HPSG LEXICON WITHOUT LEXICAL RULES

karel oliva

Department of Computational Linguistics, University of Saarland, Germany
e-mail: oliva@coli.uni-sb.de

Summary: The paper introduces an alternative to the lexical rules in a lexicon in a HPSG style by replacing them by relational constraints corresponding more directly to the standard lexicographic and morphological practice.

1. INTRODUCTION

The Head-driven Phrase Structure Grammar (HPSG, Pollard & Sag,87 and 93) came into being as and remains a predominantly syntactic theory. This is reflected not only in the main theoretical orientation, but also in the structuration of linguistic data within (the standard versions of) the theory.

As the majority, if not all, current theories, HPSG relies on an ample representation of linguistic knowledge within the lexicon rather than in the grammar rules. This knowledge, then, is in HPSG organized in a crossing inheritance hierarchy of lexical types, which should express the regularities and generalizations occurring in the lexicon and thus avoid redundant stipulations.

However, in spite of the fact that the lexicon - at least as far as the amount of information is concerned - plays an absolutely central role in the theory, the organization of data within HPSG clearly reflects the primary syntactic concern of the theory, with lexicon on a secondary position only. In particular, this comes into light when the presence of more traditional views of organization of lexical data (proved useful by long lexicographic and linguistic practice) is sought - such views are represented only marginally or not at all[1].

In this paper, I shall try

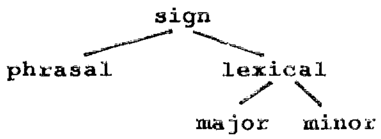
(i) to show that neglecting standard insights of the organization of lexicon is detrimental both to the linguistic adequacy and to the practical usefulness of the lexicon,

(ii) to make a proposal of an alternative reconciling the needs of HPSG with the usual lexicographic practice.

2. THE HPSG STANDARD

As the basic type of object of linguistic description HPSG adopts the sign[2], and builds a subsumption hierarchy of its subtypes, whose top level is sketched in (1).

(1)



The idea behind this hierarchy is that each node is associated with bundles of features, which are inherited by all nodes directly or indirectly subordinated to this node, by means of which a lot of redundant stipulations of features can be removed.

There are, however, two points about the hierarchy from (1) which may be worth reconsidering. First, it is the fact that linguistics in general is concerned with a broader class of objects than signs - e.g., phonemes, morphs etc. are objects having no semantics, i.e. they - by definition - cannot fall into the class of signs, and if HPSG aims at becoming a full linguistic theory, these objects have to be also taken into consideration. Second, the hierarchy reflects only the syntactically motivated division of lexical signs into major and minor ones, but the other possible - and in fact more standard - divisions into autosemantic vs. synsemantic[3] words and productive vs. nonproductive word classes are missing.

Apart from this, HPSG stipulates the lexicon to be a full-form one, i.e. a lexicon where all word-forms of a language occur as actual items in the lexicon (leaf nodes in the hierarchy). This is also an approach which hardly finds a parallel in the standard lexicographic practice (even for languages with that poor morphology as English has), and in addition an approach which at least on the explanatory level enforces the necessity of lexical (redundancy) rules - which, on the other hand, do not fit into the general image of HPSG at all.

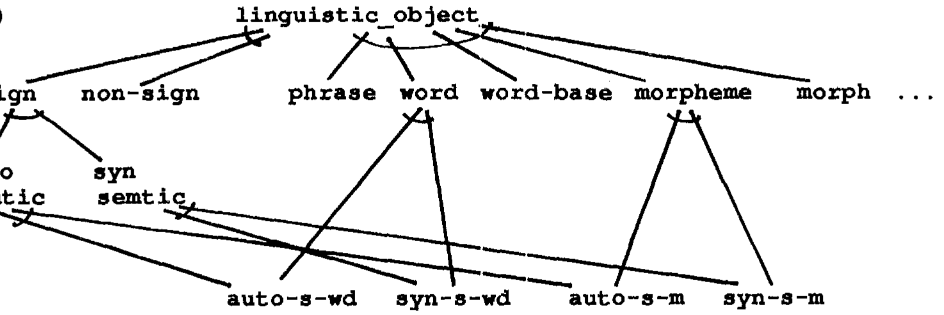
Provided that the lexicon is viewed from a "declarative" perspective[4], the lexical rules express static relationships (such as inflection) between members of word classes. These relationships are, however, not needed for the functioning of the system (since all word forms are available anyway) - they have to exist only because without them the lexicon would be linguistically clearly inadequate.

