### SYNTACTIC DESCRIPTION OF FREE WORD ORDER LANGUAGES

### Tania Avgustinova & Karel Oliva\*

Lingustic Modelling Laboratory,
Coordination Centre
for Computer Science and Computer Technology,
Bulgarian Academy of Sciences,
acad. G. Bonchev st. bl. 25A,
BG - 1113 Sofia,
Bulgaria

#### Abstract:

A framework for the description of syntactic structures of free word order languages is proposed, based on combination of intuitions underlying immediate constituent description, dependency description and communicative dynamism. The combined approach is compared to its sources and shown superior in descriptive power, esp. in the area of free intermixing of (any number of) adjuncts with complements and in coordination. Close resemblance to two other recent approaches is pointed out.

### 1. Syntactic Structures for Free Word Order

The absolute majority of current linguistic frameworks characterize syntactic structures of natural languages in predominantly static terms, paying only minimal or no attention to the communicative function of language and its reflection in the process of uttering/understanding (generation/parsing) sentences.

In the case of generative frameworks based on immediate constituent approaches to language description, this can be ascribed (at least partly) to the fact that many of them (GB, LFG, GPSG, TAGS, to mention the most widespread ones) were created primarily for the sake of description of English, a highly configurational language in which the impact of communicative functions on syntactic structure is quite limited (at least in comparison with the so-called free-word-order languages - henceforth FWOLs).

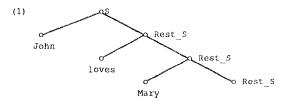
The frameworks based on the dependency syntax (e.g., the "Meaning-Text" model of Mel'chuk and Apresyan, the "Functional Generative Description" of Sgall et al., the "Word Grammar" of Hudson), on the other hand, by the very principle separate linguistic structures from the process of generation/parsing so sharply that even if any reflection of the communicative process is present in the generation/parsing procedures, it gets lost in the resulting structures and has to be added there (if needed) more or less artificially, e.g., in the form of different indices (cf. the structures in Sgall et al, 1986).

This lack of reflection of communicative aspects of language in the (syntactic) structures, together with still other features of the abovementioned frameworks (such as, for the immediate constituent based approaches, the incapability of the standard "S --> NP VP" approach to describe, e.g., the "Object-Subject-Verb" constituent order, or this obstacle overcome in some way, the problems connected with free intermixing of any number of free adjuncts among the complements, and, for the dependency based approaches, the problems involved in capturing even quite simple instances of coordination), make relatively profound adjustments in the existing frameworks or development of a new one a necessary and highly important task if a language involving broad impact of the communicative aspects on its syntactic structures (such as the FWOLs) has to be described formally in such a way that the description can be directly implemented on a computer and function as a generator or a parser.

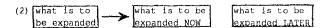
The easiest way how to overcome the difficulties connected with the current frameworks and to achieve the abovementioned goal of creating a framework suit-

able for a reasonable description of FWOLs as well as for an easy and efficient computer implementation seemed to be to augment the immodiate constituent based nontransformational approaches (which are easier to implement due to the clearcut correspondence between the rules of the grammar and the structures they generate) with the intuitions contained in more traditional descriptions of the FWOLs as well as in the descriptions of functional sentence perspective and communicative dynamism (Firbas, 1971, 1975; Sgall et al.1973).

In the unmarked case, the scale of communicative dynamism allows for splitting the sentence or any of its parts on the level of the "main" constituents (such as Subject, Object, different verbal Adjuncts etc.) at any moment into two parts, the first consisting of the constituent being processed (uttered, expanded) at the very moment, i.e. the currently least dynamic constituent, and the second one consisting of the "rest" of the sentence, i.e. of all the constituents more dynamic than the currently processed one. This results in a non-transformational account of syntactic structures, in the form of binary right-branching trees (if the division sketched above is broadened to all constituent types used in the description). An example of the structure for the notorious sentence "John loves Mary" is given in (1). (Mind the rightmost "Rest\_S" nonterminal dominating an empty string: "nothing more is to be uttered" in the sentence, "nothing is more dynamic" than "Mary".)



