TESTING THE PROJECTIVITY HYPOTHESIS

Vladimir Pericliev
Mathematical Linguistics Dpt
Institute of Mathematics with Comp Centre
1113 Sofia, bl.8, BULGARIA

Ilarion Ilarionov
Mathematics Dpt
Higher Inst of Eng & Building
Sofia, BULGARIA

ABSTRACT

The empirical validity of the projectivity hypothesis for Bulgarian is tested. It is shown that the justification of the hypothesis presented for other languages suffers serious methodological deficiencies. Our automated testing, designed to evade such deficiencies, yielded results falsifying the hypothesis for Bulgarian: the non-projective constructions studied were in fact grammatical rather than ungrammatical, as implied by the projectivity thesis. Despite this, the projectivity/non-projectivity distinction itself has to be retained in Bulgarian syntax and, with some provisions, in the systems for automatic processing as well.

1 THE PROJECTIVITY HYPOTHESIS

Projectivity is word order constraint in dependency grammars, which is analogous to continuous constituency within phrase-structure systems. In a projective sentence, between two words connected by a dependency arc only such words can be positioned which are governed (directly or indirectly) by one of these words. Or, in other words, a sentence is projective in case there are no intersections between arcs and projections in its dependency tree diagram. Thus, for instance, sentence (1) is projective, whereas sentence (2) is non-projective:

We might note that sentence (2) is ungrammatical.

The projectivity hypothesis, originally propounded by Lecerf (cf. e.g. Lecerf 1960) and later gaining wide acceptence, amounts to the following: Natural languages are projective in the sense that the non-projective constructions in them are ungrammatical. And this has an important consequence. Thus, taking into account the self-evident fact that ungrammatical phrases do not occur in texts, in the processing of texts we can rule out from consideration the non-projective parses on the basis of ungrammaticality. Projectivity thus serves as a filtering device, shown further to be of extremely powerful nature (op.cit.).

To estimate the usefulness of the projectivity hypothesis for each particular language requires the conduct of extensive empirical testings. On the basis of statistical accounts from inspection of texts French was reported by Lecerf to be almost 100% projective. The same would be true, according to him, for other languages like German, Italian, Dutch etc., although the material available (at the time) was not sufficient for statistical processing. English is also believed to be a projective language: in 30 000 phrases only two non-projective ones were found (Harper and Hays 1959); in Kareva (1965) somewhat different, but still result in the same vein was obtained (using different notation): from 10 000 phrases of connected text 620 were found to be non-projective.

Such investigations can be seen to be bound together by their approach to the testing of the projectivity hypothesis: texts are explored and statistical accounts are made of the correlation between projective and non-projective phrases. The very rare occurrence in such texts of non-projective sentences is interpreted as a confirming evidence. Such studies represent what we shall furtheron refer to as "the textual approach to the testing of the projectivity hypothesis" (or simply, "the textual approach").

2 DEFICIENCIES OF THE TEXTUAL APPROACH

The textual approach, in addition to the fact that it involves the tedious task of inspection of thousands of sentences, suffers serious methodological shortcomings which can be summarized as follows:

(i) Irrelevancy of data. The data the textual approach presents in justification of the hypothesis is, strictly speaking, irrelevant. Knowing that non-projective phrases do not occur in texts, naturally, gives us no formal right to infer that such phrases are ungrammatical as well.

(ii) Insufficiency of data. The data provided by this approach is insufficient to justify even a weaker claim to the effect that non-projective structures do not occur in texts. To justify this latter claim further steps in addition to direct inspection of certain (immaterially how large) corpora of texts should be made. In particular, a justifiable justification would have to involve both further factual confirmation (e.g. demonstration that predictions from the hypothesis in fact comply with actual data) and "systematic" confirmation (demonstration that the hypothesis is consistent with other linguistic principles, facts, etc.) (cf. e.g. Botha 1981: Ch.9; also § 3 below).

(iii) Heuristic futility. The textual approach is heuristically futile in the sense that, being confined to a mere registration of non-projective constructions within specific texts, we have no way of knowing whether the structures encountered (if some are at all encountered) are all the non-projective structures in a given language, and if not, how many more are there, and which exactly they are.

3 TESTING THE PROJECTIVITY HYPOTHESIS FOR BULGARIAN

The considerations given in § 2 seriously undermine the credulousness of the results obtained for other languages following the textual approach. What was important for our investigation however was to evade these methodological deficiencies in the study of Bulgarian. Accordingly, we had to address not texts, but rather what we had to do was to generate all logically admissible non-projective structures

in Bulgarian, and then inspect them for grammaticality.

