rangle \\
\textsc{synsem} & \textbf{indefinite article} \\
\textsc{coll} & \left\langle \left[\begin{array}{ll} \textit{np} & \\ \textsc{phon-lic} & \left\langle \left[\begin{array}{l}\textit{dom_obj}\\ \textsc{phon}\ \left[\textsc{segs}\ \boxed{1}\right]\end{array}\right], \left[\begin{array}{l}\textit{dom_obj}\\ \textsc{phon}\ \left[\textsc{sandhi-trigger}\ +\right]\end{array}\right], \ldots \right\rangle \end{array}\right] \right\rangle
\end{array}\right]$$

$$\left[\begin{array}{ll}
\textit{word} & \\
\textsc{phon} & \left[\textsc{segs}\ \boxed{1}\langle \text{ðiː}\rangle\right] \\
\textsc{orth} & \langle \texttt{the}\rangle \\
\textsc{synsem} & \textbf{definite article} \\
\textsc{coll} & \left\langle \left[\begin{array}{ll} \textit{np} & \\ \textsc{phon-lic} & \left\langle \left[\begin{array}{l}\textit{dom_obj}\\ \textsc{phon}\ \left[\textsc{segs}\ \boxed{1}\right]\end{array}\right], \left[\begin{array}{l}\textit{dom_obj}\\ \textsc{phon}\ \left[\textsc{sandhi-trigger}\ +\right]\end{array}\right], \ldots \right\rangle \end{array}\right] \right\rangle
\end{array}\right]$$

Figure 3: Lexical entries for *an* and *the*

3.2 French Adjectives

The analysis for the idiosyncratic adjective forms *bel*, *nouvel*, *vieil*, etc. is very similar to the English cases. The barrier for the licensing-constraint is defined to be of sort *np* as well and hence constrains the segment after the adjective to be a SANDHI-TRIGGER. The barrier *np* is the right choice here as these adjective forms occur only in front of nouns and cannot be used predicatively (*Cet homme est vieux/*vieil.* 'This man is old.'). The lexical entry for *vieil* (Fig. 4) serves as illustration (for an explanation of the features PHI and MAIN of CONTENT, see Fn. 15 on page 25).

The fact that French adjectives behave differently in front of vowels immediately evokes the phenomenon of liaison in general. An HPSG analysis for liaison is given by Tseng (2003), revised in Bonami et al. (2004). All ingredients of their analysis are compatible with ours, notably their "edge features" LONG-FORM and

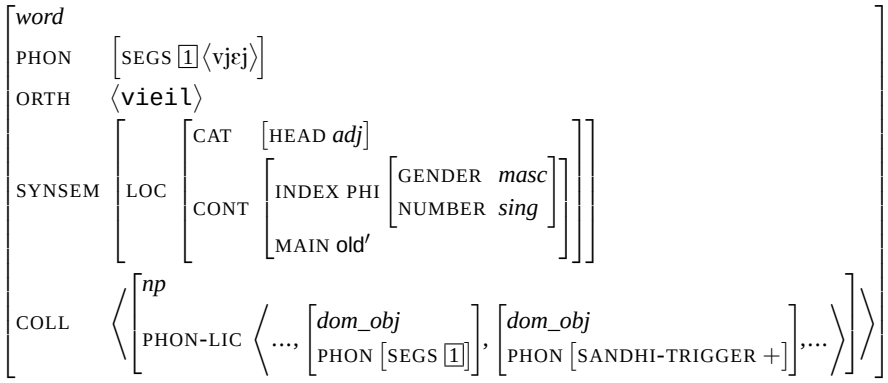

Figure 4: Lexical entry for *vieil*

LIAISON-TRIGGER (dubbed SANDHI-TRIGGER), although their position in the feature geometry has been changed to figure below *phonology* instead of *sign*. Their analysis can be combined with our approach in order to account for both the regular and the irregular data.

Bonami et al.'s approach works as follows: In the case of liaison, the right edge of a word's phonology is a long form (LFORM +), whereas the first item on the following word's phonology list must be a liaison trigger. The long form is obtained with the help of a relation between the phonology list and the so-called appendix. In the example *petit* ([pəti]) above, it would be the [t] which is appended. The fact that the appendix also plays a role in derivation and other morphological processes is independent evidence for the use of such a feature. Optional liaison is taken to be the default, and thus liaison can take place but does not have to. In certain syntactic configurations, liaison is forbidden and ruled out via constraints on specific kinds of phrases. In head-specifier structures liaison is obligatory, which in turn is also guaranteed by a constraint. The authors admit themselves that more work is needed to cast their approach in a linearization based analysis of French syntax. However, the phenomenon of liaison goes beyond the lexical level and well beyond what would count as idiosyncrasy.

As noted earlier in Section 2.2, the possessive adjectives in French deserve a closer look as well. They do not have a liaison form derived from the feminine form as was the case for adjectives such as *petit*. Instead, the feminine has a long form that is identical to the masculine. See (4) for the first person singular possessive – the same holds for the second and third person (*ton/ta, son/sa*). As for the other forms, regular liaison forms are applied, e. g. *ses* [seː]/[seːz].

(4) *le / mon bureau* (m) – 'the / my office'
la / ma maison (f) – 'the / my house'
l' / mon adresse (f) – 'the / my address'

If one wants to analyse this phenomenon in terms of referral (cf. Spencer, 1991), where the feminine possessive 'borrows' the sound form of its masculine

counterpart, a rule would be required which is diametrical to that for adjectives.

Lowenstamm (2007) has a different approach and analyses possessives in terms of insertion. In line with Kayne (2000, Ch. 8), Lowenstamm considers *mon/ma* to be bimorphemic, consisting of [m] and [õ/a]. In his approach, the nasal [õ] is inserted when the [a] falls prey to the Obligatory Contour Principle. However, this contrasts with the French definite article, where there is an alternation between *le*/*la* and *l'*: (*la adresse* → *l'adresse*, **lon adresse*). For *ma*, *ta* and *sa*, just dropping the vowel [a] does not work. They do not become *m'*, *t'* and *s'* – (*ma adresse* → *mon adresse*).[10]

Thus, French uses both possibilities for different lexical items. On the one hand, a vowel changes in front of a sandhi trigger for possessives, on the other hand, the [a] is simply dropped for the definite article. As there is no convincing explanation why in one case dropping is fine where in the other case insertion has to take place, this can be taken as an instance of idiosyncrasy which has to be encoded in the lexicon. The lexical entries for *l'* and *mon* are sketched in Figure 5.

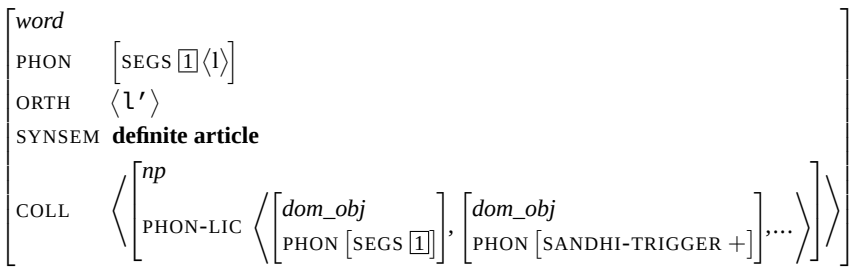

Figure 5: Lexical entries for *l'* and *mon*

3.3 Welsh Function Words

Figure 6 illustrates how to handle the kind of idiosyncrasy of the Welsh cases. The special context requirements of Welsh function words can best be captured by a disjunction in their lexical entries. The first disjunct on the DOM list in PHON-LIC

is a restriction on the phonological environment of the kind we've seen above. The barrier is specified as *utterance* because *a*/*ac* may conjoin phrases. I draw from the proposal by Carnie (2005) who assumes a flat syntactic structure in Welsh.[11] The second disjunct specifies in turn a disjunction of possible LISTEME values of the words that can occur adjacent to *ac*. The feature LISTEME is a unique identifier of lexical items (see also Section 4.1), which allows us to refer to particular words in the surrounding context of the idiosyncratic item in question. In the feature geometry, LISTEME is below [SYNSEM LOCAL CATEGORY HEAD]. In order to use LISTEME, it is necessary for a *dom*$_o$*bject* to have access to all features below SYNSEM. As shown in the data section above, there is no particular unifying feature of the possible adjoining words and a disjunction seems to be the best solution to describe the facts.

