AUTOMATIC REPRESENTATION OF THE SEMANTIC RELATIONSHIPS CORRESPONDING TO A FRENCH SURFACE EXPRESSION

Gian Piero Zarri
Centre National de la Recherche Scientifique
Laboratoire d'Informatique pour les Sciences de l'Homme
54, Boulevard Raspail
75270 Paris Cedex 06
FRANCE

ABSTRACT

The work presented here is a preliminary study concerning the automatic translation of French natural language statements into the RESEDA semantic metalanguage. The text in natural language is first (pre)processed in order to obtain its syntactic structure. The "semantic parsing" process begins with marking the "triggers", defined as lexical units which call one or more of the predicative patterns allowed for in the metalanguage. The patterns obtained are then merged, and their case slots filled with the elements found in the surface structure according to the predictions associated with the slots.

I INTRODUCTION

The work * that I intend to present here is a preliminary study concerning the automatic translation of French natural language statements into the RESEDA semantic language.

The RESEDA project itself is concerned with the creation and practical exploitation of a system for managing a biographical database using Artificial Intelligence (AI) techniques. The term "biographical data" must be understood in its widest possible sense : being in fact any event, in the public or private life, physical or intellectual, etc., that it is possible to gather about the personages we are interested in. In the present state of the system, this information concerns a well-defined period in time (approximately between 1350 and 1450) and a particular subject area (French history), but we are now working on the adaptation of RESEDA's methodology to the processing of other biographical data, for example medical or legal data.

RESEDA differs from "classical" factual database management systems in two ways:

- The information is recorded in the base using a particular Data Definition Language (metalanguage) which uses knowledge representation techniques.

- A user interrogating the base obtains not only information which has been directly introduced into it, but also "hidden" information found using inference mechanisms particular to the system ; in this respect, the most important characteristic of the system lies in the possibility of using inference procedures to question the database about causal relationships which may exist between the different recorded facts, and which are not explicitly declared at the time of data entry (Zarri, 1979;1981). For example, the system may try to explain by inference top-level changes in the State administration in terms of changes in political power.

II THE RESEDA METALANGUAGE

The biographical information which constitutes the system's database is organized in the form of units called "planes". There are several different types of plane, see Zarri et al. (1977); the "predicative planes", the most important, correspond to a "flash" which illustrates a particular moment in the "life story" of one or more personages. A predicative plane is made up of one of five possible "predicates" (BE-AFFECTED-BY, BEHAVE, BE-PRESENT, MOVE, PRODUCE) ; one or more "modulators" may be attached to each predicate. The modulator's function is to specify and delimit the semantic role of the predicate. Each predicate is accompanied by "case slots" which introduce their own arguments ; dating and space location is also given within a predicative plane, as is the bibliographic authority for the statement. Predicative planes can be linked together in a number of ways ; one way is to use explicit links of "coordination", "alternative", "causality", "finality", "condition", etc. The data representation we have chosen in the RESEDA project is basically, therefore, a kind of "case grammar", according to the particular meaning attached to the term in an AI context (Bruce, 1975; Charniak, 1981; etc.).

For example, the data "André Marchant was named provost of Paris by the King's Council on 22nd September 1413 ;he lost his post on 23rd October 1414, to the benefit of Tanguy du Châtel, who was granted this office", will be represented in three planes - that of the nomination of André Marchant, his dismissal and the nomination of Tanguy du Châtel.

The coding of information must be made on two distinct levels : an "external coding, up until

* This research is jointly financed by the "Agence de l'Informatique - A.D.I." (CNRS/ADI contract n° 507568) and the "Centre National de la Recherche Scientifique - C.N.R.S." (ATP n° 955045).

now performed manually by the analyst, gives rise to a first type of representation, formalized according to the categories of the RESEDA metalanguage ; a second automatic stage results in the "internal" numeric code. The external "manual" coding of the three events just stated is given in figure 1. The code in capital letters indicates a predicate and its associated "case slots". Every predicative plane is characterized by a pair of "time references" (date1-date2) which give the duration of the episode in question. In these three planes, the second date slot (date2) is empty because their modulators (begin, end) specify a change of state associated with a punctual event. "André-Marchant" and "Tanguy-du-Châtel" are historical personages known to the system ; "provost", "king's-council" and "letters-of-nomination" are terms of RESEDA's lexicon. The classifications associated with the terms of the lexicon provide the major part of the system's socio-historical knowledge of the period. "Paris" is the "location of the object". If the historical sources analyzed gave us the exact causes of these events, we would introduce into the database the corresponding planes and associate them with these three planes by an explicit link of type "CAUSE".

```
1) begin+soc+BE-AFFECTED-BY  SUBJ  André-Marchant
                             OBJ   provost : Paris
                             SOURCE king's-council
                             date1 : 22-september-
                                      1413
                             date2 :
                             bibl: Demurger1,273
2) end+BE-AFFECTED-BY  SUBJ  André-Marchant
                       OBJ  provost : Paris
                       date1  : 23-october-1414
                       date2  :
                       bibl: Demurger1,273
3) begin+BE-AFFECTED-BY  SUBJ   Tanguy-du-Châtel
                         OBJ    provost : Paris
                         MODAL  letters-of-
                                nomination
                         date1  : 23-october-1414
                         date2  :
                         bibl: Demurger1,273
```

