INTERFACES FOR STANDARD ARABIC IN PROLOG

A.BETARI*, P.COTE**, S. EL-KAREH***

*Dépt. Informatique Luminy,GIA
163, av.de Luminy,case 901
13288 Marseille cedex 9, France
** Université du Québec à Rimouski
300 allée des Ursulines
Rimouski (Québec)
Canada
G5L 3A1
***Faculty of Arts
Alexandria University
11, Mohamed Massoud street
Wabour El Miyah
Alexandria, Egypt

ABSTRACT

Our main objective is to deal with databases to satisfy the needs of the arab world. The conceptual frame-work we adopted is logical grammars (Modifier Logic Grammar) (MLG) and the Programmation language used is (Prolog II+ and Arity PROLOG).

INTRODUCTION

As all languages of the world , arabic has to be able to face all the technological development of the next decade. The use of the tools of information process will enable the language of the old civilization to be the major vehicle and an important factor for the economical development of the nations using it.
Specially interested in the automatic processing of the arabic language, we have tried to investigate the 3 levels: morphology, syntax and semantics.
Even though arabic has a very simplified conception of uniform structures, explicit declination system, it has not been optimized until now in the different technological developments. A most critical issue in the " informatization " of the modern society is the elaboration of the computer tools for the arabic language in order to reach the most natural man - machine communication.
Our main objective is to deal with databases to satify the needs of the arab world. The conceptual frame-work we adopted is logical grammars (Modifier Logic Grammar) (MLG) and the language used is (Prolog II+ and PROLOG Arity).
In the first stage, we have tried to define the lexical information needed to implement our database in order to facilitate the manipulation of the morphological and syntactic analyzer. Emphasis was put on the semantic formulation of the sentence.
We will try to show to which extend the adaptability of the G.E.N.I.A.L. system (Pelletier et Vaucher, 1986] is possible. This system, conceived and realized in Québec, is based on a set of software tools and techniques to control the different interfaces that has been integrated successfully to different french and english systems.
We will try to show that the different logical grammars such as Definite Clause Grammars (DCG)) [Pereira 1980], Extraposition Grammars (XG) [Pereira 1981], and the semantic representation in logical formula (Walter, McCord, Sowa and Wilson 1987; Polguère, 1984] are also suitable for the arabic language.
In this part we want to precise the different concepts in the field and the various problems encountered for adapting the different systems for the use of the arabic language.
First of all we should define what we mean by "INTERFACE". Pelletier says" Une interface homme-machine d'une base de données est un module de traduction d'une phrase en langue naturelle en une formule (requête ou commande) respectant la signification initiale de la phrase, mais qui soit directement évaluable (exécutable) dans le langage d'interrogation de la base" (1)
This meant that any system could be transfered to deal with any language. But the reality is that this was not directly transferable for the arabic language. The new formalisms developed which has been described and integrated to the PROLOG have enabled us to describe syntactic and semantic mecanisms in the language.
One of the most important application of logic grammars was done by Dahl (1981) to question databases in spanish. This system has been also adapted for the use in a library using portugese language.

The structure of G.E.N.I.A.L. is based on different levels of information:

1) A syntactic Parser.
2) Several tools and modules are needed by the Parser such as a dictionary, a module to recuparate errors, a module for the editing...
3) All tools necessary for adaptability and transportability. Figure (1) gives an idea how G.EN.I.A.L. works.

But after studying the detailed components of GENIAL we arrived to the conclusion that a simple transfer was not possible. The differences between arabic and french were far too important to enable us to benefit directly from GENIAL. Re-writing our formal grammar became a necessity. We will try to present the different modules of our interface. The first one deals with the formalism used. It is the one introduced by the Modified Logic Grammar (MLG) of McCord (1987 b).

Even if our ultimate goal is to build a grammar in order to describe the STRUCTURES OF INTERROGATIVES independly of the context used in, we had to choose a certain domain and build a " mini-database" for testing purposes. Our choice was made on UNIVERSITY ENVIRONMENT i.e. Faculty members, students, courses and departments.