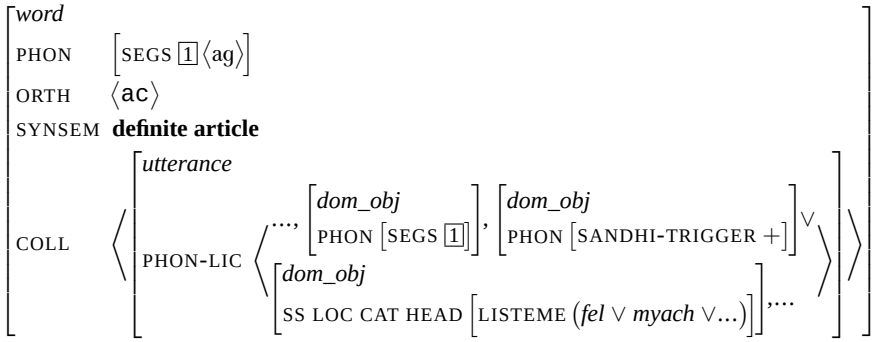

Figure 6: Lexical entry for *ac*

Asudeh and Klein (2002) mention two more phenomena. First, they refer to Welsh soft mutation (*ci* 'dog' → *ei gi* 'his dog') as an example of purely syntactic/semantic triggering of sandhi effects. Second, they regard Hausa final vowel shortening as an instance of mixed syntactic/semantic and phonologic triggering, similar to what has been discussed with respect to Welsh function words above.

The present approach can handle both kinds of sandhi triggering. However, I follow Pyatt (2003) in regarding Welsh soft mutation as a phonological phenomenon. According to Crysmann (2005), Hausa final vowel shortening (and also the clitic-affix alternation in Portuguese, cf. Crysmann, 2002) is to be handled in morphology. Although both Welsh soft mutation and Hausa final vowel shortening reveal some idiosyncrasies, these phenomena are not instances of lexical licensing and thus are not to be handled with the COLL module. In contrast, the lexical licensing cases in Welsh such as *a/ac* reveal idiosyncrasies on the phonological level as well as on the morpho-syntactic level which both can be accommodated within COLL. The phonological requirements are treated analogously to the English determiners, the morpho-syntactic ones are treated like cranberry words, which leads us directly to the next section.

4 Integrating other data

4.1 Cranberry Words and Idioms

The COLL approach has been used before to handle distributionally idiosyncratic lexical items. Among them, there are bound words (also called unique elements or cranberry words, Aronoff 1976), which are lexical units that have been "frozen" during language development over time. Dobrovol'skij (1988, p. 87) calls them relics from an earlier stage in language history. From a synchronic point of view, bound words are lexical elements which are highly collocationally restricted and are only grammatical when they co-occur with particular lexemes. The literature (Dobrovol'skij, 1988; Dobrovol'skij and Piirainen, 1994; Fleischer, 1989, 1997) mentions quite a large number of them. For German, the Collaborative Research Center 441 (Project A5) at the University of Tübingen compiled about 450 such instances.[12] Although the focus lies on German bound words, the phenomenon is by no means language specific. Dobrovol'ski's work mentions English, Dutch and Russian items. In the following examples, the bound word is underlined and the licensing context is printed in bold face.

(5) *Das "PC-Kummerbuch" ist* ***auf*** *Anhieb auf großes Interesse gestoßen.*
the PC-sorrow-book is at first-go on great interest hit
'The "PC Troubleshooting Companion" triggered interest right away.'

Combining another preposition with *Anhieb* is ungrammatical; see (6) for comparison to a non-unique noun.

(6) **auf**/***bei** *Anhieb* 'at first go' vs.
auf den/beim ersten Versuch 'at first attempt'

Similar English expressions are ***by*** *rote* 'mechanically' or ***in*** *a trice* 'as quickly as possible'. Sometimes, some variation may occur, as in *to* ***lie/go/lay*** *doggo* (Brit. slang; 'to hide oneself'), but their distribution remains nevertheless highly restricted.

An analysis requires a means to identify a particular item in a lexicon. For that, a feature LISTEME is adopted, following the idea put forth by Di Sciullo and Williams (1988). It seems in general that each word has a unique "identity" with a certain amount of idiosyncratic behavior. The possibility to select a particular word would, thus, be a useful feature for lexeme-lexeme combinations. For example, an expression like *to furrow one's brow* can be analyzed in the way that the verb *furrow* simply selects a word of the form *brow*: [SS LOC CAT SUBCAT [NP, NP_{refl_pron}, [LOC CAT HEAD [LISTEME *brow*]]]]. Another example for a selection of particular words is the perfect tense in German: a main verb has to be combined with the right auxiliary (*haben/sein*; Heinz and Matiasek, 1994, p. 222, use the attribute AUXF), cf. *er hat/*ist geschrieben* 'he wrote', *sie *hat/ist gerannt* 'she ran'.

The expression ***zu*** *Potte* ***kommen*** (“*to pot come*” – ‘to get going’ / ‘to get through’) serves as illustration for the analysis, which can be found in Soehn (2004a,b). *Potte* is a bound word that even requires more than one lexeme in its context: It needs to occur within a PP headed by ***zu***, which in turn has to be the complement of the verb ***kommen***. Because *Potte* can be regarded as the only idiosyncratic item in this expression, it encodes both criteria on its COLL list: a PP with the LISTEME value *zu* and another barrier of sort *complete-clause* with the LISTEME value *kommen*. Figure 7 illustrates the analysis, for which a separate feature LOCAL-LICENSER (LOC-LIC) is used whose value is identical to the LOCAL value of the barrier. This allows to reach the LISTEME value of the syntactic head.

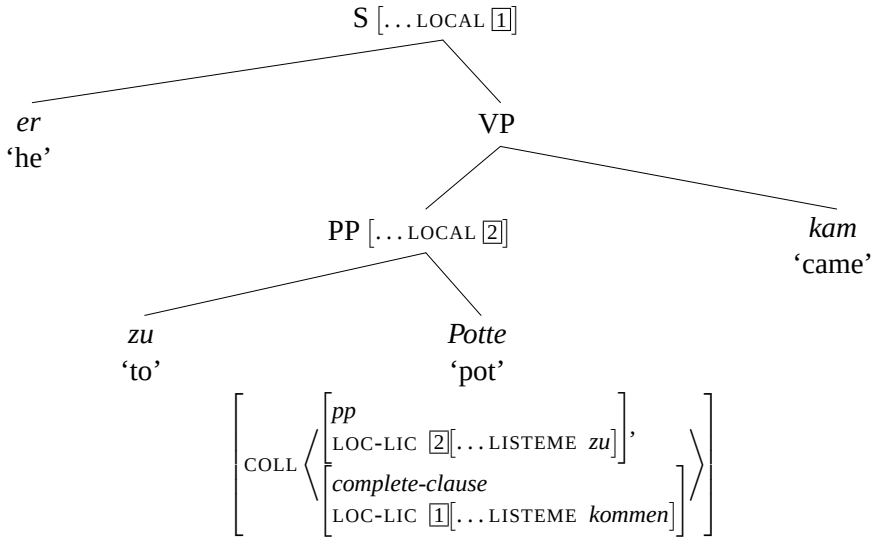

Figure 7: LE and partial context of *Potte* – The LIP guarantees the identities in 1 and 2

