

EXTENDING THE EXPRESSIVE CAPACITY OF THE SEMANTIC COMPONENT OF THE
OPERA SYSTEM

CELESTIN SEDOGBO

CENTRE DE RECHERCHE BULL
68, ROUTE DE VERSAILLES
78430 LOUVECIENNES
FRANCE

ABSTRACT:

OPERA is a natural language question answering system allowing the interrogation of a data base consisting of an extensive listing of operas. The linguistic front-end of OPERA is a comprehensive grammar of French, and its semantic component translates the syntactic analysis into logical formulas (first order logic formulas).

However there are quite a few constructions which can be analysed syntactically in the grammar but for which we are unable to specify translations. Foremost among these are anaphoric and elliptic constructions. Thus this paper describes the extension of OPERA to anaphoric and elliptic constructions on the basis of the Discourse Representation Theory (DRT).

1- INTRODUCTION

OPERA is a natural language question answering system allowing the interrogation of a data base consisting of an extensive listing of operas and their associated characteristics (composers, time of composition, recordings, etc...). The linguistic front-end of OPERA is a comprehensive grammar of French based essentially on the string grammar formalism (cf for example [SAGER, 81] or [SALKOFF, 73]); the details of the syntax used in OPERA are described in [SEDOGBO, 85]. In this paper we shall describe the semantic component of OPERA as it stands now. In addition we shall discuss a number of problematic constructions which are not handled at present in the system and we outline a possible solution for them in the setting of this system.

It has become more and more clear that the most natural and transparent formulation of the semantics of natural languages can be constructed in terms of a logical semantics i.e a semantics based on the usual model theoretic interpretation of predicate logic.

This was in fact already realized as early as 1973 by A. COLMERAUER and his group (cf [COLMERAUER & al, 73], [PASERO, 73], [COLMERAUER, 79]). In these papers the semantics for natural languages is specified in terms of the truth conditions of predicate logic formulas into which these utterances are translated in a systematic way. (It should be mentioned that in many respects the way these translations are obtained as well as the translations themselves resemble the even earlier proposals of MONTAGUE and his students (cf [MONTAGUE, 74]). The work done in COLMERAUER's group has however the additional merit that these proposals are presented in operative systems for man-machine dialogues. Indeed in these early papers we find sketched two alternatives for the systematic manipulation of the translations provided for natural language discourse:

This research was supported in part by ESPRIT, in the ACORD project (P393).

a) given a data base in relational (e.g Prolog like) form the answers to queries (which are themselves translated into logical forms) can be given in terms of the satisfaction conditions of these formulas. I.e a query like "does every student own a book? " will receive the answer "yes" if its translation " for all x (student(x) -> there exists y (book(y)& own(x,y))) " is true in the data base. Here the answers to queries are obtained by simply evaluating the translation in the data base (or model).

b) another approach however consists in regarding the data base as a set of formulas and the process of answering a query as a deduction. A query will be said to be true with respect to the data base if and only if it can be deduced from it. (Naturally as in the case of the first approach, this method applies equally well to closed sentences (i.e yes-no questions) as well as to open sentences (i.e WH-questions). For an application of this method as well as for discussion of its advantages (cf [PASERO, 73]).

For reasons which we shall not spell out here we have chosen the model theoretical approach (i.e the first approach) in OPERA. Needless to say it is not too difficult and in fact sometimes necessary, as we shall see below, to complement the pure semantic evaluation with deductive capacities.

2- SEMANTICS VIA TRANSLATION INTO LOGIC

Even if it is clear in most respects how to obtain predicate logic translations for a large variety of natural language sentences, cf [WARREN & PEREIRA, 82] we repeat here for the sake of clarity the essentials of the translation process in OPERA.

Consider for example the following sentences :

- (a) Berg a composé un opera en trois actes
(Berg composed an opera in three acts)
- (b) aucun compositeur Allemand n'a composé un opera en 1912
(no german composer composed an opera in 1912)
- (c) Chaque opera de Berg a été enregistré par Karajan
every opera by Berg was recorded by Karajan.

