PARSING FREE WORD ORDER LANGUAGES IN PROLOG

Janusz Stanisław Bień[+]
Krystyna Laus-Mączyńska[++]
Stanisław Szpakowicz[+]
[+] Institute of Informatics, Warsaw University
Warsaw, Poland
[++]Institute for Scientific Technical and Economic
Information, Warsaw, Poland

The Prolog programming language allows the user to write powerful parsers in the form of metamorphosis grammars. However, the metamorphosis grammars, as defined by Colmerauer[2], have to specify strictly the order of terminal and nonterminal symbols. A modification of Prolog has been implemented which allows "floating terminals" to be included in a metamorphosis grammar together with some information enabling to control the search for such a terminal in the unprocessed part of the input. The modification is illustrated by several examples from the Polish language and some open questions are discussed.

Metamorphosis grammars[2,3] make a convenient tool of the formal description of syntax of natural languages. Their convenience is due to their straightforward relation to the programming language Prolog. A metamorphosis grammar is an ordinary part of a Prolog program. It defines a language as well as a parser for it.

We suggest here such modifications of the way of handling the metamorphosis grammars in Prolog which allow these grammars to analyse constructions without strictly specified order of their components.

Let us consider an example. The following sentence in Polish :

(1) PRACOWAĆ BYŁO BARDZO PRZYJEMNIE
'to work' 'it was' 'very' 'nice'

"It was very nice to work."

is accepted by the metamorphosis grammar given below (nonterminals prefixed by %, terminals by #, == stands for an arrow) :

```
%S == %INF %V %ADVP.
%INF == #PRACOWAC.
%V == #BYŁO.
%ADVP == #BARDZO #PRZYJEMNIE.
```

In order to simplify the example we neglect the grammatical categories of phrases and words. The last three rules serve as "dictionary rules".

This grammar does not, however, account for many correct Polish sentences, such as:

(2) BARDZO PRZYJEMNIE BYŁO PRACOWAĆ

(3) BYŁO BARDZO PRZYJEMNIE PRACOWAĆ

To make the grammar accept these sentences we should, for example, add two rules:

```
%S == %ADVP %V %INF.
%S == %V %ADVP %INF.
```

One-third of the possible permutations of words BYŁO, BARDZO, PRACOWAĆ, PRZYJEMNIE constitute admissible Polish sentences (although sometimes stylistically marked). The complete grammar should then have 21 rules, including dictionary rules. Such a solution is obviously clumsy and not satisfactory.

Our first proposal consists in allowing two kinds of terminal symbols : anchored terminals, retrieved in the current position of a given sentence (available in metamorphosis grammars[2] and prefixed by # in our example) and floating terminals, retrieved anywhere in the unprocessed part of a sentence (we shall prefix them by @).

The easiest and most concise way of expressing a grammar for the sentences mentioned above consists in replacing every anchored terminal by a floating terminal. It is, however, not satisfactory because such a grammar accepts also deviant (syntactically or stylistically) sequences , e.g.

(4) BYŁO BARDZO PRACOWAĆ PRZYJEMNIE

(5) PRZYJEMNIE PRACOWAĆ BARDZO BYŁO

By using both the anchored terminals and the floating terminals we can define the following grammar:

```
%S == %INF %V %ADVP.
%INF == @PRACOWAC.
%V == @BYŁO.
%ADVP == #BARDZO @PRZYJEMNIE.
```

The grammar accepts only half of the incorrect sequences, but (a usual trade-off) it rejects some correct Polish sentences.

It seems that only a grammar with numerous specific rules can satisfy the strong requirement of accepting those and only those sequences which are considered correct and no others.

The formalism is, however, quite appropriate to describe e.g. the syntax of some noun phrases in Polish or syntactically unbound modifiers.

Introducing the floating terminals into the Marseille-originated Prolog interpreter requires only minor alterations of the bootstrap. The facility has been already made standard in the Prolog version for ODRA 1305 (ICL 1900 compatible) which is distributed in Poland.