Sometimes, the environments of bound words can be idiomatic as in *to* ***lie*** *doggo*, mentioned above, and in *to* ***play*** *hooky* (‘be absent from school or work’). This leads us to the discussion of idioms. “Idiom” refers to idiomatic expressions that do not form complete sentences as would be the case for e. g. *His bark is worse than his bite.*

(7) *make waves* ‘cause trouble’

(8) *pay the fiddler* ‘face the consequences of one’s actions’

(9) *spill the beans* ‘divulge a secret’

The expressions in (7) to (9) are instances of decomposable[13] idioms, i. e. their overall meaning can be derived from the (idiomatic meaning of the) idiom parts. In (7), for example, the meaning ‘cause’ can be attributed to *make* and ‘trouble’ to *waves*.

On the one hand, the idiomatic meaning of the whole idiom consists of the idiomatic meanings of its parts. On the other hand, each part of the idiom is somehow bound to its context: Only the sum of all idiom-internal words make up the idiom. The parts by themselves may only be understood literally if at all. For example, *beans* has its idiomatic meaning 'secret' only in connection with *spill*. Thus, a grammar must somehow guarantee that idiom parts cannot occur freely and still retain their idiomatic meaning. An analysis which can account for this has to meet another demand: It must be flexible enough to include all possible changes (modification, passivization, etc. as noted above) that follow from independent factors. Thus, it would not be a good idea to encode an idiom such as *spill the beans* as an unalterable string whereas one might arguably encode *The early bird catches the worm* as a fixed phrase (cf. Nunberg et al., 1994). All these questions are addressed in Sailer (2003) concerning the analysis of idioms using COLL.

4.2 Negative Polarity Items

Another domain for which COLL has been used is negative polarity items (NPIs). These are words or idiomatic phrases that typically occur in an appropriately characterized – mostly negative – environment. NPIs can be found in any part-of-speech and they can be syntactically complex and clearly idiomatic, cf. Table 1 with examples from German (taken from Richter and Soehn (2006), all polarity items are underlined).

adverbs	*jemals* ('ever'), *beileibe* ('by no means')
nouns	*Deut* ('farthing'), *Menschenseele* ('soul')
adjectives	*geheuer* ('mysterious/scary'), *gefeit* ('immune')
verbs	*brauchen* ('need'), *ausstehen können* ('can stand/bear'), *wahrhaben wollen* ('want to see the truth')
idioms	*einen Finger rühren* ('to lift a finger') *seinen Augen trauen* ('to believe one's eyes') *(nicht) alle Tassen im Schrank haben* ('not to have all cups in the cupboard' - to have lost one's marbles)

Table 1: Examples of German NPIs in different POS

Van der Wouden (1997) argues in favor of a conceptualization of polarity sensitivity that describes it as collocational restriction. One of his arguments are synonyms as *besonders* / *sonderlich* ('particularly'), of which the latter is an NPI and the former is not: *Dieses Zimmer ist **(nicht)** besonders ordentlich.* vs. *Dieses Zimmer ist *(**nicht**) sonderlich ordentlich.* 'this room is not particularly tidy'. Thus, NPI-hood seems to be unpredictable on the basis of lexical semantics, in constrast to the general view discussed in the literature.[14] In this sense, NPIs are collocates which impose idiosyncratic restrictions on their contexts. In other words, an appropriate context – their collocate – triggers NPIs. This perspective predicts

idiosyncrasies in NPIs which are similar to those observed in idiomatic expressions or other lexicalized elements with a varying degree of fixedness. I want to argue with van der Wouden that occurrences of NPIs have abstract restrictions on their contexts. They require the presence of specific triggers such as negation, downward-entailing operators in general, questions, etc. According to their logical properties, these contexts are referred to as anti-morphic (AM, e. g. ***not***, ***without***), anti-additive (AA, e. g. ***nobody***, ***nothing***, including AM), and downward-entailing (DE, e. g. ***few***, ***at most***, including AM and AA), cf. Zwarts (1996).

Some expressions can have other, perhaps more subtle context requirements in addition to those which they have being an NPI: There are adverbs (e. g. Dutch ***moeilijk***, 'difficultly') which license only a subset of NPIs (those with a meaning aspect of abilty or possibility). Once the fact that NPIs are collocations is accepted, it is no surprise that a considerable number of NPIs are idiomatic as well. Thus, NPI-hood can be regarded as just another variant of idiomatic behavior.

A closer look at NPI data reveals that (i) NPIs are not licensed by a uniform type of licensers with a varying distance between licenser and NPI and (ii) being an NPI does not allow to fully predict the context requirements of a particular item. An NPI analysis in the framework of HPSG is given in Richter and Soehn (2006) where we argue that NPIs reveal idiosyncrasies on several levels. It is thus impossible for a grammar to account for the data with conventional means, such as specifying only the semantics of an NPI in order to restrict its occurrence. Therefore, in Richter and Soehn we make use of the collocational approach and exemplify this by the expression *einen Hehl aus etwas machen* ('to make a secret out of sth.'), among others. The negation which licenses *Hehl* can be inside the NP (10) or outside (in the VP as in (11) or even higher). If this constraint would be encoded exclusively in the semantics, one would have to restrict both the NP for the first use and the utterance for the second use, thus, having two different entries for the same expression. This would be conceptually very unsatisfying and a collocational approach describes the data more elegantly.

(10) Hans macht ***keinen*** *Hehl* aus seiner Meinung.
Hans makes not-a secret out his opinion
'Hans does not make a secret out of his opinion.'

(11) ***Niemand*** macht einen *Hehl* aus seiner Meinung.
Nobody makes a secret out his opinion
'Nobody makes a secret out of his opinion.'

The lexical entry of *Hehl* is given in Fig. 8 (cf. Richter and Soehn, 2006). This expression has been chosen because it illustrates the interaction between its NPI-related restrictions and its idiomatic restrictions. The first *barrier*-object on the COLL list constrains the semantic content of *Hehl* to DE environments and to

the scope of questions (or stronger licensers). For this, the feature LF-LICENSER is used whose value is identical to the LF value of the barrier, which encodes the semantic information of a sign.[15] The relation `quest-cond-comp-op` (i. e. a question, conditional, or comparative operator), whose formal definition is omitted here, is a means to express that the semantics of the barrier contains the relevant licensers and that *Hehl* is in their scope. This relation is placed in a hierarchy of relations (`am-strenght-op` ⊆ `aa-strenght-op` ⊆ `de-strenght-op` ⊆ `quest-cond-comp-op`) which imitates the licenser hierarchy of NPIs by Zwarts (1996) or van der Wouden (1997). The unifying element in the lexical entries of all NPIs is thus that the relation `am-strenght-op` holds, which states that the NPI in question is in the scope of an anti-morphic operator (e. g. plain negation). The second element on the COLL list is of sort *complete-clause* and imposes a different kind of restriction: The value of LOCAL-LICENSER, known from the idiom cases above, is identical to the LOCAL value of the clause in which *Hehl* appears. The head verb of this clause must be *machen*. In this analysis, there is a special version of *machen* that subcategorizes for the noun *Hehl*, and a PP, thus ensuring the co-occurrence of all parts of the idiomatic expression *einen Hehl aus etw. machen*: [SS LOC CAT SUBCAT [NP, [LOC CAT HEAD [LISTEME *machen*]], PP[aus]]].

$$\left[\begin{array}{l} \textit{word} \\ \textsc{phon}\ \langle \textit{hehl} \rangle \\ \textsc{ss}\quad \left[\textsc{local}\ \left[\begin{array}{ll} \textsc{cat head} & \textit{noun} \\ \textsc{content main} & \boxed{1}\ \mathsf{hehl}' \end{array}\right]\right] \\ \textsc{coll}\ \left\langle \left[\begin{array}{l} \textit{utterance} \\ \textsc{lf-licenser}\ \left[\textsc{exc}\ \texttt{quest-cond-comp-op}(\boxed{1})\right] \end{array}\right], \left[\begin{array}{l} \textit{complete-clause} \\ \textsc{local-licenser}\ \left[\textsc{cat head listeme}\ \textit{machen}\right] \end{array}\right] \right\rangle \end{array}\right]$$

Figure 8: A sketch of the lexical entry of *Hehl*

This concludes the analyses of lexical licensing cases. As a lexicalist grammar theory forms the basis of this approach, all relevant information is encoded in the respective lexical entries. As I argued that the distributional facts cannot be deduced from independent rules, this is not an artifact of grammar theories such as HPSG but a reasonable way to go – independent of the formalism one wants to adopt. I have shown that all the different kinds of occurrence requirements can be accommodated and couched by means of the feature COLL, whose existence (or that of an equivalent in other formalisms) now seems to be strongly corroborated.

Let us sum up the discussion so far: Several instances of distributional idiosyncrasies have been described in addition to the sandhi effects from the beginning. Although the occurrence patterns are located on different levels (phonology, morpho-syntax, and semantics, respectively), their common property is that the co-occurrence of an item and its licensing context is unpredictable: Rather similar

words such as *ac* and *nac* ('and'/'neither') behave differently with respect to their distribution. Thus, the co-occurrence of the idiosyncratic items and their licensing contexts cannot be deduced from independent phonological rules or syntactic or semantic selectional features. Traditional grammars just list the phenomena discussed as special cases or exceptions. As a consequence, a distribution module within the grammar is called for.

4.3 Revising the COLL Module

In the previous subsections lexical entries were sketched in order to illustrate the use of the COLL feature. If we were to unify the accounts for the different kinds of data, the overall feature geometry of COLL would look like this:

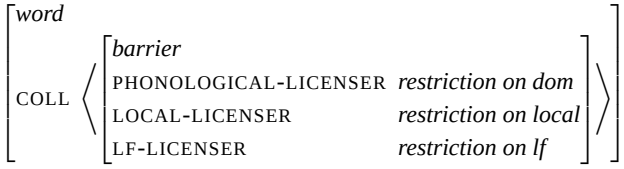

Figure 9: COLL module so far

This feature geometry does not place emphasis on the kind of licensing restriction but on the place where the restriction applies. Given the various data addressed above, this concept has to be reconsidered. Another problem with this is that the *barrier* object bears all three (and possibly more) …-LICENSER values, whereas in most cases only one of them will be used. The values of the others are identical to the respective values (e. g. the LOCAL value for LOCAL-LICENSER) – an information which is useless and which unnecessarily increases complexity. If there are two different restrictions as seen for *Hehl*, their locus might be not the same and two different barriers would have to be defined. A third issue is that this theory allows there to be any combination of restrictions per barrier. Up to now, there are interactions of PHON-LIC with LOC-LIC (for Welsh) and LOC-LIC with LF-LIC (for *Hehl*). However, is there a *barrier* object that includes restrictions on PHON-LIC and LF-LIC at once?

In order to remedy this, the COLL module must be slightly redesigned (see Fig. 10). In the new version, the COLL value is defined as a list of *licensing* objects which in turn have a LOCUS and a LICENSER attribute. Subsorts of *licensing* account for various kinds of distributional idiosyncrasies. Fig. 10 lists all subsorts at once for illustration, however, only a subset may be used in a particular entry. The feature LICENSER, appropriate for all *licensing* subsorts – with a different value according to the subsort – houses the restrictions. Barriers can still be specified - as the value of LOCUS. However, now the way is open to specify different subsorts of *barrier* which no longer have to be appropriate for all kinds of restrictions. That means, having a more detailed hierarchy below *barrier* one may distinguish

barriers that are relevant to phonology, syntax, and discourse. In (12), the comprehensive version of the LICENSING-PRINCIPLE is depicted.

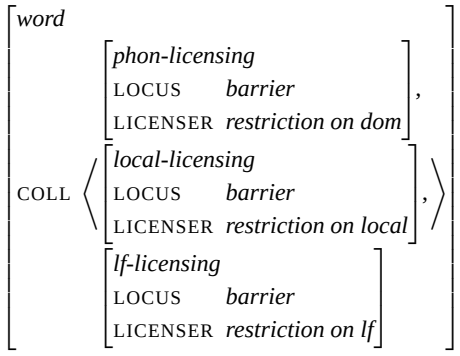

Figure 10: Redesigned COLL module

(12) LICENSING-PRINCIPLE (new version):

For each sign *x* with a *licensing* element on its COLL list and for each phrase *z*:

- the LICENSER value of *phon-licensing* is identical to the DOM value of *z*
- the LICENSER value of *local-licensing* is identical to the LOCAL value of *z*
- the LICENSER value of *lf-licensing* is identical to the EXC value of *z*

if and only if

1. *z* dominates *x*,
2. *z* can be identified as the barrier specified as the value of LOCUS and
3. *z* does not dominate any sign *y* which in turn dominates *x* and forms an equivalent barrier.

The cranberry word *Hochtouren* 'full throttle' in the phrase ***auf*** *Hochtouren* ***laufen*** 'to run at full throttle', which can be analyzed analogous to ***zu*** *Potte* ***kommen*** above, serves to illustrate the new setup, see Fig. 11 on page 21.

5 More Data: Positive Polarity Items

With the analysis set up and the collocation module redesigned, we can have a look at another set of data: positive polarity items (PPIs). The study of PPIs started notably with Baker (1970), yet they receive considerably less attention than their negative counterparts. Given their common usage and their diversity (cf. Table 2),

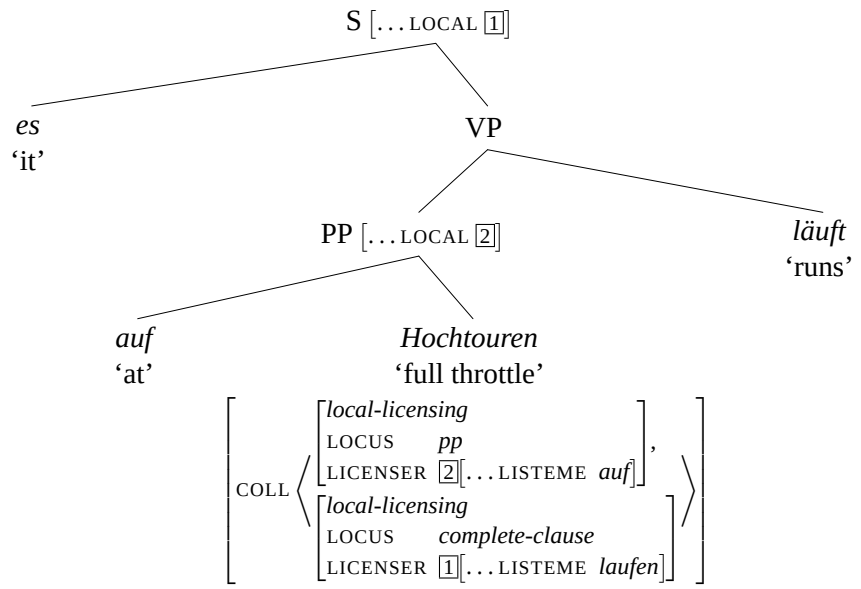

Figure 11: LE and partial context of *Hochtouren* – example for LIP's new version