The other option, namely the "procedural" perspective, seems to be even worse, since it breaks the overall "declarative" strategy of HPSG; in particular, it enforces inequality of status of different lexical items (some being "basic" and some being only "derived" - in the very procedural sense of the word).

3. RELATIONAL CONSTRAINTS AS ALTERNATIVE

Based on the preceding facts and (first of all) on the standard linguistic practice, it is possible to propose an alternative hierarchy of objects a linguistic theory should deal with, whose "top part" can be sketched as in the crossing hierarchy given in (2), where appurtenance of several subtypes of a type to classification according to the same key is marked off by an arc connecting the respective branches of the hierarchy - in particular, two divisions according to different keys are to be observed for the class "linguistic object", one corresponding to a "functional" perspective (the division into signs and non-signs), the other one corresponding to a "formalistic" perspective[5].

(2)



Among members of this hierarchy, certain parallels are to be observed. In particular, it is worth to observe the parallel between the construction of phrases (which consist of words) and the construction of words (consisting of morphemes and - in case of composed words - of bases, i.e. combinations of autosemantic morphemes). In an HPSG-like notation, the "consisting of" can be approximated by the feature structure (3), with obvious meaning of the relational constraints **phon_consists_of** and **sem_consists_of.**

(3)

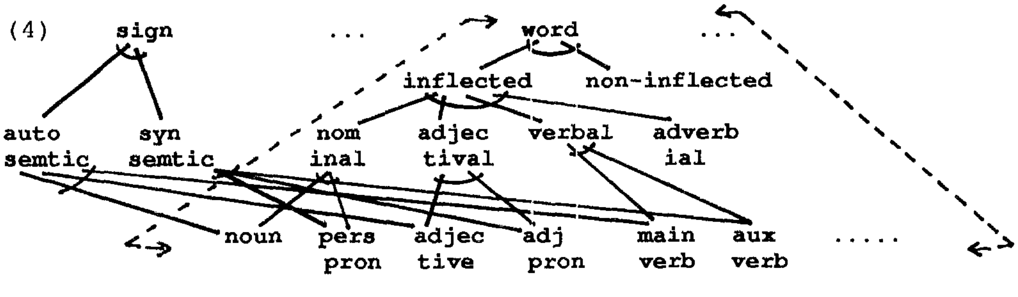
$$\left[\begin{array}{l} \text{phon: } 1 \\ \text{sem : } 2 \\ \text{constituents: } \left\{ \begin{bmatrix}\text{phon: } 3\\ \text{sem : } 4\end{bmatrix}, \begin{bmatrix}\text{phon: } 5\\ \text{sem : } 6\end{bmatrix}, \ldots, \begin{bmatrix}\text{phon: } 2n+1\\ \text{sem : } 2(n+1)\end{bmatrix}\right\} \end{array}\right]$$

$$\begin{array}{l} \& \ \text{phon\_consists\_of}(1,\{3,5,\ldots,2n+1\}) \\ \& \ \text{sem\_consists\_of}(2,\{4,6,\ldots,2(n+1)\}) \end{array}$$

The parallel, however, does not go much farther than that. In particular, the differences are that:
a. phrases consist of other phrases, but words (at least as a rule) do not consist of other words
b. for a word, at most one among its constituents, the base, is autosemantic (at least as a rule, again)

Taking into consideration this as well as still other factors, one can build up in more detail the top of the hierarchy of linguistic objects as in (4) (the "lexicon" part of the hierarchy being marked off by the interrupted line).