The syntax of OPERA (cf [SEDOGBO, 85]) yields syntactic analysis as in (d).

These trees are then translated into the following predicate logic formulas :

- (a) exist(x,opera-in-three-acts(x) & compose(Berg,x))
- (b) aucun(x,composer(x) & german(x),exist(y,opera(y) & compose-by(y,Berg) & in(y,1912)))
- (c) chaque(x,opera(x) & of(x,Berg),
recorded-by(x,Karajan))

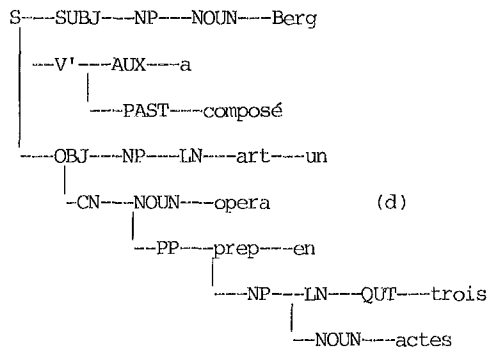


Figure 1: an example of syntactic tree in OPERA

The evaluation component of OPERA then evaluates these formulas in the data base. Quite a few syntactically complex queries can thus be formulated in this system.

3- FROM TRANSLATIONS TO DRSS

OPERA was conceived not so much as an effort towards an implementation of viable integrated natural language question answering system, as a testing ground for a rather extensive fragment of French. In fact many constructions (in particular complement constructions) which are accomodated in the syntax fragment of OPERA are such that they cannot effectively be applied in the setting of OPERA's data base.

However there are quite a few constructions which can be analysed syntactically in the grammar but for which we are unable to specify translations, even when their intended translations are essentially first-order logic formulas. Foremost among these are anaphoric and elliptical constructions. It is clear that any comprehensive treatment of such phenomena cannot restrict itself to the analysis of isolated sentences i.e to isolated queries. For what could a sentence like: when did he compose it ? mean ? or what query could be expressed by the phrase "and Britten?"?

There is not much that these sentences could convey in isolation. But consider the following dialogue :

Question: did Beethoven compose Fidelio ?
 Answer : yes.

Question: when did he compose it?
 Answer: in 1812

Or consider another very natural interaction :

Question: how many operas did Beethoven compose?
 answer: 1

Question: and Mozart?
 Answer : 15

It is clear that an adequate account of anaphoric and elliptic constructions must at least take into account the current situation of the dialogue. How to define the notion " current situation " is by no means a trivial task. For one thing we cannot simply assume that the current situation is identical to the sequence of the question-answer pairs that make up the dialogue. Much more is involved.

3.1 TOWARDS A TREATMENT OF ANAPHORA

Even though pronouns raise no problem at all as for us as far as the syntax is concerned, no one has been able to provide a unified and systematic account of anaphoric linking. As recent work has shown one must distinguish on the one hand various types of pronominal reference and on the other hand show how the resolution of anaphora must appeal to common underlying mechanisms.

It is one of the achievements in [KAMP, 81] to show the way towards that unified treatment .

In fact we shall base our proposals for both the treatment of anaphora and of ellipsis on the concepts and technics introduced by KAMP in his theory of discourse representation.

Since the current translation component of OPERA does not take into account pronouns at all we must first provide an extension which is also able to deal with pronominal reference. Syntactically pronouns can occur in all NP positions characterized by the grammar.