this is quite surprising. Unlike NPIs, PPIs (such as *already*, *pretty* or *would rather*) cannot occur in the scope of negation, or in other words, NPI-licensing contexts have an anti-triggering effect on PPIs. However, theories differ with respect to their distribution. Von Bergen and von Bergen (1993) and, primarily, Szabolcsi (2004) suggest that PPIs are forbidden only in anti-morphic or anti-additive contexts whereas they are felicitous elsewhere. In contrast, van der Wouden (1997) argues that their distribution is more complex and that there are PPIs that are forbidden in downward-entailing contexts as well. As to the explanation of PPI-hood, Giannakidou (To appear) maintains that PPIs can be divided into several subcategories. Firstly, there are speaker-oriented items, mostly adverbs such as *unfortunately*, *probably*, and *unbelievably*. Their use in negative contexts is odd for the following reason (according to Ernst, 2005): Speaker-oriented adverbs express subjectivity and thus the speaker's commitment to the truth of the embedded proposition. Negation would require the proposition to be false in all possible worlds which would lead to a contradiction regarding the speaker's commitment. Secondly, van Os (1989) contends that all intensifiers in German are PPIs. Their use in the scope of negation would be odd for similar reasons as for speaker-oriented items. Thirdly, PPIs such as *some* as in *"I didn't eat something."* always have or trigger a referential interpretation and thus only a wide scope reading is available (cf. Giannakidou, to appear). A German example is *durchaus* 'definitely': The meaning of (13) is that there are definitely useful results but nobody was happy with them. Therefore, the PPI is felicitous in spite of being c-commanded by a negative element.

(13) Niemand war mit den *durchaus* brauchbaren Ergebnissen zufrieden.
nobody was with the definitely useful results happy
'Nobody was happy with the definitely useful results.'

However, idiosyncratic PPIs remain, which is in line with van der Wouden (1997). Just like the pair *besonders* / *sonderlich* there is also *sehr* 'very'/ *ziemlich* 'pretty'. The latter is a PPI and thus cannot occur in an NPI-licensing context:

(14) *Montags ist der Zug immer sehr / ziemlich voll.*
'On mondays the train is always very / *pretty* crowded.' vs.
*Montags ist der Zug **nie** sehr / *ziemlich voll.*
'On mondays the train is ***never*** very / **pretty* crowded.'

It doesn't follow from the meaning of *ziemlich* that it is a PPI which has to be encoded as idiosyncratic behaviour. A second argument can be built upon one of van der Wouden's PPI examples *verdienstelijk* 'meritorious'. In our research[16] it could be shown that the German counterpart *verdient* is not a PPI which demonstrates that there are differences concerning PI-hood within closely related languages (German and Dutch in this case).

intensifying adv.	*ausgesprochen* ('notedly'), *durchaus* ('definitely'), *geradezu* ('downright')
colloquials	*abgefahren* ('wacky'), *affengeil* ('phat'), *rattenscharf* ('red-hot'), *volle Kanne* ('full throttle')
speaker-oriented adv.	*erstaunlicherweise* ('astonishingly'), *glücklicherweise* ('fortunately'), *tragischerweise* ('tragically'), *zweckmäßigerweise* ('expediently')
idioms	*jmd. den Buckel runter rutschen* ('sb. can slide down one's back – sb. can take a slow boat to China'), *jmd. den Lebensfaden abschneiden* ('sb. the life-thread cut off – to kill sb.'), *eine Meise haben* ('to have a tomtit – to be stupid'), *einen in der Krone haben* ('to have one in the crown – to be drunk'), *sattsam bekannt sein* ('widely known be – to be notorious')
others	*leidlich* ('fair'/'passable'), *erstmals* ('for the first time'), *munkeln* ('to rumour'), *ungeachtet* ('notwithstanding'), *grassieren* ('to rage'), *lieber* ('rather'), *sowieso* ('in any case'), *ziemlich* ('pretty')

Table 2: Some examples of German PPIs

The item *ziemlich* (meaning also *rather* or *quite*) expresses a certain grade on a scale. As it is not an endpoint of the scale, all scalar NPI theories (cf. Fauconnier, 1975) fail to explain PIs like that. In addition, the item does not seem to convey a

speaker-commitment. Even worse, *ziemlich* has two equivalents: *besonders*, which is not sensitive to polarity, and *sonderlich*, an NPI. For this reason, its PPI-hood will have to be encoded in the lexical entry. Similar examples are *sattsam* 'well' – a bound word at the same time –, *grassieren* 'to rage', *munkeln* 'to rumour', *schlichtweg* 'utterly', and idioms of course, cf. Table 2.

For the lexical encoding of PPI-hood, the hierarchy of NPI licensing environments as sketched in Richter and Soehn (2006) can be used. For example, *ziemlich* is not allowed in DE contexts and thus it is forbidden in anti-additive and anti-morphic ones as well. The specification that *ziemlich* cannot occur in the scope of a `de-strength-operator` entails its ban from the other two contexts due to the hierarchy. In contrast, *ziemlich* can occur in conditionals and questions (cf. 15) which would allow *ziemlich* to be in the scope of a `quest-cond-comp-op` and an imperative.

(15) Na, geht Ihnen das Wetter in Österreich auch schon *ziemlich* auf die
well goes you the weather in Austria also already pretty on the
Nerven**?**
nerves
'Well, is the weather in Austria already getting on your nerves as well?'
Kleine Zeitung, 21.04.1997, Ressort: Lokal;

Analogously to NPIs, the information about positive polarity sensitivity is specified in the COLL list, sketched in Fig. 12. Inspired by (Pearce, 2001, p. 43) who names "those words which must **not** be used with the target word since they will lead to unnatural readings" as **"anti-collocations"**, I suggest that the relation between NPI-licensing contexts and PPIs is anti-collocational, because the former block the occurrence of the latter. This seems to be in line with the notion of "anti-triggering" in the NPI literature. Such an anti-collocation is expressed with a negated collocational restriction in Fig. 12.

$$\left[\begin{array}{l} \textit{word} \\ \text{PHONOLOGY} \left\langle \textit{ziemlich} \right\rangle \\ \text{SYNSEM LOCAL} \left[\begin{array}{ll} \text{CAT HEAD} & \textit{adjective} \\ \text{CONTENT MAIN} & \boxed{1}\ \text{ziemlich}' \end{array}\right] \\ \text{COLL} \left\langle \left[\begin{array}{ll} \textit{lf-licensing} & \\ \text{LOCUS} & \textit{utterance} \\ \text{LICENSER} & \left[\text{EXC} \ \neg\, \texttt{de-strength-op}(\boxed{1})\right] \end{array}\right] \right\rangle \end{array}\right]$$

Figure 12: Sketch of the lexical entry of *ziemlich*

6 Conclusion

Lexical idiosyncrasies can be found at several levels of language. From phonological via morpho-syntactic to semantic restrictions, there doesn't seem to be any area

where everything can be explained with rules and generalizations. In this paper, sandhi effects, cases of co-occurrence of particular lexical items and restrictions on the semantic environment have been discussed. Whether there are idiosyncrasies on the pragmatic level will have to be brought to light by further research. However, why shouldn't this be the case?

