

# Lexicalising Word Order Constraints: a feature-oriented specification

This paper presents a way in which a lexicalist HPSG grammar can handle discontinuous constituency in freer word order languages without recourse to an additional ('phenogrammatical') component responsible for constituent ordering, or linearisation, such as Reape's Word Order Domain (Reape, 1993; Reape, 1994) amongst others. Our key proposal is to incorporate into *lexical heads* the WOCs (Word Order Constraints) feature, which encodes the way that the head itself and its complements should be linearised in its projection. Although this setup would require some modifications to other components of the grammar to achieve the same coverage as the DOM-oriented framework, I believe it is worth pushing the boundary, since we could then describe the word order information *within* HPSG's standard TFS, retaining the process-neutrality and reusability of constraint-based grammars. An implementation of a parsing system based upon our formalism is also briefly described.

The main motivation for linearisation grammar comes from the fact that freer word order languages consistently display interleaving of constituents from different phrases, such as extra-clausal scrambling, extraposition, etc. (Reape, 1993; Yatabe, 1996; Chung, 1998). These constructions cannot be adequately handled within the context-free phrase structure rules (Suhre, 2000) and call for some non-CFG machinery for constituent ordering, such that (1) discontinuity/interleaving can be allowed and (2) appropriate LP constraints are enforced. For this purpose Reape introduces into HPSG some mechanisms outside the standard TFS logics and grammars, as well as some features, including DOM, the locus of linearisation. Firstly, Reape's 'default' combinatorial operation for a phrasal projection is *domain union*, which is essentially discontinuity-allowing but order-preserving merging of lists. Secondly, LP constraints are stated in a separate component of the theory, though applied locally in phrasal projections in the main. Coupled with the local 'UNIONED' feature for phrasal types, which indicates the intervenability of a phrase, the interaction of these mechanisms controls the way that constituents are linearised in DOMs.

What I find rather striking about this intricate framework is that although additional devices outside the standard TFS are invoked to enforce the word order related

constraints, most of their checks are carried out *locally*. One exception is naturally discontinuity, which cannot be determined locally, but even here the intervenability information originates from a local feature, UNIONED. Thus if all LP constraints were rendered locally applicable,<sup>1</sup> all that DOM would be doing then is percolation of intervenability information. Hence this opens the possibility to bring the relevant information back into TFS specifications, dispensing with the non-TFS components as well as a linearisation-specific feature like DOM. My proposal is to encode the LP and intervenability of a phrase in its lexical head, as properties that hold amongst the head itself and its complements (including subject). As a locus of percolation of this information, we use an enriched PHON feature attached with word order constraints.

The AVM on the next page shows how our lexically encoded WOCs are enforced in a projection with an example, the English VP *inform him of the news*, where one finds the WOC feature in its head verb, *inform*: here the binary ADJ (adjacency) and LP relations, represented with the infix notation  $\sim$  and  $\prec$  respectively, encode the constraints. We are assuming, somewhat simplistically, the following: first, this verb and its indirect object must be adjacent (*\*inform yesterday him of the news*); secondly, the indirect object must precede the of-PP (*\*inform of the news him*); and the general head-initial constraint of English VPs. Notice that there is no ADJ requirement between the indirect NP and of-PP. This allows an intervention between the two constituents, as in *'inform him yesterday of the news'*.

The job of constraining the phonological string is borne by the CONSTRS subfeature of the now compound PHON, into which the WOC values emanating from lexical heads are percolated. In this example the PHON feature of the VP encodes the legitimate word order patterns which obey the specified constraints in an underspecified manner. Also notice that the WOCs are passed up *all the way through* to the upper nodes into PHON|CONSTRS (notice that the NHD-

---

<sup>1</sup>This is a contentious proposition, however. The standard assumption is that there *are* LP constraints applicable beyond local domains, but I will argue that such cases are either manageable by slight extension of our treatment (e.g. type constraints) or disputable (e.g. locality-violating constraints).