SPECIFIC PROBLEMS TO ARABIC

Before exposing the diferent components of our interface we would like to talk about certain problems specific to arabic language.

a) Lemmatization

What do we mean by lemmatization?

Lemmatization is the process to group different forms of occurences of a certain word under a canonical form. This is possible because our objective is not to study the different values given by the distinction of the inflected forms. Our need is to relate at a certain stage different forms to the predicate containing the syntactic and semantic information, to a terminal symbol. (2)

The recognition of the word in arabic is not an easy problem to solve. Space delimiters are inadequate because several affixes could be joined to one word. We had two solutions either to include a morphological analyzer or to adopt a simple frame-work to determine the affixes that are potentially joined to our forms and try to find generalizations and formulate certain rules.

We have opted for the second solution because the first is time-consuming in performing the user's request. We refer the reader to our article (3) where we introduced the concordances program and its use in the arabic text processing. And the second reason is that here is on the market certain morphological analyzer integrated in several software related to the Quoran so it would be useless to redo the work.

2) Extraposition

Linguists usually talk about extraposition when a word or a group of words appear in front of their normal position in the sentence. We have to recall that in arabic, complements usually follow he verb.

Since arabic is a language permitting different word order we had to take it in consideration when structuring our grammar. The extraposition grammars of Perreira systematically deals with this problem. But we found that the implementation of such grammar would add to the complexity of the grammer.

We have limited ourselves with 2 conditions in the extraposition:

1) The extraposed element should be unified with the argument of the predicate of the verb corresponding to the role played by this item.

2) The second type of extraposition is related to the relative phrases:

ex: (almawaad allati yudarisuhaa al? ustaaz mahdi

المواد التي يدرسها الاستاذ مهدي

The role played by the relative is detected by the presence or the absence of the pronoun (suffixe) joined to the verb. If there is a suffixe the role of the relative is determined as object.

In our grammer the "slots" help verify if the extraposed element fits or not, after satisfying these conditions the analysis is performed.

Non-contextual grammars are inadequate for natural language processing, modifications have been brought in order to answer to different constraints.

The simple modifier logic grammar (SMLG) is a simplified version of the MLG meaning that the grammar as a whole is not used only needed rules to deal with the question / answer system are retained and some idiosyncratic features of the English language are disregaded in our formal system (i.e. multiple possessions structures: my brother's teacher's wife).

CONCEPTUAL FRAME WORK.

Our gammar consists essentially of a syntactic and a semantic component:

A) SYNTACTIC COMPONENT:

The output of this component is:
-A syntactic tree
- A restructuring of the tree in order to keep track of the dominance and the scope of the quantifiers and determiners.
-An analysis based on the tree, produces the logical formula needed for the semantic component.
The most important modification to the rules of MLG in comparison to the Definite Clause Grammars is the adjunction of the 2 arguments to every non-terminal in order to represent of the sentence dealt with.

*****

Our mini-database is represented by six types of predicates:

fac(NoProf, NomProf, Dept, EtatCivil).
étudiant (NoEtudiant, étudiant, Dept, Année).
dept(Nom).
directeur(Nom,NomDudépartement).
cours (NoCours,Titre,Dept,NoProf).
inscription(NoEtudiant,NoCours).

Our grammar is able to process the following types of questions:

I. Y/ N QUESTIONS

In arabic these questions are generally introduced by particles like (/?a/ and /hal/) followed either by a verb or by a (thematic group) nominal sentence. The nominal sentence are of the following types:

a) Participle	[hal nabil mossagal fi .....]	..... هل نبيل مسجل في
b) Predicative	[hal nabil telmiz.....]	..... هل نبيل تلميذ
c)Prepositional phrase	[hal nabil fi qism.....]	..... هل نبيل في قسم
d)Noun phrase (Idasfa)	[hal nabil ra?is qism.....	..... هل نبيل رئيس قسم