On such an approach, both the generation of all possible constituent orders and free intermixing of any number of adjuncts between any two complements is guaranteed for FWOLs, and this without using the Kleene star in the rules, metarules generating an infinite number of rules or any other way of using (explicitly or implicitly) an infinite rule set. Just on the contrary, the approach results in a drastic simplification of the number and shape of rules needed: one gross rule scheme (2) is sufficient for the whole grammar:



In this scheme, the second constituent on the right hand side is always the phrasal head; according to the nature of the left daughter, the rule set can be further factorized into the following subsets reflecting the classical linguistic wisdom: rules expanding the lexical head, rules expanding a complement, rules expanding a free adjunct, rules expanding an extraposed constituent, rules expanding a member of a coordinated structure, rules expanding minor categories (conjunctions, particles etc.). Such a division is im-

portant not only because it brings along some more purity and perspicuity, but also because it allows for a straightforward implementation of different feature inheritance principles of the framework (such as the Head Feature Principle, Subcategorization Principle etc.) in the computer variant of the grammar; on a reasonable formal notation of the grammar rules allowing marking off the type of the rule as the property of the rule itself, it is possible to bound the application of the principles to the whole rule types rather than to each rule separately, as the case often is in many current parsers (e.g., for a head daughter in a rule, it is not necessary to stipulate explicitly the sharing of its head features with the mother, since this is provided for by listing the rule in the class of head expanding rules).

#### 2. Relation to Other Syntactic Frameworks

The proposed structures might seem rather unconventional at first glance; however, their relation to structures used in more usual syntactic frameworks can be shown to be quite straightforward in simple cases. All what is needed to obtain dependency trees is to factorize the set of nodes of the described structures by all bar projections of a single terminal node. An X-structure can be obtained by factorizing the set of nodes by projections of the same bar-level of a single terminal node.

In more complicated cases, however, the factorizations sketched above cannot be performed. Exactly in these cases, the structures proposed rank better in describing at least the following phenomena of FWOLs:

- in relation to dependency syntax (traditionally used for description of FWOLs), first of all in describing coordination, but also the so-called non-projective constructions (e.g., unbounded dependencies) as well as cases where contact position of certain words or constituents is required or positions are to be strictly fixed even in FWOLs (e.g., the Wackernagel's position of clitics), which both is difficult to achieve in dependency descriptions if non-projective constructions are allowed to occur since these interfere with the "basic" projective ordering generated

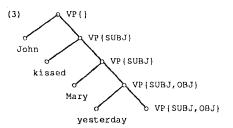
- in relation to standard variants of X-syntax, the approach adopted solves the problems with the position of subject, with free intermixing of complements and adjuncts and, in addition, it is able to cope with certain cases of "heavy" coordination (see below) on a context-free basis

## 3. Subcategorization and Coordination

Generally speaking, the intuitions (as opposed to the formalism) standing behind the framework are very close to (if not the same as) those supporting dependency approaches (certainly more so than to the intuitions of the majority of current immediate constituent approaches, cf., e.g., the nonexistence of the "NP/VP" division of a sentence), but the structures developed for the formal incarnation of these intuitions have by far more descriptive power than the standard formalizations proposed for the dependency approaches. This extra power (even in comparison with the standard X-approaches) stems mainly from the increased number of nonterminal symbols: the greater number of nonterminals allows for a more subtle structuring of the terminal string.

The crucial point of this refinement of structural information is the one concerning subcategorization of phrases. In accordance with the treatment of subcategorization in HPSG and other frameworks, subcategorization can be informally viewed as the number and shape of constituents to be added to a particular phrase for it to become a saturated projection of its lexical head (e.g., for a VP, this subcategorization is the number and shape of constituents to be added for the VP to become a sentence; thus, a sentence is just an alias for a VP with empty subcate-

gorization). In the example (3), it is important to notice the "sharing" of the subcategorization requirements (depicted schematically as sets of subcategorized-for elements associated with the nonterminal nodes of the structure) between the lexical head of the sentence (the verb) and its rightmost phrasal projection, as well as the stepwise right-to-left reduction of the subcategorization requirements of the VP's, and also the fact that the expansion of a lexical head or a free adjunct does not affect the subcategorization.