We shall mark each occurrence of a pronoun by indicating its gender/number features; As a first step in the translation procedure we shall therefore simply take these translations to be identical with themselves. Thus the translation of sentences like :

Berg l'a composé
 (Berg composed it)

will be :
 compose(Berg, [it,null,sing])

similarly a sentence like :
 chaque compositeur qui a composé un opera l'a enregistré
 (every composer who wrote an opera recorded it).

will be translated into :
 chaque(x,exist(y,opera(y)&composer(x)&compose(x,y),record(x,[it,null,sing])))

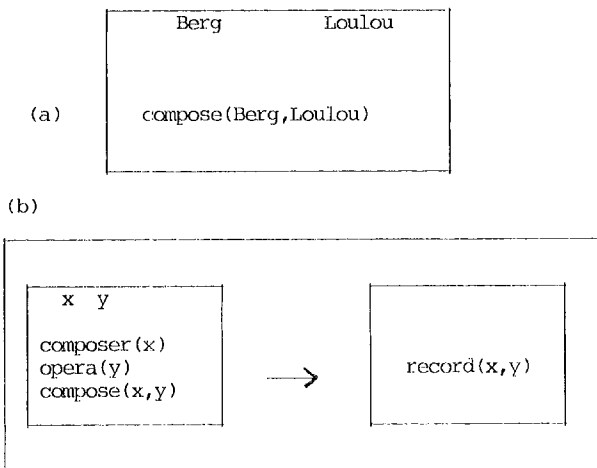
We shall call translations containing occurrences of pronouns "unresolved translations". Notice that contrary to the practice in [MONTAGUE, 74] we do not index pronouns either in the syntax or the initial translation phase. The resolution of anaphora will take as input "unresolved translations" and yield as output a rather different type of semantic representation namely a discourse representation structure (DRS).

What is DRS? In general we shall say that a DRS is pair consisting of a (possibly empty) domain of discourse referents U and a set of conditions CON. We shall for the present discussion take into consideration only three types of conditions :

- a) atomic conditions, which consists of n-ary predicate P and terms.
 a term is either a discourse referent or a proper name; among the predicates we single out the equality predicate "=".
- b) conditional conditions which have the form =>(k1,k2) where k1 and k2 are again DRSSs.
- c) negative conditional conditions which has the form ≠>(K1,k2).

In a more comprehensive treatment it is clear that we shall need further types of conditions.

Thus for the final translations of unresolved translation we want to arrive at DRSs like the following :



There are precise truth definitions for DRSs which we shall not spell out here however (cf KAMP for details). In any case the first DRS is logically equivalent to the formula : $compose(berg, loulou)$ and the second DRS is logically equivalent to the predicat logic formula :

$chaque(x, (chaque(y, (composer(x) \& opera(y) \& compose(x,y)) \rightarrow record(x,y)))$

What is interesting in the second example is of course a fact that the pronoun "it" has as its syntactic antecedent the noun phrase "an opera ". The semantic force of this noun phrase cannot however be rendered in terms of an existential quantifier at least not if we want to establish an anaphoric link between the existential quantifier and the variable representing the pronoun in the consequent of the conditional.

MONIAGUE for instance runs into this problem in PIQ where the pronoun is either left unbound or either is bound by the existential quantifier when the latter occur outside the conditional all together. This wide scope reading of the indefinite NP obviously gives them the wrong interpretation. As does of course the translation leaving the pronoun unbound. In the DRS on the other hand we get a universal reading for an opera as the result of the interpretation of the conditional.

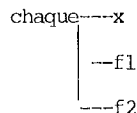
Even though inside the antecedent of the conditional we treat the occurrence of the indefinite noun phrase "an opera " as we would in an ordinary indicative sentence, where of course the interpretation of the indefinite article corresponds more naturally to an existential quantifier. This is one of the features of KAMP theory that we shall take advantages of in the present proposal.

As we said above we shall transform unresolved translations stepwise into DRSs. Needless to say we could of course set up the translation procedure in such a way that we obtain DRSS directly from the output of the syntactic component. But this would entail major revision of the entire translation algorithm in any event the way we propose to derive DRSS can be

regarded as DRS construction algorithm in its own right. On this view "our unreduced translations" play the role of an intermediate structure between syntax and DRSs.

Assume that t is an unresolved translation of a sentence . We first generate K(t) the discourse representation structure corresponding to t, in the following way :

if t is a universal tree i.e a formula of the form

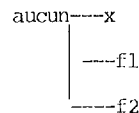


we create a DRS K(t) with an empty domain and the condition $\Rightarrow(K1,K2)$.