The interaction between regularity and irregularity is a challenging issue for grammar development. Consider the syntactic "fixedness" of decomposable idioms (cf. Fraser, 1970), notably passivization, where in addition to idiosyncratic behavior general grammar rules apply: Passivization of an idiom is basically only possible if the idiomatic verb has the appropriate valence structure to passivize (Dobrovol'skij, 1999). Moreover, there is not only an interplay between regular and idiosyncratic rules but also between idiosyncrasies on different levels. Data from Welsh have been presented that reveal context restrictions regarding both phonological and lexical features and data from German where lexical items can be NPIs and parts of idiomatic expressions at the same time.

A formal grammar is usually regarded as a means to generate all the valid strings of a language. It can also be used to determine whether any given string belongs to the language and to analyze it – i. e. to describe its internal structure. For this reason, formal linguists strive to describe language in full detail. While their main focus is on searching for generalizations and rules that cover the regular cases, idiosyncratic phenomena have to be taken into account in order to get the full picture. As one of the possible application areas of formal linguistics is natural language processing, the software has to "know" not only the rules but also the idiosyncrasies. Otherwise, it would generate phrases like *A uncle of mine spilled the secret about my birthday present for grandma. I can stand him a lot but for that, he'll pay the violinist!*[17]

Acknowledgements

The research on this paper was done during my work at the Collaborative Research Center 441, Tübingen University, Germany, and funded by the *Deutsche Forschungsgemeinschaft* and the *Landesstiftung Baden-Württemberg*. Parts of this paper has been presented as a poster ("Not just an analysis – Phonological Licensing of Cranberry Words") at "Linguistic Evidence 2008" in Tübingen. I am very grateful to the staff of the *Laboratoire de Linguistique Formelle* in Paris (in particular to Olivier Bonami, Bernard Fradin, Danièle Godard, Jean Lowenstamm, and François Mouret) for their warm welcome and for many fruitful discussions. In addition, I am indebted to Mingya Liu, Frank Richter, and Manfred Sailer for helpful comments on earlier versions of this paper. I would also like to thank three anonymous reviewers of "Research on Language and Computation"' for their comments and Lucas Ogden for the English language revision of this paper.

Notes

[1]Throughout this paper, bound words and polarity items are underlined, while licensing contexts are printed in bold face. In continuous text, examples appear in italics.

[2]The fact that [ðiː] is sometimes called the stressed form of *the* is ignored here, simply because [ðə] can also bear emphatic stress in some variants of English.

[3]This is rather archaic. Postposition (*un hasard fou, un oreiller mou*) is more commonly used.

[4]The masculine form *mou* does not occur before a noun except together with other adjectives: *en français elle aurait dû dire « on t'a abandonnée », ce qui ne serait qu'un mou, exsangue équivalent des mots russes...* N. Sarraute "Enfance" (1983, p. 182).

[5]Thanks to Olivier Bonami (p.c.) for pointing out this distinction.

[6]This means in contrast that non-idiosyncratic words and regularly built phrases have an empty COLL list.

[7]One minor issue is the inclusion of orthographic information of a sign. The orthography is – in addition to phonological properties – important surface information of linguistic entities, a sign's identification mark. Linguistic motivation for taking orthography into account is given by Fradin (2003). Orthographic information is relevant for morphological derivations affecting acronyms. For example, JOC (Jeunesse Ouvrière Chrétienne, [ʒɔk]) plus *-iste* becomes *jociste* [ʒɔsist], not [*ʒɔkist]. Thus, in order to formulate morphological rules, orthographic information has to be available.

[8]Predeterminers as in *quite an odd example* are not considered here. However, the linearization approach adopted here will be able to handle these cases.

[9]Even if parts of that NP will be extraposed, no linearization component would ever tear apart the *dom_obj* of an NP in such a way that the determiner would "get" a different neighbour on its right. Thus, we are on the safe side in having an *np* as barrier.

[10]Just for the sake of preciseness, [õ] is denasalized in front of the vowel as an assimilation effect.

[11]However, the exact definitions of *barrier*s in Welsh will have to be subject to closer scrutiny.

[12]See http://www.sfb441.uni-tuebingen.de/a5/codii/index.xhtml

[13]Cf. Raymond W. Gibbs et al. (1989) for the distinction between decomposable and non-decomposable idiomatic expressions from a psycholinguistic perspective.

[14]For reasons of conciseness, a discussion of NPI approaches in the literature is omitted here (cf. the overviews in Krifka, 1995; van der Wouden, 1997; Richter and Soehn, 2006).

[15]Richter and Soehn (2006) adopt Lexical Resource Semantics (LRS) for their analysis.

This semantic module uses expressions of Ty2 for logical representations of the meaning of natural language expressions. The EXC (external content) feature which appears in the lexical entry contains the logical form of a phrase, more precisely, the semantic content of its maximal projection. CONTENT MAIN houses the non-logical constant which is the nuclear semantic contribution of a lexical sign. The traditional ϕ-values person, number, and gender are located below INDEX PHI. A new and interesting idea is put forward by Sailer (2007), who uses a representational language, DRT (Discoures Representation Theory). Although a lexicalist or collocational approach as well, the theoretical significance of the representation of meaning is used (their logical form) to classify NPI licensers and not their denotation. However, it remains to be seen how weaker licensers such as interrogatives can be captured and whether there are plausible representations for all kinds of licensing contexts for a unified analysis.

[16]Psycholinguistic acceptability judgement experiments have been conducted which are reported in Liu and Soehn (2009).

[17]The grammatical alternative would be: *An uncle of mine spilled the beans about my birthday present for grandma. I like him a lot ($\neq$ can't stand him at all) but for that, he'll pay the fiddler!*

References

Androutsopoulos, I. and R. Dale: 2000, 'Selectional Restrictions in HPSG'. In: *Proceedings of COLING 2000*. Saarbrücken, pp. 15–20.

Aronoff, M.: 1976, *Word Formation in Generative Grammar*. MIT Press, Cambridge MA, third printing 1985. Linguistic Inquiry Monographs.

Asudeh, A. and E. Klein: 2002, 'Shape Conditions and Phonological Context'. In: F. van Eynde, L. Hellan, and D. Beermann (eds.): *Proceedings of the 8th International HPSG Conference*. pp. 20–30, CSLI Publications. http://cslipublications.stanford.edu/HPSG/2/.

Baker, C. L.: 1970, 'Double Negatives'. *Linguistic Inquiry* **1**, 169–186.

Bird, S. and E. Klein: 1994, 'Phonological analysis in typed feature systems'. *Computational Linguistics* **20**, 455–491.

Bloomfield, L.: 1935, *Language*. George Allen & Unwin, London. Linguistic Inquiry Monographs.

Bonami, O., G. Boyé, and J. Tseng: 2004, 'An Integrated Analysis of French Liaison'. In: G. Jäger, P. Monachesi, G. Penn, and S. Wintner (eds.): *Formal Grammar 2004 Preproceedings*. Nancy, France.

Carnie, A.: 2005, 'Flat Structure, Phrasal Variability and VSO'. *Journal of Celtic Linguistics* **9**. http://dingo.sbs.arizona.edu/∼carnie/publications/.

Chomsky, N.: 1965, *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Crysmann, B.: 2002, 'Constraint-based Coanalysis. Portuguese Cliticisation and Morphology-Syntax Interaction in HPSG'. Ph.D. thesis, Saarland University.

Crysmann, B.: 2005, 'Hausa Final Vowel Shortening — Phrasal Allomorphy or Inflectional Category?'. In: *On-line Proceedings of the 4th Mediterranean Morphology Meeting (MMM4)*, Catania, Sicily.

Curry, H. B.: 1961, 'Some Logical Aspects of Grammatical Structure'. In: R. O. Jakobson (ed.): *Structure of Language and its Mathematical Aspects*, Vol. 12 of *S*ymposia on Applied Mathematics. Providence: American Mathematical Society, pp. 56–68.

Di Sciullo, A.-M. and E. Williams: 1988, *On the Definition of Word*, Linguistic Inquiry Monographs. MIT Press, Cambridge, Mass, second printing.

Dobrovol'skij, D.: 1988, *Phraseologie als Objekt der Universalienlinguistik*. VEB Verlag Enzyklopädie Leipzig. Linguistische Studien.

Dobrovol'skij, D.: 1999, 'Gibt es Regeln für die Passivierung deutscher Idiome?'. In: I. Bäcker (ed.): *Das Wort. Germanistisches Jahrbuch*. DAAD, Bonn.

Dobrovol'skij, D. and E. Piirainen: 1994, 'Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative'. In: *Folia Linguistica XXVIII/3-4*. Mouton de Gruyter, Berlin, pp. 449–473.

Ernst, T.: 2005, 'On Speaker-Oriented Adverbs as Positive Polarity Items'. Electronic Poster for the Workshop: Polarity From Different Perspectives, New York University, 11.–13.03.2005.

Fauconnier, G.: 1975, 'Pragmatic Scales and Logical Structure'. *Linguistic Inquiry* **6**, 335–375.

Fillmore, C., P. Kay, and M. O'Connor: 1988, 'Regularity and Idiomaticity in Grammatical Constructions: The Case of *L*et 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*. Niemeyer, Tübingen, second revised edition.

Fradin, B.: 2003, *Nouvelles approches en morphologie*, Linguistique nouvelle. Presses Universitaires de France.

Fraser, B.: 1970, 'Idioms within a Transformational Grammar'. *Foundations of Language* **6**, 22–42.

Frontier, A.: 1997, *La Grammaire du Français*. Paris: Belin.

Giannakidou, A.: To appear, 'Negative and Positive Polarity Items: Variation, Licensing, and Compositionality'. In: C. Maienborn, K. von Heusinger, and P. Portner (eds.): *Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter.

Heinz, W. and J. Matiasek: 1994, 'Argument Structure and Case Assignment in German'. In: J. Nerbonne, K. Netter, and C. Pollard (eds.): *German in Head-Driven Phrase Structure Grammar*. CSLI Publications, pp. 199–236. Lecture Notes 46.

Höhle, T. N.: 1999, 'An Architecture for Phonology'. In: R. D. Borsley and A. Przepiórkowski (eds.): *Slavic in Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, pp. 61–90.

Kathol, A.: 2000, *Linear Syntax*. New York: Oxford University Press.

Katz, J. J. and P. M. Postal: 1963, 'Semantic Interpretation of Idioms and Sentences Containing Them'. In: *Quarterly Progress Report*, No. 70. Massachusetts Institute of Technology, Research Laboratory of Electronics: pp. 275–282.

Kayne, R. S.: 2000, *Parameters and Universals*, Oxford Studies in Comparative Syntax. Oxford University Press.

Krifka, M.: 1995, 'The Semantics and Pragmatics of Weak and Strong Polarity Items'. *Linguistic Analysis* **25**, 209–257.

Ladusaw, W.: 1980, *Polarity Sensitivity as Inherent Scope Relations*. Garland Press, New York.

Lapointe, S. G.: 2001, 'Stem selection and OT'. In: G. Booij and J. van Marle (eds.): *Yearbook of Morphology 1999*. Kluwer Academic Publishers, London, pp. 263–297.

Liu, M. and J.-P. Soehn: 2009, 'Empirical Perspective on Positive Polarity Items in German'. In: S. Featherston and S. Winkler (eds.): *Fruits: Process and Product in Empirical Linguistics*, Vol. 2: Product. Berlin: de Gruyter, pp. 197–216.

Lowenstamm, J.: 2007, 'On Little N, $\sqrt{}$, and Types of Nouns'. Manuscript. Université Denis Diderot, Paris.

Nunberg, G., I. A. Sag, and T. Wasow: 1994, 'Idioms'. *Language* **70**, 491–538.

Pearce, D.: 2001, 'Synonymy in Collocation Extraction'. In: *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*. Carnegie Mellon University, Pittsburgh, pp. 41–46.

Penn, G.: 1999, 'An RSRL Formalization of Serbo-Croatian Second Position Clitic Placement'. In: V. Kordoni (ed.): *Tübingen Studies in Head-Driven Phrase Structure Grammar*. Tübingen, pp. 177–197, Universität Tübingen. http://www.sfs.uni-tuebingen.de/∼gpenn/cards/rsrl.html.

Pollard, C. and I. A. Sag: 1994, *Head-Driven Phrase Structure Grammar*. Stanford University: CSLI/The University of Chicago Press.

Pullum, G. K. and A. M. Zwicky: 1988, 'The syntax-phonology interface'. In: F. J. Newmeyer (ed.): *Linguistics: The Cambridge Survey*, Vol. I of *Linguistic Theory: Foundations*. Cambridge University Press, pp. 255–280.

Pyatt, E. J.: 2003, 'Relativized Mutation Domains in the Celtic Languages'. In: *Proceedings from the Penn Linguistics Colloquium 26*.

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

Reape, M.: 1994, 'Domain Union and Word Order Variation in German'. In: J. Nerbonne, K. Netter, and C. J. Pollard (eds.): *German in Head-Driven Phrase Structure Grammar*, No. 46 in CSLI Lecture Notes. Stanford University: CSLI Publications, pp. 151–197.

Reape, M.: 1996, 'Getting things in order'. In: H. Bunt and A. van Horck (eds.): *Discontinuous Constituency*, No. 6 in Natural language processing. Berlin, New York: Mouton de Gruyter, pp. 209–253.

Richter, F.: 1997, 'Die Satzstruktur des Deutschen und die Behandlung langer Abhängigkeiten in einer Linearisierungsgrammatik. Formale Grundlagen und Implementierung in einem HPSG-Fragment'. In: E. Hinrichs, D. Meurers, F. Richter, M. Sailer, and H. Winhart (eds.): *Ein HPSG-Fragment des Deutschen, Teil 1: Theorie*, No. 95 in Arbeitspapiere des SFB 340. Universität Tübingen, pp. 13–187.

Richter, F.: 2007, 'Closer to the Truth: A New Model Theory for HPSG'. In: J. Rogers and S. Kepser (eds.): *Model-Theoretic Syntax at 10 (Workshop Proceedings of MTS@10, organized as part of ESSLLI'07 at Trinity College in Dublin, Ireland)*. pp. 99–108.

Richter, F. and J.-P. Soehn: 2006, '*Braucht niemanden zu scheren*: A Survey of NPI Licensing in German'. In: S. Müller (ed.): *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, pp. 421–440.

Sag, I. A.: 1997, 'English Relative Clause Constructions'. *Journal of Linguistics* **33**, 431–483.

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

Sailer, M.: 2007, 'NPI Licensing, Intervention and Discourse Representation Structures in HPSG'. In: S. Müller (ed.): *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*. Stanford, pp. 214–234, CSLI Publications.