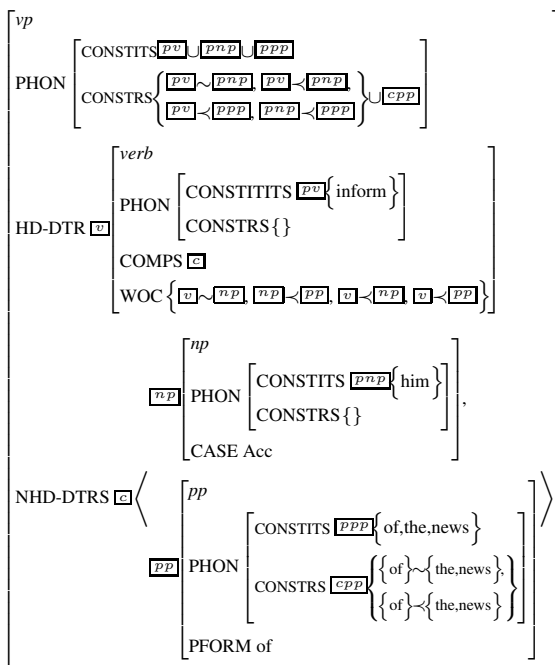


Figure: Example of verbal projection with WOC

DTR's PHON|CONSTRS values, in this example  $\langle \overline{pp} \rangle$ , are unioned), to ensure that all the WOCs can be enforced at any point of successive projections. For example, it is not at the stage of the head's projection but at some upper node that an ADJ constraint may be violated: the ungrammatical *inform yesterday him of the news* should be blocked in the application of  $VP \rightarrow VP$  adv, one step after the projection of *inform*. Yet, due to the cumulative inheritance of WOCs, the relevant WOC,  $\langle \overline{vp} \sim \overline{pp} \rangle$ , is also found in this upper node, blocking the above sequence there.

This is only an example with Head-Complement Structure, and hence extension is very much in order for general applicability, to Head-Adjunct Structure and Head-Specifier Structure at the very least. This is by no means a trivial exercise, however. I will therefore detail my envisaged strategy for this extension. Roughly speaking, my strategy is to introduce a generalised supertype *functor-word* for those lexical items with selectional properties (valence potential) and write the relevant WOCs there, as well as a generalised Schema.

Also, it is not just words but also some of their projections that should carry WOC information. For example, bar-level projections like nominals keep its SPR valence undischarged and hence should retain the WOC for this valence. This does not affect the fact that all the WOCs ultimately originate from lexical heads, but means the WOCs feature should be included in this level of projections and an appropriate feature-passing mechanism should be in place. Furthermore, now that all the word order information is in lexical heads, their appropriate subtyping in terms of WOCs would also be crucial for succinct and non-redundant description of word order patterns for individual languages.

One counterintuitive aspect of my proposal is that even those apparently 'clausal' word order constraints

– e.g. the AUX-V inversion for polar questions – would have to be stated in lexical heads – auxiliary in this case. As Kathol (2000) argues, the issue of clause types may be a matter that should not be determined on the level of the head a clause is a projection of but on the level of clause itself. However, I defer this question for later consideration, as our first priority is to examine whether our approach is technically extensible at all to other principal constructions, or Schemata.

The outline of our implementation of a parser using this formalism is then briefly discussed. It is a modified chart parser based on Gazdar and Mellish (1989) but enhanced by a non-CFG algorithm that can parse discontinuous scrambling relatively efficiently (Reape, 1991; Daniels, 2005). Now, the advantage of our formalism over the existent ones is that since word order information is now written inside a pure TFS grammar, we do not need a separate word-order component, e.g. in the form of phrase structure rules or LP statements: we can directly connect such a grammar to a parser, if its algorithm is appropriately re-designed. This re-designing involves firstly a mechanism whereby rules are generated dynamically during the parse using lexical heads' COMPS information. Secondly, word order information, also found in lexical heads, can be carried into edges and referred to in every dot move, to restrict search space. Due to this design, the parser needs to be bottom-up and head-corner, as the lexical head must be parsed first (van Noord, 1997).

## References

- C. Chung. 1998. Argument composition and long distance scrambling in Korean. In A. Hinrichs, T. Nakazawa, and A. Kathol, editors, *Complex Predicates in Nonderivational Syntax*. Academic Press.
- M. Daniels. 2005. *Generalized ID/LP Grammar*. Ph.D. thesis, Ohio State University.
- G. Gazdar and C. Mellish. 1989. *Natural Language Processing in Prolog*. Addison Wesley.
- A. Kathol. 2000. *Linear Syntax*. OUP.
- M. Reape. 1991. Parsing bounded discontinuous constituents: Generalisation of some common algorithms. *DIANA Report, Edinburgh University*.
- M. Reape. 1993. *A Formal Theory of Word Order*. Ph.D. thesis, Edinburgh University.
- M. Reape. 1994. Domain union and word order variation in german. In *German in Head-Driven Phrase Structure Grammar*.
- O. Suhre. 2000. Computational aspects of a grammar formalism for languages with freer word order. Diplomarbeit, Eberhard-Karls-Universität Tübingen.
- G. van Noord. 1997. An efficient implementation of the head-corner parser. *Computational Linguistics*.
- S. Yatabe. 1996. Long-distance scrambling via partial compaction. In M. Koizumi, M. Oishi, and U. Sauerland, editors, *Formal Approaches to Japanese Linguistics 2 (MIT Working Papers in Linguistics 29)*. MIT Press, Cambridge, Mass.