K1 will have x in his domain and the result of transforming f1 in K1 as its condition

K2 will have an empty domain and the result of transforming f2 as its conditions.

Let t be a tree with the determiner "aucun" as its top node, i.e a formula of the form :



we procede as in the case above except that the condition we enter into K(t) is now a negative universal condition, i.e a condition of the form $\Rightarrow(K1,K2)$.

Suppose t is dominated by exist i.e t has the form :



we create a DRS K(t) whose domain contains x and we add the result of transforming f1 as the conditions to K(t).

Suppose t is a tree dominated by "et" i.e a formula of the form :



we create a DRS K(t) with an empty domain whose conditions are the result of transforming f1 and f2 with respect to K(t).

Finally suppose the unresolved translation is non quantified then we enter it as is into K and we add all occurrences of proper names into the domain of K.

This provides the induction basis for the transformation. Let K' be a DRS with domain U' and conditions CON' and let t be a tree i.e a formula. The result of transforming t in K' is the application of the above three rules to t. When we no longer have any tree to process all conditions in the principle K i.e the DRS representing the sentence to be transformed, as well

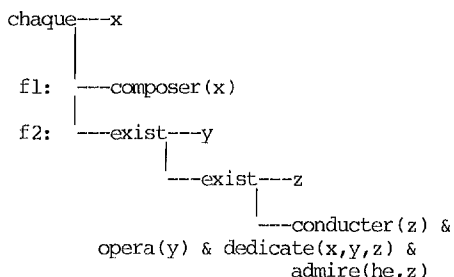
as all conditions occurring in the sub-DRSs of K will now be conditions in the language of DRSs or atomic conditions containing occurrences of pronouns.

How are these be eliminated ?

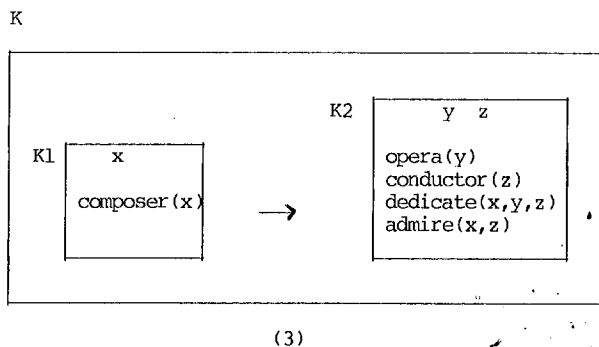
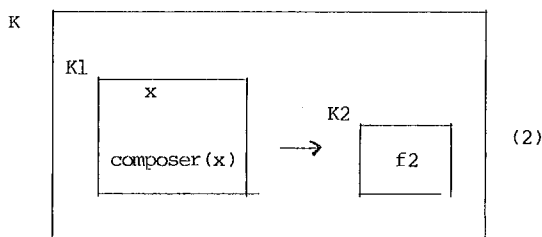
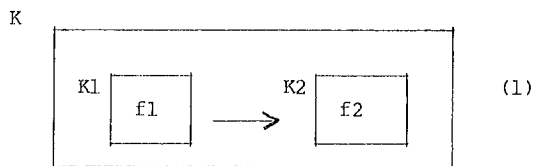
Let us first consider an example the sentence

"every composer dedicated an opera to a conductor that he has admired"

has as its unresolved translation:



We indicate the construction of the DRS stepwise



3.2 INCORPORATING DRS CONSTRAINTS INTO TRANSLATIONS

We will give below the exact method for translating formulas of OPERA into DRSs.

One of the most important features of DR-theory is the precise constraints on the antecedents of pronouns.

Let K' be a DRS embedded in K (which is a DRS too); the antecedent of a pronoun occurring in K' is the list consisting of the union of $U(K)$ and $U(K')$ if and only if K is accessible to K' .

For a precise definition of the notion of accessibility of [GUENTHNER & LEHMAN,85].

Let us now illustrate the notion of accessibility by giving the table of accessibility of the DRS above:

DRS	DRS accessible
K1	K
K2	K1

(1)

By the transitive closure of the accessibility relation we obtain for example all the possible antecedents of a pronoun occurring in $K2$ (e.g a pronoun occurring in $K1$ cannot have as antecedent a referent of $K2$).

For the clarity of what will follow, we will call "unresolved predicate" (abbreviated UP) a predicate whose arguments included at least one pronoun. Then to resolve an UP, one must replace the pronoun arguments with appropriate referents accessible.

The idea is to transport during the translation of formulas, a list of antecedents accessible according to DRSs constraints.

How is this list to be constructed?

As shown in (3.1) a universal tree is translated into a DRS $K = \Rightarrow(K1, K2)$; the antecedent list of K which we note $L(K)$ is empty.

The antecedent list of $K1$ is $L(K1)$ and $L(K2) = U2 + L(K1)$ (we denote the union of sets by the symbol +).

The existential tree is translated into $K = [U, CON]$ and the antecedent list $L(K) = U$.

Let f be a formula with "aucun" at its top node; f is translated into the DRS $K \not\Rightarrow(K1, K2)$. The list of antecedents $L(K)$ of K is empty and $L(K1)$ is $U1$ and $L(K2) = U2 + L(K1)$.

Let us call a DRS containing "unresolved predicates" an "unresolved DRS". Thus each unresolved DRS is a pair of the form $\underline{K} = [K, L(K)]$ where $L(K)$ is the antecedent list of the DRS K .

To resolve an unresolved DRS, each unresolved predicate (UP) must be resolved according to the following rule :

for an unresolved DRS $\underline{K} = [K, L(K)]$ with $K = [U, CON]$ a UP P of CON is transformed into the logical predicate P' which is obtained by unifying pronoun arguments of P in $L(K)$.

The example "every composer dedicated an opera to a conductor that he has admired" treated in session 3.1 will illustrate how unresolved DRSs are resolved.

Given the unresolved DRS $\underline{K} = [K, nil]$ with

$K = \Rightarrow(k1, k2)$, $\underline{K1} = [K1, x]$, $K1 = (x, composer(x))$

and

$K2 = [K2, (x, y, z)]$ with $K2 = ((y, z, he), CON2)$, $CON2 =$
 $opera(y) \& conductor(z) \&$
 $dedicate(x, y, z) \& admire(x, he)$

K will be resolved by application of the rule described above, i.e.: The pronoun he is unifiable in the list (x,y) of antecedents, to z. Therefore the unresolved predicate will be translated into $admire(x, z)$.

4- EXTENDING THE DIALOGUE CAPACITY OF OPERA

A problem arises in interrogating a data base in natural language which is the problem of dialogue situation. For example :

- (a) qui a composé Loulou (who composed Loulou ?)
- (b) où est-il né? (where is he born?)

This dialogue can be translated into a DRS containing the semantic representation of the two sentences. But if there is no interaction with the data base, the anaphoric link of the pronoun in (b) will be an unbound variable and not for example the individual "Berg" (as it is Berg who composed Loulou).

What we prefer for a question-answering system is to take into account the situation of a question and its answer.

A possible solution could be to consider that after the evaluation of a query, its corresponding DRS is therefore instantiated (i.e. its unbound variables are now bound); in such a situation we loose the reading of the DRS as a logical formula and moreover we cannot represent the instantiation of a splitted DRS (e.g. a universal formula).

4.1 THE NOTION OF A "CURRENT SITUATION DRS"

The solution we proposed is to separate the DRS of a formula from the DRS for the situation.

Let t be a formula and $K(t) = [U, CON]$ its translation into a DRS.

The DRS of situation is the DRS $K\$(n) = [U\$, CON\$]$

with

U\$ containing the instantiation of referents of U,
 CON\$ empty, and n denoting the current state of the dialogue (i.e. the occurrence of the question during the dialogue).

The rule for UP resolution presented in the last session, must now be modified in the following way:

let n be the current state of the dialogue and K be an unresolved DRS; the antecedent of a pronoun occurring in K is contained in the list consisting of the concatenation of $L(K)$ and $L(K\$(n-1))$.

The antecedent list of $K\$(i)$ is defined in the same way as that of a normal DRS, i.e. $L(K\$(i)) = U\(i) .

To illustrate what is above, we will treat the dialogue above:

(a) is translated into the logical formula $t1 = Wh(x, compose(x, Loulou))$; its translation into a DRS will produce K1.

K1

?x Loulou compose(x, Loulou)

Since there is no pronoun in K1, we can then evaluate the formula t1. After evaluation, we can build the DRS of situation $K\$:$

$K\$(1)$

Berg Loulou

(b) will be translated into the logical formula $t2 = Wh(y, born([he, mas, sing], y))$

The unresolved DRS of t2 is $K2 = [K2, L(K2)]$ with

K2

[he, mas, sing] y born(he, y)

and $L(K) = y + L(K\$(2-1)) = [y, Berg, Loulou]$.

To resolve K2, an antecedent must be substituted to "he"; this antecedent will be Berg (because arguments in OPERA are typed).

The DRS of situation becomes :

$K\$(2)$

Berg Vienne

4.2 TREATMENT OF ELLIPSIS

One of the most common phenomena of dialogue is ellipsis such as in :

(a) Berg a composé Loulou
 Berg composé Loulou

(b) Britten aussi
 Britten too

or

(c) et une symphonie
 and a symphony

The interpretation of (b) is "Britten a composé Loulou" (i.e. VP-ellipsis); and (c) is to be interpreted as "Berg a composé une symphonie".

The ad hoc treatment proposed in [SEDOGBO, 85] fails in many cases as we mentioned, since the logical formulas in OPERA are equivalent to first-order logic formulas.

An interesting extension to VP-ellipsis in DR-theory is described by KLEIN (cf [KLEIN 84]). KLEIN introduced the notion of predicate-DRS and pro-

posed an indexation of NP predicate-DRS and VP predicate-DRS in a DRS.

We will not propose here how this extended DRS can be implemented; but will exploit the parallelism between our logical formulas and DRSs.

Each sentence will be translated into two partial-logical formulas (noted PLF) of the form $\langle x, f \rangle$, where x is a variable or an individual and f a logical formula.

We assume that the composition of PLF(NP) and PLF(VP) gives the translation of the sentence.

The first PLF is implicitly indexed by the NP and the second PLF is indexed by the VP.

The DRS of current situation must be modified in the following way : $K\$(n) = [U\$, CON\$\]$ with

$U\$\$ defined as above, and $CON\$\$ containing PLF(VP).

Given a VP-ellipsis s , it will be translated into the PLF: $\langle i, f1 \rangle$. The DRS of current situation contains in its $CON\$\$ a PLF $\langle j, f2 \rangle$.

The VP-ellipsis is treated in the following way:

- 1) unification of i and j
- 2) the composition of $\langle i, f1 \rangle$ and $\langle j, f2 \rangle$ produces the translation t
- 3) t is translated into a DRS K and $K\$\$ is built as described in session (4.1)

We will illustrate the VP-ellipsis treatment by processing the dialogue above:

(a) is translated into the formula $\text{compose}(\text{Berg}, \text{Loulou})$ and $\text{PLF}(\text{NP}) = \langle \text{Berg}, \text{Berg} \rangle$, $\text{PLF}(\text{VP}) = \langle i, \text{compose}(i, \text{Loulou}) \rangle$.

The translation t of (a) is then obtained by composition of $\text{PLF}(\text{NP})$ and $\text{PLF}(\text{VP})$; thus $t = \text{compose}(\text{Berg}, \text{Loulou})$.

The evaluation of t will augment the dialogue situation of a new DRS of situation containing $\text{PLF}(\text{VP})$.

The translation of (b) will produce a $\text{PLF}(\text{NP}) = \langle \text{Britten}, \text{Britten} \rangle$ and a $\text{PLF}(\text{VP}) = \langle i, p \rangle$.

$\langle i, p \rangle$ will then be unified to the $\text{PLF}(\text{VP})$ contained in the $CON\$\$ of $K\$\$, i.e $\langle i, p \rangle$ is unified with $\langle i, \text{compose}(i, \text{Loulou}) \rangle$; the composition of the two $\text{PLF} \langle \text{Britten}, \text{Britten} \rangle$, $\langle \text{Britten}, \text{compose}(\text{Britten}, \text{Loulou}) \rangle$ will produce the logical formula " $\text{compose}(\text{Britten}, \text{Loulou})$ ".

5- CONCLUSION

The extensions to the OPERA system proposed here give a powerful dialogue capacities to OPERA.

On the basis of the DR-theory, we propose an extension for the treatment of anaphora. We do not treat here the definite article as a definite reference since in our system the definite article must be interpreted as an indefinite article. However notice that even if in the framework of DR-theory the definite article is explained as a definite anaphora, its anaphoric link requires often the use of deduction. We prefer therefore not to treat the definite anaphora.

The extension to ellipsis described in this paper is limited to VP-ellipsis; in fact the other kinds of elliptic sentences can be seen as conjoined to the preceding sentences. Even if the treatment proposed is not the one described by KLEIN, the notion of partial logic formula is equivalent to that of partial-DRS.

In order to handle the dialogue we introduce the notion of a current situation DRS. But, is the level of logic translations (i.e our logical formulas) necessary, since we translate these formulas into DRSs?

The reason to maintain this intermediate representation is that our use of DRSs is only justified by the accounting of the dialogue, so that we do not need the complex features of a DRS system. Then for reason of efficiency we think that it is better to evaluate translations on the data base (this enables the use of an optimization algorithm before executing the queries).

ACKNOWLEDGMENTS

The author would like here to thank F. GUENTHNER for his help during the specification of these extensions and for his comments on this paper.

REFERENCES

- [COLMERAUER & al,73] A. COLMERAUER, H. KANOUI, R. PASERO, P. ROUSSEL: "Un système de communication homme machine en Français". GIA, Université de Marseille 1973
- [COLMERAUER, 79] A. COLMERAUER, "An interesting subset of natural language". in Logic Programming, pp. 45-66, eds: Clark & Tarnlund Academic Press, 1982.
- [GUENTHNER & LEHMAN, 85] F. GUENTHNER, H. LEHMAN: "A theory of the representation of knowledge". IBM Heidelberg report, 1985
- [KAMP, 81] H. KAMP " A Theory of Truth and Semantic Representation" in Formal Methods in the Study of Language Amsterdam, 1981
- [KLEIN, 84] E. KLEIN: "VP ellipsis in DR theory" report, CSLI, 1984
- [MONTAGU,74] R. MONTAGU : " The proper treatment of quantification in ordinary English" in Formal Philosophy : selected papers of Richard MONTAGU. RICHARDSON ed., Yale University Press, 1974
- [PASERO, 73] R. PASERO: "Représentation du Français en logique du premier ordre en vue de dialoguer avec un ordinateur" These de 3ème cycle, Marseilles, 1973.
- [PEREIRA & WARREN, 82] F. PEREIRA, D. WARREN "AN Efficient and easily adaptable system for interpreting Natural Language Queries" American Journal of Computational Linguistics vol.8, n° 3-4 1982.
- [SAGER,81] N. SAGER: "Natural language information processing: a computer grammar of English and its applications". Addison-Wesley Publishing company, 1981
- [SALKOFF,73] SALKOFF M. "Une grammaire en chaîne du français" Editions Dunod, Paris, 1973.
- [SEDOGBO, 85] C. SEDOGBO: "The semantic representation of discourse in a natural language application" report BULL, 1985.