Sailer, M. and F. Richter: 2002, 'Collocations and the Representation of Polarity'. In: G. Alberti, K. Balogh, and P. Dekker (eds.): *Proceedings of the Seventh Symposium on Logic and Language*. Pécs, pp. 129–138.

Soehn, J.-P.: 2004a, 'About Spilled Beans and Shot Breezes. A New Word-level Approach to Idioms'. In: G. Jäger, P. Monachesi, G. Penn, and S. Wintner (eds.): *Proceedings of Formal Grammar 2004, Nancy*. pp. 125–140.

Soehn, J.-P.: 2004b, 'License to COLL'. In: S. Müller (ed.): *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*. Stanford, pp. 261–273, CSLI Publications. http://cslipublications.stanford.edu/HPSG/5/.

Soehn, J.-P.: 2006, *Über Bärendienste und erstaunte Bauklötze – Idiome ohne freie Lesart in der HPSG*, No. 1930 in Deutsche Sprache und Literatur. Frankfurt: Peter Lang Publishing Group. Phil. Dissertation (2005), Friedrich-Schiller-Universität Jena.

Spencer, A.: 1991, *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Blackwell, Oxford.

Szabolcsi, A.: 2004, 'Positive Polarity — Negative Polarity'. *Natural Language and Linguistic Theory* **22**(2), 409–452.

Thorne, D. A.: 1993, *A Comprehensive Welsh Grammar*. Oxford: Blackwell.

Tseng, J. L.: 2003, 'EDGE Features and French Liaison'. In: J.-B. Kim and S. Wechsler (eds.): *The Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, pp. 313–333. http://cslipublications.stanford.edu/HPSG/3/.

van der Wouden, T.: 1997, *Negative Contexts. Collocation, polarity and multiple negation*. Routledge, London and New York.

van Os, C.: 1989, *Aspekte der Intensivierung im Deutschen*. Tübingen: Gunter Narr.

von Bergen, A. and K. von Bergen: 1993, *Negative Polarität im Englischen*. Tübingen: Gunter Narr.

Zwarts, F.: 1996, ‘A Hierarchy of Negative Expressions’. In: H. Wansing (ed.): *Negation: A Notion in Focus*. Berlin/New York: de Gruyter, pp. 169–194.

Zwicky, A. M.: 1985, ‘Rules of allomorphy and phonology-syntax interactions’. *Journal of Linguistics* **21**(2), 431–436.

Appendix

Figure 13 on page 32 contains the relations that identify barriers as phrases in the structure. These relations are necessary in order for the LICENSING-PRINCIPLE to work. The sort names are straightforward except *vp_ne* which refers to a VP that does not contain any extracted element (empty NONLOCAL INHERITED SLASH value). Concerning the feature geometry cf. also 13; *barrier* itself is below the topmost sort *object*. An *utterance* refers to an unembedded phrase which has – in addition to its syntactic completeness (STATUS *complete*) – the property of being the smallest unit with an illocutionary force (bearing the feature ILLOCUTION). The features STATUS and ILLOCUTION are taken from Richter (1997, S. 68f and 136) who creates an interface to a theory of illocution: phrases with an illocutionary force are always unembedded and do not contain any unbound traces. In order to make this distinction, Richter enhances the sort hierarchy introducing *embedded-sign*s and *unembedded- sign*s (ibid., p. 135, cf. also Richter, 2007, p. 102). Unembedded signs bear the feature ILLOCUTION with *question*, *assertion*, *exclamation* and others as possible value. Richter (ibid.) argues that a more fine-grained sort hierarchy below sign and the introduction of the sort *unembedded-sign* are necessary because unembedded signs as independent utterances are a central concept of grammar theory which is an empirically describable and perceptible object of linguistics.

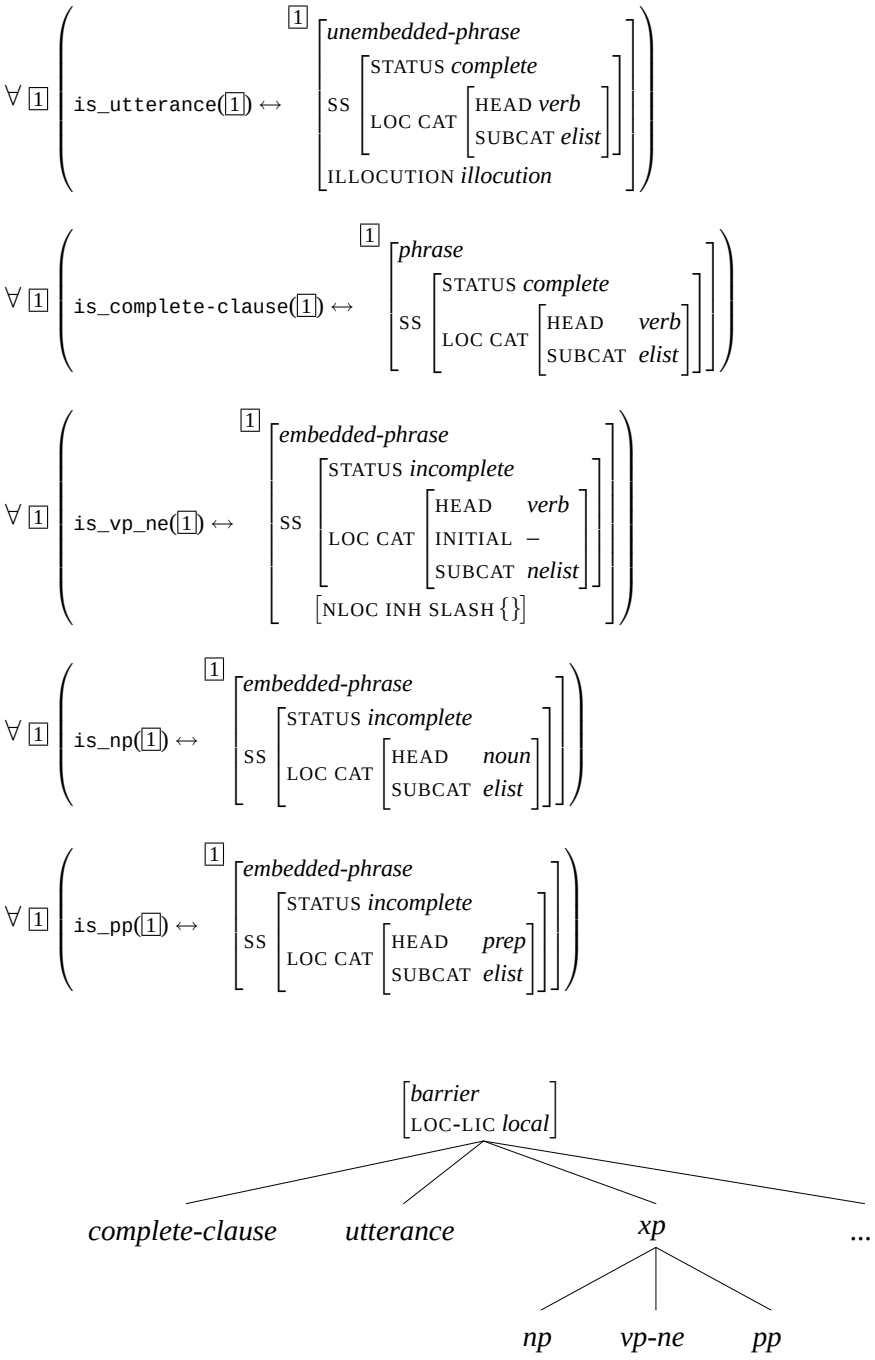

Figure 13: Relations for *barrier*-subsorts and sort hierarchy for *barrier*