(4)



The class which is of particular interest in connection with the effort of removing the deficiencies of the lexical rules is the class **inflected**. By the very fact that this class - by definition - subsumes all inflected words and by parallel to (3), this class should be associated with the constrained feature structure (5).

```
(5) ┌                                                                   ┐
    │ phon:   1                                                         │
    │ prefixes:{[phon.   3],[phon:   5],...,[phon:   2n+1]}             │
    │ base:   2                                                         │
    │ suffixes:{[phon:   4],[phon:   6],...,[phon:   2k+2]}             │
    └                                                                   ┘

  &   phon_consists_of(1,{3,5,...,2n+1},2,{4,6,...,2k+2})
```

The definition of the relational constraint in (5) is actually the formal definition of inflection in the respective language[6]. In particular, this allows for expressing the part-of-speech independent regularities of inflection on one spot of the language model rather than repeatedly as with standard lexical rules (an example of such regularity might be the infix *-e-* in English words "*goes*" and "*potatoes*").

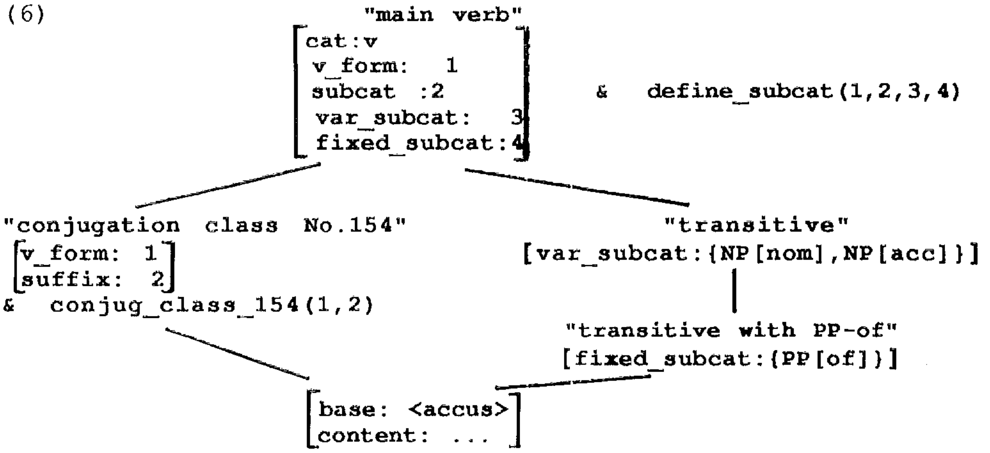
Having viewed the "top" part of the lexical hierarchy, let us turn our attention now to its bottom and middle.

The leaf elements of the hierarchy are lexemes - feature bundles encoding the most idiosyncratic information about a particular word. As a rule, a lexeme consists of the base ("root") of the word and of the semantic information associated with this base.

In the typical case, a lexeme of an inflected word has two immediate superclasses. These correspond to the cross-classification of (inflected) words according to, first, subcategorization requirements of the respective word (as in standard HPSG), and, second, according to the inflection class of the word. The respective subcategorization class then assigns to the word those subcategorization requirements which are not influenced by its morphology, as well as basic information about the nature of subcategorization requirements which do undergo changes due to the morphological form of the word. The inflection class, then, in the form of a relational constraint ties together the morphological features of the word with the respective affix(es), the constraint being in fact the (traditional) inflection table.

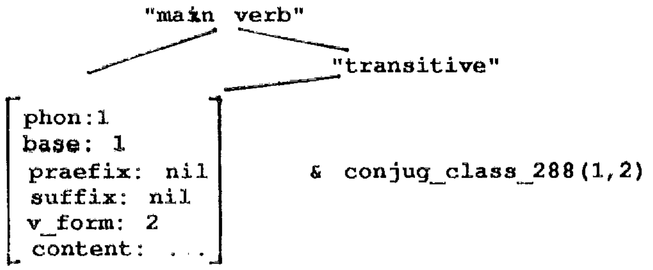
Finally, when these two classes meet upper in the hierarchy, the class where they meet defines, again via a relational constraint, the exact relation between the subcategorization and the morphological form of the word. The example (6) showing the verb "accuse" might clarify this.

(6)



In the case of totally irregularly inflected words, it would be of course inadequate to postulate a class expressing the inflection of this very one word only. The simplest solution in this case is to associate the respective relational constraint expressing inflection directly with the lexeme, cf. example (7) (technical variations concerning the base and affixes are possible, but insubstantial).