II. OPEN QUESTIONS

1) /man/ AND /maadha/

Our grammar is able to recognize questions like:

a) [man yar?as qism.....] ..... من رئيس قسم
b) [maadha yudarris alostaaz.....] ..... ماذا يدرس الاستاذ

It will also recognize sentences like:

c) [man howa ra?is qism.....] ..... من هو رئيس قسم
d) [ma hiwa almawad allati yudarrissuha ?alostaaz.....] ..... ما هي المواد التي يدرسها الاستاذ

2) /?ay/ and participle + /?ay/

a) [?ay maada rassaba fiha .....] ..... اي مادة رسب فيها
b) [fi ?ay qism ya:mal alostaaz.....] ..... في اي قسم يعمل الاستاذ

3) NUMBER / kam / AND / maa.caadad /

a) [kam (min) maaada yudarris alostaaz.....] ..... كم (من) مادة يدرس الاستاذ
b) [maa (howa) cadad ?aqsaam alkoliyah.....] ..... ما (هو) عدد اقسام الكلية

III. The sentences beginning with " I want "

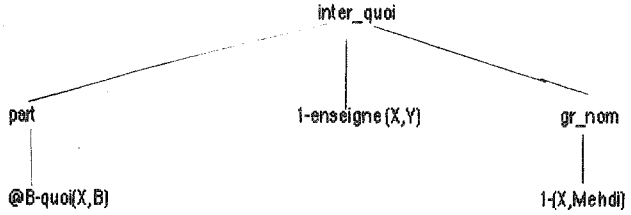
The " I want sentences " are the sentences which have the form of a declarative structure with the illocutionary force of a question.

[?urido ?an ?acref man yudarrissu madat .....] اريد ان اعرف من يدرس ماده
[I want to know who teaches the course.....]

There has been several comments when we exposed our system to informants , the most important one, I think, was raised by an arabic professor who didn't want any toleration in the correctness of the arabic formulation of the request.
In other words which degree of grammaticality should be tolerated and when should we reject the question?
We think that only formulation leading to mis-interpretation should be refused otherwise all requests should be accepted.

B) THE SEMANTIC COMPONENT:

The semantic component performs the semantic analysis through certain predicates. The first argument of the predicate is composed of the element SYN's output of the syntactic analysis and the logical formula representing the meaning of the sentence



The semantic predicate representing the whole sentence is
sem(inter_quoi,1,quoi(Y,X=Mehdi et enseigne(X,Y))
Our interface is implemented with different modules which render the use of the program more friendly.

SOFTWARE AND HARDWARE USED.

The interface has been developed in two different systems in order to facilitate the access:
-Arity Prolog with an arabic MS-DOS (version 3.3 Alis Technologies).
-The system works on micro-computer IBM or compatible with live memory of 640 K.
and it has been also developed in PROLOG II+ (developed by Colmaurer in Luminy, Marseille)
--The system used is Mac II.
The two interfaces are dealing with the same kind of data but with a difference that with PROLOG II we have to give a transcription of the question while Arity PROLOG accepts the arabic characters directly.

CONCLUSION

The project has helped in developing different tools used as aids for the linguistic analysis and has enabled to evaluate the different formal grammars and different form of arabization. But we are still hoping to have an arabic computer and not an arabized system.
Our grammar should be able to cover all the interrogatives structures of arabic.
A module should be added whose function is to detect and display the source of error in the request.

NOTES

1)PELLETIER, B. (1986). *Système d'interrogation de banque de données en langue naturelle*, mémoire de maîtrise, Université de Montréal, p.3.

2) Marie-Anne, "La Lexicométrie" to be published.
3) EL KAREH, S. COTE, P., MAAMOURI, M., (1992) De la prédictibilité au déterminisme, C.I.L., Québec,

ACKNOWLEDGMENTS: This research has been developed with the financial support of the Agence de coopération culturelle et technique ACCT, the Informatics research center of Montréal, and the Universities of Alexandria and Québec at Rimouski.

REFERENCES

ARITY CORPORATION (1988). *The Arity/Prolog. Language Reference Manual.* Mass: Arity Corporation.