As mentioned above, this "extra descriptive power" can be made use of for description of (among other) certain "heavy" coordinations. The instances we have in mind are "Right Node Raising" and "Across the Board" coordinations exemplified in (4) and (5), respectively.

- (4) Mary baked and John ate an apple pie.
- (5) the pie Mary baked and John ate

Before presenting the treatment proper, two matters have to be pointed out:

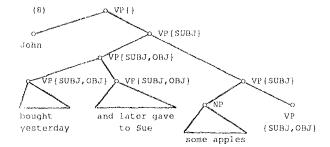
- first, in FWOLs "Right Node Raising" and "Across the Board" are exactly the same cases of coordinative constructions (due to the free-word-order, the position of the "extracted" constituent plays no syntactic role)

- second, the grammaticality of other cases of coordination can be order dependent, even in FWOLs: typical case is "Gapping" (cf. the contrast shown for English in (6)a,b but holding also in (at least) Bulgarian, Czech, Polish, Russian and Slovak), somewhat unclear is the situation with "Non-Constituent Coordination", where speakers of the abovementioned languages seem to have different opinions about the grammaticality of the respective counterparts of (7)b.

- (6) a. John loves Mary and Jim Sue.
  - b. \* Jim Sue and John loves Mary.
- (7) a. John gave a book to Mary and a bunch of flowers to Sue.
  - b. ?? A book to Mary and a bunch of flowers to Sue John gave.

This corroborates the view that (6)a,(7)a are instances of some extragrammatical communicative processes (i.e. processes not reflected in the grammar of the language - such as the tendency to avoid uttering identical parts of coordinated structures etc.) rather than true cases of "coordinated predication" which seems to be the case with "Right Node Raising" and "Across the Board".

The treatment of "Right Node Raising" and "Across the Board" relies fully on the refinement of subcategorization into the increased number of nodes of the structure, but on the other hand it does not require any augmentation of the coordination mechanisms of the framework, the only coordination rule being the "coordination of likes". The approach even allows for description of constructions where "Right Node Raising" and "Across the Board" cooccur. The structure assigned to such cases is given in (8) (the terminal string of which is quite probably no good English, but translations into the FWOLs tested are considered fully acceptable).



Note that the term "lexical head" should be taken with a grain of salt for the VP in (8) (as well as for all other coordinative constructions) — this is, however, a purely terminological matter which can be coped with easily in a fullfledged exposition of the theory and has no bearing on its validity. Similarly, for coordinations consisting of more than two members, the exemplified construction would not conform to the scheme from (2); this is again due to simplifications adopted for the purpose of the current presentation, and in a more detailed exposition coordinative constructions would be also expanded in the "one-member-at-a-time" manner.

It might be also interesting to observe that "Gapping" and "Non-Constituent Coordination" cannot be treated in the framework, unless it is augmented with some "deletion" processes operating on the structures generated by the context-free base.

### 4. Conclusions

The framework presented in this paper was created in the course of preparatory work for an implementation of a parser for Bulgarian, a free-word-order language from the Slavonic group. The main idea standing behind the structures as presented was merging the insights concerning communicative dynamism contained in the works of linguists of the Prague School with the intuitions underlying the dependency descriptions of language, and implementing the whole in an immediate constituent based formalism. The result might seem rather unorthodox in many respects, but the deviations contained can be sanctioned by at least two remarkable advantages of the framework proposed over the more standard approaches:

- first, on the theoretical side, the use of (increased number of) nonterminal symbols makes the framework superior in descriptive (and, let us hope, also explanatory) adequacy concerning such phenomena as coordination and the so-called "non-projective" constructions in FWOLs, while simultaneously keeping the generative power on the context-free level (for space limitations, no examples of the non-projective constructions were given, but due to the presence of nonterminals in the structures, they can be treated in the way broadly used in other context-free based immediate constituent approaches, e.g., by the "SLASH" mechanism of GPSG or HPSG)

- second, from a more practical viewpoint, while the overall approach allows for keeping virtually all linguistic intuitions contained in the dependency approaches traditionally used for the description of FWOLs, the formalization adopted allows for using a generative grammar (i.e. a set of declaritively stated rewriting rules) for the description of the language. which in turn quarantees a clear correspondence between the structures of the language and the mechanism that generates them. This correspondence is never so straightforward in "pure" dependency approaches which, as a rule, use some exclusively procedural machinery rather remote to the structures that are to be generated or parsed by it (cf. the framework described in Apresyan et al,1989 for implementation of the Meaning-Text model or the "transducing automata" of Functional Generative Description described in Sgall et al, 1969). Needless to add, the possibility of keeping the linguistic information in a separated declarative format (such as a set of rewriting rules) makes the job of creating and, mainly, debugging the generator/parser easier (though, of course, not at all easy).

Partly also as an indirect support for the formalism presented, it is further necessary to say that structures in many respects similar were proposed independently in (Uszkoreit, 1986) for the description of "complex fronting" in verb-final German clauses and in (Gunji, 1987) in the framework of JPSG (which, however, advocates the existence of the "NP/VP" division of the sentence and also, similarly to Uszkoreit, heavily relies on the fact than Japanese is a verb-final language). What is interesting and, hopefully, also important to observe is the fact that this happened in spite of the fact that both Uszkoreit and Gunji started from intuitions different from the ones incorporated in the presented framework (and also different from each other) and aimed at description of phenomena often marginal, remote or even nonexistent in the languages on the base of which the currently presented framework was developed and for the description of which it is intended to serve.

# Bibliography:

Apresyan J.D., I.M. Boguslavskij, L.L. Iomdin, A.V. Lazurskij, N.V. Pertsov, V.Z. Sannikov and L.L. Tsinman: The Linguistic Support for the System ETAP-2 (in Russian), Nauka, Moskva 1989

Firbas J.: On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective, in: Brno Studies in English 7, University of Jan Evangolista Purkyne (formerly and now again University of Tomas Garrigue Masaryk), Brno 1971, pp.12-47

Firbas J.: On the Thematic and the Non-Thematic Section of the Sentence, in: H.Ringbom (ed.): Style and Text - Studies Presented to N.E. Enkvist, Stockholm 1975

Gunji T.: Japanese Phrase Structure Grammar, Reidel, Dordrecht 1987

Pollard C. and I.Sag: Information Based Syntax and Semantics, vol.1: Fundamentals, CSLI Lecture Notes No. 13, CSLI, Stanford, California 1987

Sag I., G. Gazdar, T. Wasow and S. Weisler: Coordination and How to Distinguish Categories, CSLI Report No.3, CSLI, Stanford, California 1984

Sgall P.,L.Nebeský, A.Goralčíková and E.Hajičová: A Functional Approach to Syntax in Generative Description of Language, Elsevier, New York 1969

Sgall P.,E.Hajičová and E.Benešová: Topic, Focus and Generative Semantics, Kronberg, Taunus 1973

Sgall P.,E.Hajičová and J.Panevová: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Academia, Prague 1986

Shieber S., S.Stucky, H.Uszkoreit and J.Robinson: Formal Constraints on Metarules, in: Proceedings of the 21st Annual Meeting of the ACL, Cambridge, Massachusetts 1983

Uszkoreit H.: Linear Precedence in Discontinuous Constituents: Complex Fronting in German, CSLI Report No.47, CSLI, Stanford, California 1986

Lehrstuhl für Computerlinguistik Universität des Saarlandes Im Stadtwald D-6600 Saarbrücken (West) Germany

<sup>\*</sup> since 1st April 1990: