# A MODULAR PARSER FOR FRENCH

Eric Wehrli

Geneva University Hospital

ABSTRACT

In this paper, we describe an efficient parser for French based on an adaptation of Chomsky's Government-Binding (GB) theory. Reflecting the modular conception of a GB grammar, the parser consists of several distinct procedures corresponding to the subsystems of the grammar (e.g. phrase-structure rules, binding, control, 'theta'-theory, etc.). The interaction of these fairly simple modules produces the kind of complexity required in order to build all the linguistically motivated structures for a given sentence.

## I INTRODUCTION

This paper is a report on work in progress at the University Hospital of Geneva. It describes a general parser for French developed as part of a project to design a natural language interface for a data base system. The originality of this parser lies both in its linguistic foundations and in the way they have been implemented. The grammar on which this work is based is an adaptation of Chomsky's Government-Binding Theory (cf. Chomsky, 1981). It can be viewed as a system of components and a system of principles falling into various subsystems such as the X theory, the theta theory, the binding theory, control, government, etc. There is a large amount of interaction among these subsystems, as well as between the subsystems and the rule components, leading to the intricate system of the grammar.

The modular conception of the grammar was carried over to the parser, which consists of several procedures corresponding to subtheories of the grammar. This approach makes it possible to break the analysis into distinct tasks which can be achieved following different strategies.

This paper is organized as follows : the next section concerns the specifications of the grammar. Section 3 gives an overview of the parser and presents some examples of grammatical constructions that the parser handles at the present stage. Finally, the various syntactic procedures are discussed in section 4.

*This research is directed by Prof. J-R. Scherrer and -sponsored by the Swiss National Science Foundation.

## II LINGUISTIC SPECIFICATIONS

An important feature of our approach is the central role of the lexicon. Following the spirit of Chomsky's Projection Principle, we assume that the structure of a sentence is essentially a projection of the lexical specifications of its elements. Furthermore, we assume that the range of possible non-terminal categories is largely restricted to projections of the lexical categories N, V, A and P, in accordance with the X system. We also make use of the results of trace theory.

We have adopted a version of the X theory according to which a non-terminal category, say V dominates its head, V, and two (possibly empty) sets of constituents : Spec V (the specifier system of V) and Comp V (the complement system of V). Compl V is completely determined by the lexical features of V (i.e. the subcategorization features). As for the Spec V set of constituents, we take it to be an inherent property of the given category. Thus, for instance, we will take auxiliaries and clitic pronouns to be part of the specifier of V. Similarly, the specifier of N will include determiners, adjectives, etc.

As for the morphological component, we assume it to be internal to the lexicon, which is structured in such a way as to minimize redundancy.

The grammar contains :

(i) a lexicon, which includes a list of all the words (including compounds) of the language, along with their categoria1 , morphological, syntactic and semantic features.

(ii) a phrase-structure component, which specifies the set of possible structures of the language. It is to a large extent redundant with the lexical information (i.e. the subcategorization features).

(iii) a rule of binding, which associates an element in non-argument position with an empty category (i.e. a trace) in argument position.

(iv) a rule of control, which assigns an interpretion to the empty (i.e. non-lexically realized) subject of infinitival clauses.

(v) a condition on well-formedness of the structures, essentially an adaptation of Chomsky's 1981 theta-criterion, which rules out structures in which obligatory arguments are missing.

## III OVERVIEW OF THE PARSER

The parser reads a sentence and returns the set of possible analyses for this sentence.The analyses are given both in a phrase structure representation (using the familiar labelled bracketing notation) and in a functional structure representation. The former, which corresponds to an anotated surface structure, emphazises the structural relations between constituents, while the latter emphasizes the grammatical functions. An illustration is given in (1), where a is the input sentence, b the phrase structure and c the functional structure :

(1) a. de quel livre Jean a-t-il parlé à Marie ?
'about which book did Jean talk to Marie'

b. Syntactic structure :
$[_{\bar{S}}[_{COMP}[_{PP}de[_{NP}quel\ livre]]]_i[_S[_{NP}Jean]_j$
$[_{VP}a\text{-}t\text{-}il_j parlé[_{PP}e]_i[_{PP}à[_{NP}Marie]]]]]$

c. Functional structure :
CATEGORIE : $\bar{S}$ TETE : parlé
TEMPS : PASSE COMPOSE INTERROGATIF
SUJET : TRACE DE Jean
COMPLEMENT PREP. TETE : Marie
COMPLEMENT PREP. TRACE DE quel livre

The functional structure not only provides the list of arguments ordered with respect to their grammatical function, it also specifies the type of sentence (interrogative, negative, etc.) as well as tense, mood and voice of the verbs. The element e_ in (lb) stands for 'trace' and the indices represent the binding relations .

The parser can handle a substantial number of grammatical constructions, including wh and yes/no questions, imperatives, relative clauses, "tough-movement, tensed and infinitival sentential complements, passives, several cases of coordination, clitic pronouns, causatives, etc. Some of these constructions are illustrated in the printout given in figure 1 below, where a is the input sentence and *b* the syntactic structure provided by the parser. For reason of space, the functional structures have been removed.

Example (2), 'has he given them to him', contains three clitic pronouns, les, lui and il. All three of them are coindexed with a trace in argument position. Les is coindexed with an NP trace in direct object position, lui with a PP trace in indirect object position and il with an NP trace in subject position.

Sentence (3), 'which book has Jean persuaded Marie to give to Paul' is an example of long distance dependency. The wh-element quel livre is linked to a trace in the direct object position of the embedded sentence. It is, therefore interpreted as the direct object of donner. Notice, also, that the direct object of persuader controls the empty subject of the infinitival sentence.

Finally, in (4) we have an example of ambiguous sentence. In fact, this sentence has the peculiar property that each one of its words is ambiguous : le and la can be determiners or pronouns, belle is an adjective or a noun, ferme a noun or a verb, voile a noun or a verb. Among the numerous potential combinations, the parser correctly returned only two structures : (4bl) 'the beauty closed the veil' and (4b2) 'the nice farm is hiding if .

(2)a. les lui a-t-il donnés ?

b. $[_S[_{NP}e_k][_{VP}les_i\ lui_j\ a\text{-}t\text{-}il_k donnés\ [_{NP}e_i][_{PP}e_j]]]$

(3)a. quel livre Jean a-t-il persuadé Marie de donner à Paul ?

b. $[_{\bar{S}}[_{COMP}[_{NP}\ quel\ livre]]_i[_S[_{NP}\ Jean]_j[_{VP}\ a\text{-}t\text{-}il_j\ persuadé\ [_{NP}\ Marie]_k$
$[_{\bar{S}}de[_S[_{NP}\ e]_k\ [_{VP}\ donner[_{NP}\ e]_i\ [_{PP}\ à[_{NP}\ Paul]]]]]]]]]$

(4)a. la belle ferme le voile.

bl. $[_S[_{NP}\ la[_{ADJ}\ belle]\ \ ferme][_{VP}\ [_{CLI}\ le_i]\ voile\ [_{NP}\ e]_i\ ]]$

b2. $[_S[_{NP}\ la\ belle\ ]\ [_{VP}\ ferme\ [_{NP}\ le\ voile\ ]]]$

Figure 1. Samples of analyses

## IV THE SYNTACTIC ANALYSIS

The analysis can be sketched as follows : first, a lexical analysis of the input sentence is performed. It produces a chart data structure (in the sense of Kaplan, 1973) consisting of n+1 vertices (where n is the number of words) related by as many edges as there are readings of the words.In other words, each edge in the chart corresponds to a particular reading and bears all the lexical information relevant to it.

The chart then undergoes the syntactic analysis, which is divided into several distinct procedures corresponding to the different components of the grammar described in section 2.

The reduction process, which roughly corresponds to the phrase-structure component consists of two procedures. A first procedure takes care of lexical heads such as N, V and Adj and builds up respectively NP, VP and AP constituents combining the head and its specifier system. This first step is achieved by means of a predictive analysis. As for the second procedure, it attempts to reduce these constituents into larger and larger constituents until it reaches the top level. The strategy, this time, is striclty bottom-up.

The rationale for this division between a bottom-up analysis for the complement systems and a predictive analysis for the specifier systems is based on a very interesting property of the grammar of French (and presumably of many other natural languages). Verbs, nouns and adjectives take essentially the same range of complements (i.e.pps, Ss, NPs), a property which leads to potential ambiguities. On the other hand, they usually have distinct specifier systems. This latter property makes a predictive analysis both convenient and appropriate.

Edges corresponding to non-terminal categories inherit the lexical specifications of their lexical heads. They also have registers in which both the phrase structure and the functional structure are built. Each syntactic rule is associated with a set of conditions and a set of actions. The conditions verify that the rule is compatible with the features of the constituents to be combined. A typical example of a condition is number and gender agreement. Another example is the condition on interpretation. This condition states, for instance, that a VP and an NP can only be combined into a larger VP if there is a possible interpretation for NP in VP. This rules out attempts to build a transitive structure with an intransitive verb.

The actions associated with a rule specify the effect of the rule on the phrase structure and on the functional structure. Thus, if the rule combines a VP and an NP, the action specifies (i) that the NP has to be inserted in the phrase structure as a daughter of VP, and (ii) that it will fill up the direct object slot in the functional structure.

The binding procedure takes care of elements which are not in argument positions and assigns them an interpretation. Following standard practice in Generative Grammar, grammatical functions are determined on the basis of structural properties. Thus, for instance, subject is defined as NP immediately dominated by S, object as NP immediately dominated by VP, and so on (cf. Chomsy, 1965). These positions are called argument positions. However, there are cases where a constituent interpreted, say, as a direct object does not appear in the canonical position for direct object. This is typically the case of question words and clitic pronouns. Following the spirit of trace theory, we assume that these constituents get their interpretation through the binding of an empty category (i.e. a 'trace') in argument position.In the phrase structure representation this binding relation is expressed by means of coindexing.

The binding procedure works like this. Take the case of wh-binding. The procedure is activated by the rule combining a wh-element and a sentence. The procedure looks for an appropriate slot in the functional structure associated with the S edge. If there is such a slot, an empty element is created and inserted both in the phrase structure and in the functional structure. In case there is no appropriate slot in the matrix sentence, or if the slot is an optional one, the search continues in the next embedded S.

To give a concrete example, suppose that we have a wh-word, such as qui ('who') and an adjacent sentence, such as est-ce que Jean a rencontre ('has John met'). The functional structure attached to the sentence contains the argument slots specified in the lexical properties of the lexical head of the sentence, that is the verb rencontrer ('to meet'). This follows from the Projection principle. Since this verb is transitive, the functional structure contains two slots : a subject slot filled by the NP Jean, and a direct object slot, which has not yet been satisfied. Because this empty slot is compatible with the wh-word, the binding procedure assigns it an empty category ($[_{NP}e]$) bound (i.e. coindexed) to the wh-word.

Notice that, since a new edge combining a wh-element and an S can only be created if a trace can be assigned to the wh-element, the procedure guarantees that no wh-elements in a given structure are left uninterpreted. The search for an appropriate slot is interrupted if there is an intervening COMP position containing a wh-word, or if the embedded S is dominated by an NP node. In other words, Chomsky's Subjacency condition is built into the procedure.

The binding procedure applies in a very similar fashion in structures containing clitic pronouns. Like wh-words, clitic pronouns do not occur in argument position and must, therefore, bind a trace in order to get an interpretation.

The control procedure takes care of the control phenomenon, i.e. the interpretation of the subject of infinitival sentences. It is well-known that in the following examples the subject of the infinitival verb is interpreted as being Jean in (5a) and Marie in (5b) :

(5) a. Jean a promis a Marie de partir.
'Jean promised Marie to leave'
b. Jean a persuade" Marie de partir.
'Jean persuaded Marie to leave'

Following again standard assumptions in Generative Grammar, we assume that an infinitival complement has essentially the same structure as a tensed complement, namely a sentential structure. However, whereas tensed sentences usually have lexical subjects, infinitival sentences exhibit an empty subject. This empty category, commonly referred to as PRO, is interpreted either as co-referential with an argument oi the; matrix verb (as in (5)), or as arbitrary in reference, depending on the lexical properties of the matrix verb. Activated by the rule attaching an infinitival S to a VP, the procedure determines the value of the controller (i.e. the antecedent of PRO) and creates an anaphoric relation between PRO and its controller.

Both the binding and the control procedures establish a relation between an antecedent and a bound anaphoric element. In the syntactic structure, this relationship is expressed by means of co-subscripting, which is achieved as follows : each constituent is attached a register containing a pointer to its antecedent and a pointer to an anaphor. Notice that an anaphoric element can itself function as the antecedent of an anaphor, thus determining a chain of anaphoric relations. At the end oi the syntactic analysis, a simple mechanism assigns an index to each anaphoric chain.

Finally, the 'theta' procedure verifies that the analyses satisfy the condition that all the obligatory arguments of a verb be assigned a constituent. Structures that do not fulfill this requirement are ruled out. Notice that this procedure only enforces half of Chomsky's Theta-criterion. The other part (i.e. each argument is assigned a theta-role' follows from the fact that an argument can only be attached to a given constituent if the functional structure has an appropriate slot available.

## V. CONCLUSION

We have presented a parser based on recent developments in theoretical linguistics. Reflecting the modular conception of the grammar, it consists of several procedures corresponding to subsystems of the grammar. This strategy proved highly successful : the parser is powerful enough to build all the linguistically motivated structures for a given input sentence (for a substantial subset of the language). But it is also restricted enough so that it will not provide alternative analyses when the input sentence is not truly ambiguous. This is by no means a trivial task, and should greatly reduce the burden of the semantic analysis.

## REFERENCES

[1] Chomsky, N. Aspects of the Theory of Syntax, Cambridge, MIT Press, 1965.

[2] Chomsky, N. Lectures on Government and Binding, Dordrecht, Foris Publications, 1981.

[3| Kaplan, R. "A General Syntactic Processor", in Rustin, R. (ed.)
Natural Language Processing, Algorithmics Press, 1973.