(7)



4. CONCLUSIONS

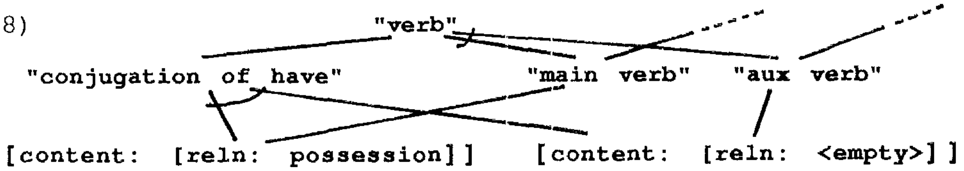
In this paper I tried to show how relational constraints can take over the bulk of (if not all) the work which is in standard HPSG assigned to the operation of lexical rules.

Such an approach has at least the following two major advantages:

1. it is much more straightforwardly related to the standard lexicographic and morphological usage, a matter which is of a remarkable theoretical and even greater practical importance
2. the disadvantages of the lexical rules approach to lexicon mentioned in the first part of this paper (either inherent procedurality of the description or inherent redundancies in the description) disappear with the replacement of the (also formally poorly understood) lexical rules by the standard machinery of relational constraints.

As a minor point in favour of the sketched approach it might be noted that it also easily avoids such counterintuitive stipulations as postulating the division of English verbs into classes "auxiliary", "main" and a singleton class "have-as-abstract-verb" (cf. Pollard and Sag,87, p.215) which was necessary due to the fact that both the main verb "*have*" and the auxiliary "*have*" are of the same - idiosyncratic - inflection class. All that is necessary to say on the given approach is that the two "*have*"s differ as to their contents and as to their subcategorization (i.e. they constitute two lexemes at the bottom of the lexical hierarchy), while simultaneously belonging to the same conjugation class, cf. (8).

(8)



As for applications, a large computational lexicon of Czech, based on ideas presented in this paper, is currently under preparation in the framework of a project aiming at the development of a prototype of a grammar-checker. This project is being carried out jointly with the Charles University in Prague.

BIBLIOGRAPHY

Flickinger D.: Lexical Rules in the Hierarchical Lexicon, PhD Dissertation, Stanford University 1987

Kathol A.: Passive Without Lexical Rules, paper presented at the workshop "HPSG and German Grammar", University of Saarland, August 1991

Pollard C. and I.A Sag: Information-based Syntax and Semantics, vol.1:Fundamentals, CSLI Lecture Notes Nr. 13, Stanford University 1987

Pollard C. and I.A Sag: Information-based Syntax and Semantics, vol.2, unpublished manuscript, summer 1993

Riehemann S.: Word Formation in Lexical Type Hierarchies - A Case Study of *bar*-Adjectives in German, MA Thesis, University of Tübingen 1993

Saussure F.de: Course de linguistique generale, Geneve 1915

FOOTNOTES

[1] *Cf. also the following complete quotation of Sect 8.3. Suggestions for further reading from (Pollard & Sag,87) p. 218:*

The key ideas in this chapter (i.e. "The Lexical Hierarchy and Lexical Rules") arise not from previous linguistic studies, but rather from algebraic approaches to datatype theory and the semantics of programming such as ...; similar ideas are widely employed in frame-based knowledge representation systems (see ...).

I just want to point out that hardly anyone would protest if "the key ideas ... arise not ONLY from previous linguistic studies, but ALSO from ...". However, the current wording explicitly excludes any reference to previous linguistic work.

[2] *In accordance with the tradition going back at least to de Saussure (Saussure,15), a sign is an object relating a (phonetic) form and a (semantic) content.*

[3] *In different terminology, "content vs. function".*

[4] *Cf. (Pollard & Sag,87), p.209 ff.*

[5] *Though (2) corresponds rather well to the traditional picture, it is worth remarking that a more simple hierarchy can be achieved by making phrase, word, word-base and morpheme subtypes of sign (in a division different from the autosemantic/synsemantic opposition) while making morph and smaller units sub-classes of non-sign. For our purpose, this point is, however, immaterial.*

[6] *Taken strictly, as it is, the definition does not take into consideration introflective languages. This, however, can be remedied simply by reformulating the constraint accordingly.*