COELHO, H.M.F. (1979). A Program Conversing in Portuguese Providing in Library Service. Ph.D. thesis, University of Edimburgh.

COLMERAUER, A. (1973). *Un système de communication homme-machine en français.* Groupe d'Intelligence Artificielle, Université d'Aix-Marseille.

COLMERAUER, A. (1978). Metamorphosis Grammars dans Bolc, L., Ed., Natural *Language Communication with Computers*, Springer-Verlag, New York, 133-189.

DAHL, V. (1981). Translating Spanish into Logic through Logic. *American Journal of Computationnal Linguistics*, 13, 149-164.

DAHL, V. et Mc CORD, M. (1983). Treating Coordination in Logic Grammars, *American Journal of Computationnal Linguistics*, 9:2, 69-91.

EL KAREH , S., *L'interrogation dans le dialecte égyptien. Une analyse pragmatique*, Thèse , Université d'Alexandrie, 1981.

EL KAREH, S. et EL-SAYED, N. ,T.GHOWEIL,(1989): A parser for analysing declarative sentences in egyptian newpaper, a new approach, *Cahier du participant, Colloque international sur l'informatique cognitive des organisations; l'impact de l'intelligence artificielle et des sciences cognitives dans les organisations des années 90*, Québec: Ed. GIRICO, 172-175.

EL KAREH, S., EL-SAYED, N., A model for a parser for declarative sentence, *Second Conference on Arabic Computational Linguistics*, KOWEIT, 1989

EL KAREH, S. COTE, P., MAAMOURI, M., "Development of computer tools to analyze arabic questions", in *Planning for the Informatics Society, the 12 National Computer Conference and Exhibition, Conference Proceedings, vol.II*, King Saud University, Riyadh, Saudi Arabia, October 21-24, 1990.

EL KAREH, S., De la prédictibilité au déterminisme, vers une analyse automatique de l'arabe, *Colloque texte et Contexte, Université du Caire*, Caire , 1991

EL KAREH, S. COTE, P., MAAMOURI, M., De la prédictibilité au déterminisme, C.I.L., Québec, 1992

J.-H. JAYEZ, P. LEGRAND, Y. SIMON, *Outils informatiques pour le traitement de la langue française*, in I.C.O.' 89, Québec, Canada.

GAZDAR, G., MELLISH, C., *Natural Language Processing in PROLOG*, Addison-Wesley Publishing Company, Wokingham, England, 1989, pp. 504

GIROUX, S. (1988). *Interfaces en langue naturelle, GENIAL II*, Centre canadien de recherche sur l'informatisation du travail et Université de Montréal, document no 200.

LALLICH-BOIDIN, G., *Analyse syntaxique automatique du français*, Thèse de Doctorat, Grenoble, 1986

MANKAI, C. et MILI, A. (1989). Analyse automatique de la langue, *Informatique cognitive des organisations* (textes réunis par B. MOULIN et G. SIMIAN), L'Interdisciplinaire (Informatique), 62-84.

Mc CORD, M.C. (1982). Using Slots and Modifiers in Logic Grammars for Natural Language, *Artificial Intelligence*, 18, 327-367.

Mc CORD, M., SOWA, J., WILSON, W. et WALKER, A. (editor) (1987). *Knowledge Systems and Prolog, a logical approach to expert systems and natural language processing*, Mass. Addison-Wesley.

MELONI, H. (1983). Traitement des contraintes linguistiques en reconnaissance de la parole, *TSI*, 2: 1, 349-363.

PELLETIER, B. (1986). *Système d'interrogation de banque de données en langue naturelle*, mémoire de maîtrise, Université de Montréal.

PELLETIER, B. et VAUCHER, J. (1986). GENIAL: un générateur d'interfaces en langue naturelle, *Actes de la sixième conférence canadienne sur l'intelligence artificielle*, École polytechnique de Montréal, 21-23 mai, 235-239.