figure 1

This manual procedure for converting information in natural language into one or more planes has at least two major disadvantages which the proposed study intends to deal with :

- The manual representation of biographical information in the terms of the metalanguage can only be performed by a specialist. This is done, at the moment, by the researchers themselves who have constructed the prototype system. Such a method is obviously out of the question if the system is to be used routinely by an uninitiated public, especially as RESEDA was conceived as a system supplied continuously with biographical information extracted from many different sources.
- In spite of the fact that the syntax of RESEDA's metalanguage imposes strict constraints on the forming of predicative schemata accepted by the system and that these are then thoroughly checked, we cannot completely exclude the possibility of two coders translating the same information differently.

III DESCRIPTION OF THE METHOD OF AUTOMATIC CODING

To describe our methodology, I will use the example given in the preceeding section. The initial text in natural language is first (pre) processed to obtain its constituent structure. For this purpose, we have used in a first approach the French surface grammar implemented in DEREDEC, a software package developed at the University of Québec at Montréal by Pierre Plante (1980a;1980b). This system, comparable to an ATN parser, permits a breakdown of the surface text into its syntactic constituents, and establishes, between these constituents, syntagmatic relationships of the type "topic-comment", "determination" and "coordination". This preliminary analysis provides a context for subsequent processing, without necessarily removing all the ambiguities : in the same vein, see Boguraev and Sparck Jones (1982).

The specific tools that we intend to develop for this project are of two types : a general procedure which can be likened to a sort of semantic parsing, and a system of heuristic rules.

A. Semantic Parsing

The first stage of the general procedure consists of marking the "triggers", defined as lexical units which call for one or more of the predicative patterns allowed for in RESEDA's metalanguage. Thus we do not take into consideration every one of the lexical items met in the surface text, retaining only those directly pertaining to the "translation" to be done.

However, we do not limit ourselves to a simple keyword approach, since a number of operations utilizing data provided by the morpho-syntactic analysis executed by DEREDEC are necessary before the predicative patterns which will be actually used afterwards can be selected.

One of the results of the DEREDEC analysis is a kind of lemmatization enabling the reduction of surface forms in the text to a canonical form ; for example, infinitive in the case of verbs. The canonical forms found in the text under examination are compared with a list of potential triggers stored permanently in the system. In the case of the sentence we are analyzing we can construct from this list the following sub-list : verbal forms = "name", "loss", "grant" ; terms pertaining directly to the metalanguage or terms which have a direct correspondence with elements of the metalanguage : "office", synonymous with "post" in RESEDA ("post" is a "generic" term, a "head" of a "sub-tree" in RESEDA's lexicon), and its specification "provost". The results of the pre-analysis executed by DEREDEC enable the elimination of potential patterns associated with the triggers "name" and "grant" which would correspond to surface constructions of type "active", as in the hypothetical example "The Duke of Orléans named André Marchant provost of Paris ...". The patterns which will be actually utilized afterwards are therefore those shown in figure 2. Note that in the case of a trigger "name (active form)" the personage who figures as surface object would have found as the "SUBJECT" of "BE-AFFECTED-BY",whilst the surface subject would have been associated with the slot "SOURCE" of "BE-AFFECTED-BY".

```
name (passive form)  →  begin+(soc+)BE-AFFECTED-BY  SUBJ  <personage>-surface subject of the trigger
                                                    OBJ   <post>-surface complement
                                                    (SOURCE <personage>|<social-body>-surface
                                                          complement of the agent of the trigger)
                                                    date1 : obligatory
                                                    date2 : prohibited
                                                    bibl. : obligatory

provost     →   (soc+ )BE-AFFECTED-BY  SUBJ  <personage>
                                       OBJ   <post>-trigger
                                       (SOURCE <personage>|<social-body>)
                                       date1 : obligatory
                                       date2 : optional
                                       bibl. : obligatory

loss      →      end+BE-AFFECTED-BY    SUBJ  <personage>-surface subject of the trigger
                                       date1 : obligatory
                                       date2 : prohibited
                                       bibl. : obligatory

post      →      (soc+)BE-AFFECTED-BY    SUBJ  <personage>
                                         OBJ   <post>
                                         (SOURCE <personage>|<social-body>)
                                         date1 : obligatory
                                         date2 : optional
                                         bibl. : obligatory

grant (passive form)  →  begin+(soc+)BE-AFFECTED-BY  SUBJ  <personage>-surface subject of the trigger
                                                     OBJ   <post>-surface complement
                                                     (SOURCE <personage>|<social-body>- complement of
                                                           the surface agent)
                                                     MODAL letters-of-nomination
                                                     date1 : obligatory
                                                     date2 : prohibited
                                                     bibl. : obligatory

office      →      (soc+)BE-AFFECTED-BY  SUBJ  <personage>
                                         OBJ   <post>
                                         (SOURCE <personage>|<social-body>
                                         date1 : obligatory
                                         date2 : optional
                                         bibl. : obligatory
```