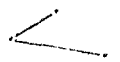
It was appropriate to accomplish our testing in two phases: preliminary (manual) testing, in which the plausibility of the projectivity hypothesis was to be estimated for the Bulgarian language, and testing proper (automated testing), in which the non-projective structures in Bulgarian were to be automatically generated, and then checked for grammaticality/ungrammaticality.

3.1 Preliminary testing

The preliminary (manual) testing comprised: (i) factual testing, and (ii) systematic testing (cf. § 2 (ii)).

In the factual testing it was inspected whether certain predictions from the projectivity hypothesis are consistent with actual data. That is, we take an arbitrary non-projective situation, say, a situation of the form:

(3)



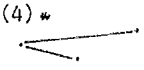
X1 X2 X3

and then, substituting X1, X2, and X3 with appropriate word classes, check whether the resultant construction is well-formed in Bulgarian or not.

In the systematic testing it was inspected whether the projectivity hypothesis in fact fits in with other known word order principles, rules, etc. (of universal or language-specific nature).

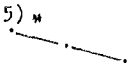
By way of illustration, consider the generally recognized universal principle: In all languages there exist classes of words occupying a rigidly fixed position in the sentence (the particular words and positions of course being language-specific). On inspection, this principle turns out to contradict the projectivity hypothesis. This is so, since such situations may occur in which this fixed position of certain words leads to non-projectivity. Thus, one manifestation of this principle in Bulgarian syntax is reflected in the fact that the verb sam 'be' never occurs in sentence-initial or sentence-final position. Now, assume that we have a three-word sentence containing be in which moreover: (a) be governs another word, X2, X2 being positioned to the left or right of be; and (b) X2 governs X3, X3 being obligatorily positioned to the right of X2. This being the case, three structures are theoretically admissible, two projective and one non-projective:

(4) *



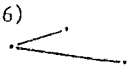
X2 X3 be

(5) *



be X2 X3

(6)



X2 be X3

However, structures (4) and (5) will be ungrammatical, as predicted by the principle mentioned (notice the position of be). This latter fact, in turn, predicts the grammaticality of the non-projective structure (6) (knowing of course that there is nothing to forbid in Bulgarian the occurrence of three-word sentences containing be). As another illustration, this mode of testing would have to lead to the discovery of non-projectivities of the type: "A procedure is discussed which..." in English which are due to the sentence-initial position of the subject in the English sentence.

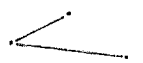
In summary, the results obtained from our preliminary testing showed the implausibility of the hypothesis for Bulgarian: we easily found numerous and diverse kinds of counterexamples to it. We further noticed that the counterexamples belonged, informally speaking, to two stylistic layers which could be labeled as stylistically marked and stylistically unmarked.

3.2 Testing proper

As a next step in our investigation, the non-projective constructions in Bulgarian had to be generated, and then assessed for well-formedness. More specifically, non-projectivity in triples and quadruples was to be examined (in so far as non-projectivity in more than four-words constructions is reducible to triples or quadruples).

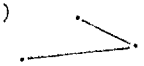
In triples, there are two possible non-projective situations, viz. (the mirror-images):

(7)



X1 X2 X3

(8)



X3 X2 X1

In quadruples, these non-projective situations are 30 in number. That is, the total number of non-projective situations is 32. The number and content of constructions in Bulgarian conforming to these situations will be language-specific, i.e. it will depend on the specific Bulgarian word classes and the possibilities for their mutual positioning. E.g. the constructions conforming to situation (7) will be the set of all triples X1 X2 X3 such that X2 governs X1, X1 being positioned to the left of X2, and X1 governs X3, X3 being positioned to the right of X1.

Then, a program was written in BASIC implemented on the Bulgarian microcomputer "Pravetz" (a machine compatible with Apple II) which generated the constructions conforming to the non-projective situations. The input to the program was a fragment of the dependency grammar for Bulgarian given in Pericliev 1983. In particular, 30 rules were stored, each rule consiting of a pair of word classes, a master and a slave, and their mutual position(s). For obvious reasons the rules were not arbitrarily chosen, but rather it was required that they be maximally diverse in syntactic nature. That is, they included pairs of notional and/or functional words (particles, pronouns/adverbs introducing clauses, paired conjunctions, duplicating parts of the sentence, etc.). The generated constructions were then inspected for well-formedness.

The results from our experiment may be summarized as follows. From about 300 non-projective constructions generated, approximately 15% turned out to be ungrammatical. The remaining part of the constructions were grammatical. As already expected, they could be classed into two groups according to their stylistic value: stylistically unmarked and stylistically marked constructions.