PEREIRA, F. (1981). Exposition Grammars, *American Journal of Computationnal Linguistics*, 7: 4, 243-256.

PEREIRA, F. et WARREN, D. (1980). Definite Clause Grammars for Language Analysis, *Artificial Intelligence*, 13, 231-278.

PIERREL, J.-M., *Dialogue oral Homme-Machine*, Hermès, Paris, 1987, pp. 239.

PIQUE, J.F. (1981). *Sur un modèle logique du langage naturel et son utilisation pour l'interrogatio des banques de données*, thèse de doctorat de 3e cycle en Intelligence artificielle, Université d'Aix-Marseille II, Faculté des sciences de Luminy.
POLGUÈRE, A. (1984). *Programmation logique des interfaces*, Rapport de stage aux Laboratoires de Marcoussis, France.
J.RAOULT, Linguistique automatique et informatique documentaire, Colloque Franco-Anglais, Déc. 1984
ROUSSEL, P.L. (1975). *Prolog. Manuel de référence et d'utilisation*. Université d'Aix-Marseille.
SAAD, A.M. (1986). Arabic Language Parser, *International Journal of Man-Machine-Studies*, 25, 593-611.
SIROIS-DUMAIS, R. (1989a). Projet: "Transfert technologique d'outils informatiques de langue française pour utilisation en langue arabe". *Dossier de faisabilité*. Université du Québec à Rimouski, mars 1989, 31 p.
SIROIS-DUMAIS, R. (1989b). *Elaboration d'outils informatiques pour l'analyse des interrogations arabes*. Université du Québec à Rimouski, août 1989, 49 p.
SMART, J.R. (1986). *Arabic*, Teach yourself books, Holder and Stoughton.

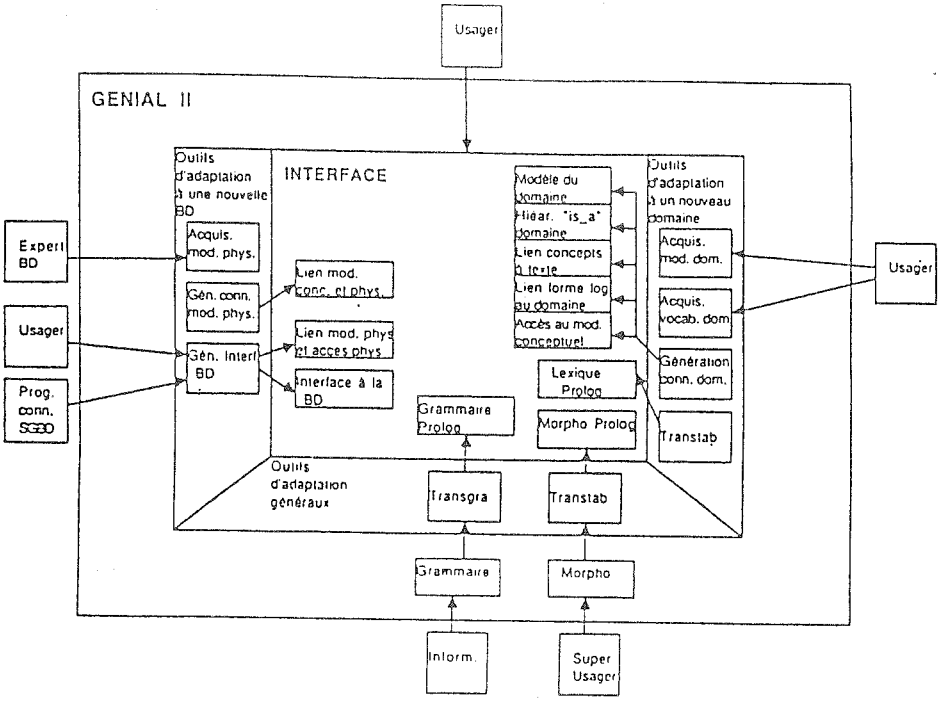


Figure 1 - Architecture de GENIAL (version II) (tirée de S. Giroux, 1988, p. 19)