To illustrate deficiencies of the proposed mechanism in parsing certain kinds of free word-order constructions we shall consider the following Polish sentences:

(6) TRZEBA BY CZEGOŚ WIĘCEJ
`is needed` `something` `more`
[present, impersonal] [conditional formative] [genitive]

(7) CZEGOŚ BY WIĘCEJ TRZEBA

"Something more would be needed."

The sentences (6), (7) consist of the impersonal conditional verb-like phrase TRZEBA BY and the noun phrase CZEGOŚ WIĘCEJ. The words CZEGOŚ and WIĘCEJ may occupy any position, but the order of TRZEBA and BY is restricted. If BY precedes TRZEBA then BY must not be the first word of a sentence, otherwise, BY must be adjacent to TRZEBA.

Therefore in order to make a concise grammar accepting all correct Polish sentences built of the words TRZEBA, BY, WIĘCEJ, CZEGOŚ, we must introduce a more selective information concerning the order of words. We supply selected terminals and nonterminals with control items restricting their scopes of floating. The lack of such an item means the restrictions inherited from the left-hand nonterminal (in particular no restrictions).

For example, such restrictions could be:

a terminal should be the last (the first),
a terminal must follow (immediately follow) the recently retrieved terminal.

Coming back to our example we should specify:
either BY follows a verb immediately,
or BY must not be the first and must precede a verb.

We can now write the grammar accepting the sentences (6), (7). The grammar is as follows (variable paramaters prefixed by asterisks, control items separated by commas).

```
%S(*TENSE,*MOOD) ==
    %VPIMPERS(*TENSE,*MOOD).

%VPIMPERS(*TENSE,*MOOD) ==
    %VIMPERS(*TENSE,*MOOD,*SYNTREQ)
    %REQ(*SYNTREQ).

%VIMPERS(*TENSE,COND,*SYNTREQ) ==
    %VERB( IMPERS,*TENSE,*SYNTREQ)
    @BY,NEXT.

%VIMPERS(*TENSE,COND,*SYNTREQ) ==
    @BY,NOTFIRST
%VERB(IMPERS,*TENSE,*SYNTREQ),AFTER.

%VERB(IMPERS,PRESENT,NP(GEN)) ==
    @TRZEBA.

%REQ(NP(*CASE)) == %NP(*CASE).

%NP(*CASE) == %NPRON(*CASE) %MOD.

%NPRON(GEN) == @CZEGOS.

%MOD == @WIECEJ.
```

In order to make the example clear we use only the categories relevant for the sentences under discussion. We omit, for instance, the number and gender of a noun phrase; the parameter *SYNTREQ expresses a single syntactic requirement (in general a verb can have more then one requirement; for details, see Szpakowicz[5]). The rule for NP is also very simplified. From the point of view of the description of Polish syntax the grammar presented above is, in fact, unsophisticated and fragmentary. It is sufficient, however, to illustrate some linguistic phenomena mentioned earlier.

An experimental version of the ODRA-Prolog accepts the metamorphosis grammar

rules with control items (syntactically just Prolog terms). The inventory of the word order restrictions has yet to be established by the research on word order in Polish. Thus, for the time being, the interpreta - tion of the control items is implemented in an ad hoc manner.

A formal description of the syntax of a natural language of free word-order type, as for example Polish and other Slavonic languages, requires, however, some addi - tional technical and linguistic problems to be solved.

We want to present now those problems which we find to be the most impor- tant.

In some cases the occurence of a word - form depends on particular proper- ties of the word which immediately prece- des it (usually it is the phonetic shape of the preceding word which influences the choice of the proper word-form). For example, agglutinative present tense form of the verb BYĆ in second person, singu- lar, masculine can be realized either by Ś or by EŚ. The forms Ś, EŚ are written jointly with the preceding syntactic item but on the level of syntactic descrip- tion they are clearly distinguishable.

Let us illustrate this problem by the following sentences:

(8)	NAROBIŁ	+ EŚ	ŁADNEGO	KŁOPOTU
	'to cause`		'cute` here: 'big`	'trouble`
	[sg, masc]	[2p, sg, masc]	[sg, masc, gen]	[sg, masc, gen]

(9) ŁADNEGO + Ś KŁOPOTU NAROBIŁ

"You've caused quite a lot of trouble."

The very simple grammar presented below accepts these two sentences but it accepts also some incorrect sequences because the rules do not express the dependency phenomena mentioned above.

```
%S == %PP(*GENDER ,*NUMBER ,NP(GEN))
  %VPT(*GENDER ,*NUMBER ,*PERSON,*X)
  %NP(*NUMBER2,*GENDER2,GEN).
%VPT(MASC,SING,2P,VOW) == @S.
%VPT(MASC,SING,2P,CON) == @ES.
%PP(MASC,SING,NP(GEN)) == @NAROBIL.
%NP(SING,MASC,GEN) ==
  @LADNEGO @KLOPOTU,AFTER.
```

(VPT - the abbreviated present tense form of the verb BYĆ; VOW and CON mean "used after a vowel" and "used after a consonant").

So far we do not see the simple and satisfactory way of relating the parameter *X of %VPT to the other words and phrases. Provisionally the agreement of the agglutinative forms of the verb BYĆ with the corresponding words may be re- solved during dictionary lookup in the pre-parsing phase.

The other purely linguistic problems are related to influence of the free word- order on accomodating the verb phrase to the gender of a compound noun phrase. For example, the verb phrases in the apo- sition agree in gender with the last consti- tuent of the noun phrase, as in:

(10)	JAN	LUB	MARIA	PRZYSZŁA
	'John`	'or`	'Mary`	'came` [fem]

Similarly, the gender of the verb phrase in the postposition may agree with the first constituent of the noun phrase , for example:

(11)	PRZYSZEDŁ	JAN	LUB	MARIA
	'came` [masc]			

It is only recently that this difficult problem has been a subject of a partial research. The formal syntax description of written sentences in Polish with neutral word-order is available[5,6]. It accepts practically all nonelliptical declarative and negative sentences, as well as the majority of interrogative sentences, nevertheless , we can propose only a provisional solution of this problem.

Another complicated question consists in the discontinuity of the phrases which constitute the sentence, as for example interpenetration of the verb phrase and the noun phrase :

(12)	NOWĄ			KSIĄŻKĘ		/NP/
		DAŁ	JAN		MARII	/VP/
	'new`	'gave`	'John`	'book`	'Mary`	
	[acc]		[nom]	[acc]	[dat]	

"It is a new book that John gave to Mary".

Therefore the control information should allow the search of missing constituents of the phrases even far off the main component. On the other hand it should protect against "borrowing" an inappropriate constituent from a quite different phrase, e.g. from the subordinate clause.

It is now clearly visible that parsing free word-order languages is really different from the syntactic analysis of, say, English. Although the presented modifications of metamorphosis grammars do not solve all the problems discussed above, they provide a useful instrument for further experimental studies.

Finally we want to emphasize that we were aware of the semantic and pragmatic functions of free word-order, which are studied e.g. by Sgall[4] and Szwedek[7]. But we believe that, from the methodological point of view, it is justified to prescind from them in the syntax description. A reader interested in some notions of the impact of word-order on semantico-pragmatic level, may wish to consult Bień[1].

References

[1] Bień J.S. Multiple Environments Model of Natural Language [in Polish, unpublished Ph.D.thesis], 1977.

[2] Colmerauer A. Metamorphosis Grammars. In Bolc L.(ed) Natural Language Communication with Computers, Lecture Notes in Computer Science 63, 1978.

[3] Pereira F., Warren D.H.D. Definite Clause Grammars Compared with Augmented Transition Networks. Dept.of AI Report 58, University of Edinburg, 1978.

[4] Sgall P., Hajičová E., Benešová E. Topic, Focus and Generative Semantics. Kronberg Taunus: Scriptor Verlag GmbH, 1973.

[5] Szpakowicz S. Automatic Syntactic Analysis of Polish Written Utterances [in Polish, unpublished Ph.D. thesis], 1978.

[6] Szpakowicz S. Syntactic Analysis of Written Polish. In Bolc L.(ed) Natural Language Communication with Computers, Lecture Notes in Computer Science 63, 1978.

[7] Szwedek A. Word Order, Sentence Stress and Reference in English and Polish. Edmonton: Linguistic Research Inc.,1976.