TOWARD A PARSING METHOD FOR FREE WORD ORDER LANGUAGES[x]

Janusz S. Bień, Stanisław Szpakowicz

Institute of Informatics, Warsaw University
P.O.B. 1210, 00-901 Warszawa, Poland

1. Introduction

"Free word order" is a traditional term that should not be taken literally. However, we shall retain the term for its conciseness.

Formal descriptions of syntax have been usually based either on the immediate constituents or on the dependency philosophy. Neither of them seems directly applicable to free word order languages. The intertwining phrases cannot be described naturally by IC rules. Some coordinate constructions are difficult to describe by means of dependency relations. In our opinion, parsers for free word order languages should not be based on the methods developed within the IC framework. Scarce experiments with parsers based on the dependency formalism, eg. /5/, do not seem promising. Therefore, we decided to take a fresh start and to attack the problem by reanalysing the basic notions of syntax and parsing. We focus our attention on those formal aspects of a language system which might be most useful for automatic text processing. We assume that the morphological level is described along the lines of /2/.

---

x) This paper is an extended abstract of /3/.

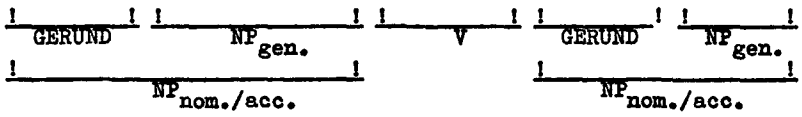
2. The Notion of Syntax

In this paper, we understand syntax as the domain of formal relations between words, i.e. roughly as so-called surface syntax. We define the notion using a morphology-based criterion, described below.

The outcome of morphological analysis can be ambiguous for an isolated word. In most situations, however, the morphological features of a word are uniquely determined by some formal properties of its context.

Sometimes the ambiguity remains, as in the following sentence:

Opóźnienie brygad piecowych spowodowało potępienie wuja Jana.



There are five independent ambiguities in this sentence, yielding 32 coherent readings. Two of them are due to the neutralization of agent/patient function during nominalisation. For example, "potępienie x" means "disapproval of x" (either "x disapproves y" or "y disapproves x"); such an ambiguity can be resolved only by examining the meaning of a given phrase, so we call it semantic one.

The next ambiguity occurs in the phrase "wuja Jana", that means either "uncle John"$_{gen.}$ or "John's uncle"$_{gen.}$. Here we can see two kinds of syntactic relations: case agreement (the former interpretation) or government (the latter one), which both require "Jana" to be in genitive case. Such an ambiguity we consider as purely syntactic one.

In the phrase "brygad piecowych" we can discern either case agreement ("piecowych"$_{gen.}$ is then an adjective) or government ("piecowych"$_{gen.}$ is then a noun). Here, the elimination of morphological homonymy gives rise to alternative constructions, thus increasing the syntactic ambiguity.

The last ambiguity stems from the nominative/accusative neutralization both of a virtual subject and a virtual object of the sentence. It suffices to assign a syntactic function to one of them; the function of the other and the morphological characteristics of both of them will be fully determined.

The example demonstrates how certain relations between sentence components allow to disambiguate the morphological properties of individual words without resorting to their meanings. In our approach, these relations constitute the level of syntax /3/.

3. Syntactic Relations

Syntactic relations (eg. agreement, government) consist in matching syntactic properties (eg. case, gender) of respectiveu units. The basic unit is a morphological word /2/.

By the syntactic structure of a sentence we understand some explicit representation of all the syntactic relations between its components, usually - a graph. Such a graph need not necessarily be connected. For example, some modifiers are linked to their heads only by semantic relations and not by syntactic ones. Similarly, some elliptic sentences may have disconnected syntactic representations.

4. The Notion of Parsing

We understand parsing as a process of establishing all syntactic structures of a given text. Although such structures are rather unsophisticated, they are practically very important for low-level text processing.

In search of an adequate parsing method, we found the idea of Marcus /4/ most appealing. He claims that natural languages are designed to be deterministically parsed from left to right and that writing a grammar should consist in finding out local clues which enable the parser to select properly what to do next. This idea seems even more advantageous for free word order languages. Rich inflection makes

the local clues much more explicit and the parser's expectations more precise. Besides, such an organisation of the parsing process is compatible with the resource control hypothesis /1/ which is hoped to account for semantic implications of free word order.

5. Conclusion

As a practical consequence of the considerations given above, we adopt the following research program. As a starting point we take the existing IC-based syntactic description of Polish sentences with neutral word order /6/, consisting of about 500 rules (some parts of it have been rewritten in greater detail /7/, with the number of rules increasing 5-10 times). We are going to restructure the description to obtain an index of expectations related to each syntactic unit. We shall incorporate the clues, thus obtained, into some Marcus-style parsing strategy. We expect that it will lead to an efficient and linguistically sound parser for Polish.

References

/1/ Bień J.S.: A Preliminary Study on Linguistic Implications of Resource Control in Natural Language Understanding. ISSCO Working Paper 44, Geneve 1980.

/2/ Bień J.S., Saloni Z.: The notion of morphological word and its application to the description of Polish inflection (preliminary version) /in Polish/. Prace Filologiczne XXXI, to appear.

/3/ Bień J.S., Szpakowicz S.: Toward a Parsing Method for Free Word Order Languages. In: Papers in Computational Linguistics II. IInf UW Reports, to appear.

/4/ Marcus M.P.: A Theory of Syntactic Recognition for Natural Language. MIT Press 1980.

/5/ Panevová J., Sgall P.: On Some Issues of Syntactic Analysis of Czech. In: The Prague Bulletin of Mathematical Linguistics 34, 1980, 21-32.

/6/ Szpakowicz S.: Formal syntactic description of Polish sentences /in Polish/. Wydawnictwa Uniwersytetu Warszawskiego, in press.
/7/ Szpakowicz S., Świdziński M.: An outline of sentence schemes classification in contemporary written Polish /in Polish/. Studia gramatyczne V, Wrocław, to appear.