figure 2

In reality, the predicative structures selected are not limited to those shown in figure 1. They are in fact repeated with predicative patterns of the type "BE-AFFECTED-BY" which have as "SUBJECT" "<social-body>",and as "OBJECT" "<personage>" accompanied by the specification ("SPECIF") of a "<post>". These constructions each correspond to the description : "A personage receives a post in a certain organization (the organization in question, SUBJECT, is "augmented", BE-AFFECTED-BY, by the personage, OBJECT, in relation, SPECIF, to a given post)". A corresponding surface expression would be, for example, the following : "André Marchant (personage) is named secretary (post) of the papal court (social body)". Therefore, for example, the pattern in figure 3 is also associated with the trigger "name (passive form)". The patterns in this second set will be eliminated at the end of the construction procedure since, as it is not possible to obtain a surface realization of the concept "<social-body>" in the position "SUBJECT", they cannot provide complete predicative structures.

The last stage of the general procedure consists of examining the triggers belonging to the same morpho-syntactic environments, as defined by the results of the DEREDEC analysis. If there are several triggers pertaining to the same environment, and if the predicative patterns triggered are the same - which means that the predicates and case slots must be the same and that the modulators, dates and space location information must be compatible - then it can be said that the triggers refer to the same situation. As a

```
name (passive form) → begin+(soc+)BE-AFFECTED-BY  SUBJ   <social-body>
                                                  OBJ    <personage>-surface subject of the trigger
                                                         SPECIF <post>-surface complement
                                                  (SOURCE <personage>|<social-body>)
                                                  date1 : obligatory
                                                  date2 : prohibited
                                                  bibl. : obligatory
```

figure 3

result, the predicative patterns are merged as to obtain the most complete description possible ; the predictions about filling the slots linked with the cases of the resulting patterns together govern to search for fillers in the surface expression.

Thus, the first two triggers in figure 2, recognized as relevant to the same environment, are combined in the formula in figure 4, which gives the general framework of plane 1 in figure 1.

```
begin+(soc+)BE-AFFECTED-BY  SUBJ   <personage>-surface subject of "was named"
                            OBJ    <post>-"provost"
                            (SOURCE <personage>|<social-body>-surface complement of the
                                   agent of "was named")
                            date1 : obligatory
                            date2 : prohibited
                            bibl. : obligatory
```

figure 4

The example we are considering illustrates a particularly simple case, in which it is not necessary to establish links between the planes to be created. If we had to process the sentence "Philibert de St Léger is nominated seneschal of Lyon on the 30th of July 1412, in lieu of the late A. de Viry", three planes should be generated : one for the nomination of Philibert de St Léger, one for the death of A. de Viry, and another establishing a weak causality link ("CONFER", in our metalanguage) between the first two planes. Surface items such as conjunctions, prepositions and sentential adverbs can be used to infer links between planes : causality, finality, coordination, etc. More precisely, in the last example, "in lieu of" is a potential trigger according to the following rule : if the main noun group of the surface prepositional phrase contains a trigger, this phrase constitutes a plane environment and "CONFER" introduces the plane created.

B. Heuristic Rules

The process I have outlined so far requires a corpus of heuristic rules - organized in the form of "grammars" associated with the predicative patterns of RESEDA's metalanguage - which will enable the slots in these patterns to be filled using the surface information in accordance with the predictions which characterize the slots. In the case of the pattern in figure 4, this filling-in poses no real problems, since the surface elements "André Marchant", "provost", "King's Council" and "22nd September 1413" - standardized according to RESEDA's conventions, see figure 1 - will take up the slots "SUBJECT", "OBJECT", "SOURCE" and "date1" directly. The filling-in operations are usually much more complicated, and require the use of complex inference rules. I shall say just a few words here about the heuristic rules designed to solve cases of anaphora (as in our example, "he", "this office", "who").