The unmarked constructions, informally speaking, included diverse kinds of structures: some questions (with the question particle li 'do' or with li together with a notional questioning word), some exclamatory sentences (with structure of questions), di-

fferent complex sentences (a word belonging to some subordinate clause, most often objective and attributive clause, is positioned somewhere in the main clause), sentences containing clitics (be, short possessive and dative pronouns, etc.), various constructions with "strongly linked" parts (paired conjunctions/particles, duplicating parts, Bulgarian equivalents of more ... than, such ... that , etc.) and many others. The ratio between stylistically unmarked and stylistically marked constructions was about 1:5.

4 DISCUSSION OF RESULTS

In principle, a hypothesis, in empirical sciences such as linguistics, may be said to be: (a) absolutely true (i.e. true without exceptions), (b) on the whole true (i.e. true, but with certain exceptions), and (c) false.

The projectivity hypothesis would have been usable as a filtering device (in Bulgarian) if it fell under cases (a) or (b), case (b) presupposing further that there is a list available of all exceptions. The results obtained however unambiguously class it under case (c). Indeed, the great majority of non-projective constructions in Bulgarian are well-formed rather than unacceptable, as implied by the hypothesis. This seems to be the only correct conclusion, despite the fact that the results themselves should not be considered as absolutely final, this being due to the following circumstances: (1) We did not in fact inspect literally all admissible non-projectivities but only those that were obtainable from the fragment of grammar stored for the experiment; and (2) the presense/absense of projectivity significantly depends on the conventions chosen for dependency arcs distribution. Still, our investigation on the whole is to be viewed as sufficiently reliable (another experimental testing in which a larger grammar was used and slightly different notation where linguistically justified gave similar results).

Another point deserves special attention. Despite the fact that non-projective structures in Bulgarian are grammatical, the prevalent part of them are "stylistically marked". Whatever that means, the latter circumstance implies that, quite probably, such constructions will not occur in some kinds of texts at least. Our chances to pinpoint such texts to a great extent depend on our understanding of the constructions in question (the label "stylistic markedness" in itself is no great progress in this direction). So we focused our attention on these constructions, having already completed the experimental testing described. The preliminary results of this latter study were quite interesting: non-projectivity in fact turned out to finely describe sentences with a special type of logical emphasis in Bulgarian, characterized by:

(i) presense of both non-projectivity and contrastive stress (CS) (in contrast to other emphatic sentences in Bulgarian, having a CS and "normal", projective word order);
(ii) presense of CS only on one of the words (immaterially which) from the non-projective "interval" (an "interval" is constituted by a pair of words connected by an arc such that intersects with a projection); the non-projective constructions in this sense contrast with other emphatic constructions with projective word order , which can take a CS on any word in them;
(iii) having synonymous constructions of two types: with a projective word order and CS, or such analogous to the English cleft sentences ("It is he who ..., where the underlined words are connected by an arc); perhaps, not by coincidence the latter constructions, both in Bulgarian and in English, are non-projective;
(iv) having a relation to the topic-comment distinction; more specifically, one of the words from the non-projective interval (immaterially which) is necessarily the comment of the sentence (though, naturally, not each comment requires non-projectivity).

These findings rehabilitate the projectivity/ /non-projectivity distinction (thogh not the hypothesis) for Bulgarian syntactic theory: non-projectivity happens to be the word order part of a very general mechanism for formation of sentences with logical emphasis. Thus, though Bulgarian is generally characterized as a free word order language, sentences with logical emphasis are not with a "free", but rather with a precisely specified word order coinciding with non-projectivity.

As to the need to retain the above distinction in syntactic processors of Bulgarian, things will be determined by the concrete applied goals: some small applications can certainly ignore non-projective structures (their emphatic part), whereas robust systems cannot do without them. That is, the former systems would have to keep the distinction and the latter reject it. Still, it will not be surprising if the distinction might be usefully accomodated even in the latter case. However, this needs further investigations, mostly in the line of textual linguistics: the study of context, topic/comment, etc. One is reminded in this train of thought of the remark of the philosopher K. Popper to the effect that a scientific investigation begins and ends with problems.

REFERENCES

Botha R., 1981, The Conduct of Linguistic Inquiry, Mouton: The Hague.

Harper K. and D. Hays, 1959, The use of machines in the construction of a grammar and computer program for structural analysis, Rapport UNESCO.

Kareva N., 1965, A classification of non-projective constructions, Scientific & Technical Information, No.4 (in Russian).

Lecerf Y., 1960, Programme des conflits, modèle des conflits, Traduction automatique, 1, No.4.

Pericliev V., 1983, Syntactic Ambiguity in Bulgarian and in English (Ph.D. Dissertation; in Bulgarian).