In the approach that we propose, marks of anaphora are identified during the general analysis procedure ; the actual solving brings into play a number of criteria from simple pairing off and morphological agreement to more subtle criteria, like contextual proximity, persistence of theme, etc. Thus, morphological agreement and contextual proximity are used to replace "who" by "Tanguy du Châtel" in our example ; persistence of the theme enables us to fill in the missing date for Tanguy du Châtel's posting with the date "23rd October 1414" appearing in the surface expression.

We would like to integrate this approach, which has been purely empirical up to now, into the framework of a more general theory. Two directions of enquiry seem particularly interesting in order to develop our own philosophy of the subject.

The PAL system of Candace Sidner (1979;1981), is a top-down anaphora resolution method which makes use of the notion of focus (likened to the theme of the discourse). By searching in the text for "focuses" which refer to a system of representation organized as a series of "frames", it is able to solve references. If the reference is not found by using the frames themselves, it is inferred from other frames contained in the database. The interest in this study lies in the fact that RESEDA already has, as permanent data, a certain amount of general knowledge organized in a form very similar to that of frames. Thus, in my example, the nomination and dismissal of André Marchant refers to the context of the "civil war at the beginning of the 15th century" which is one of those frames (Zarri et al., 1977). The approach used by Klappholz and Lockman (Lockman, 1978) depends on the hypothesis that there is a strong link between co-reference and the cohesive links of a discourse. These links, when marked progressively in the text, become indices of the structure of the discourse, organized as a tree structure and created dynamically. These cohesive links (effect, cause, syllogism, exemplification, etc.) are very similar to the logical connections between planes in RESEDA (causality, finality, condition, etc.).

IV CONCLUSION

The study that I have described here is intended to automatically achieve a representation of fundamental underlying semantic relationships corresponding to a French surface expression. I have already pointed out the benefits that we hope to obtain from this work as far as RESEDA is concerned. I should like to add that, on a more general level, solving the problem of automatically recording natural language data would obsiously allow us to face, with a certain amount of confidence, the analogous problems of natural language interrogation of RESEDA's database ; the advantages of this, from the point of view of widespread use of the system, are obvious. But the results of this study can, in principle, be used not only in the framework of RESEDA, but in a number of different applications such as, for example, automatic abstraction, paraphrase, machine translation and the direct coding of natural language documents in a factual database.

V REFERENCES

BOGURAEV, B.K., SPARK JONES, Karen (1982) "A Natural Language Analyser for Database Access", Information Technology : Research and Development, I, 23-29.

BRUCE, B. (1975) "Case Systems for Natural Language", Artificial Intelligence, VI, 327-360.

CHARNIAK, E. (1981) "The Case-Slot Identity Theory", Cognitive Science, V, 285-292.

LOCKMAN, A.B. (1978) Contextual Reference Resolution, Technical Report DCS-TR-70. New Brunswick: Rutgers University Department of Computer Science.

PLANTE, P. (1980) DEREDEC - Logiciel pour le traitement linguistique et l'analyse de contenu des textes, manuel de l'usager. Montreal: Université du Québec à Montréal.

PLANTE, P. (1980) Une grammaire DEREDEC des structures de surface du Français, appliquée à l'analyse de contenu des textes. Montréal: Université du Québec à Montréal.

SIDNER, Candace L. (1979) A Computational Model of Co-reference Comprehension in English, Ph.D. Thesis. Cambridge: MIT Artificial Intelligence Laboratory.

SIDNER, Candace L. (1981) "Focusing for Interpretation of Pronouns", American Journal of Computational Linguistics, VII, 217-231.

ZARRI, G.P. (1979) "What Can Artificial Intelligence Offer to Computational Linguistics ?" The Experience of the RESEDA Project", in Advances in Computer-aided Literary and Linguistic Research, Ager, D.E., et al., eds. Birmingham: University of Aston.

ZARRI, G.P. (1981) "Building the Inference Component of an Historical Information Retrieval System", in Proceedings of the Seventh International Joint Conference on Artificial Intelligence - IJCAI/81 (Vancouver 1981). Menlo Park: The American Association for Artificial Intelligence.

ZARRI, G.P., ORNATO, Monique, KING, Margaret, ZWIEBEL, Anne, ZARRI-BALDI, Lucia (1977) Projet RESEDA/0: Rapport Final. Paris: Equipe Recherche Humanisme Français.