COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

CONVERSION OF A FRENCH SURFACE EXPRESSION INTO ITS SEMANTIC REPRESENTATION ACCORDING TO THE *RESEDA* METALANGUAGE

Jacqueline Léon

Centre National de la Recherche Scientifique
Laboratoire d'Informatique pour les Sciences de l'Homme
Paris, France

Daniel Memmi

Centre National de la Recherche Scientifique
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Orsay, France

Monique Ornato

Centre National de la Recherche Scientifique
Equipe de Recherche sur l'Humanisme Français des XIVe et XVe siècles
Paris, France

Joanna Pomian

Université de Paris I
Centre National de la Recherche Scientifique
Equipe de Recherche sur le traitement automatique de l'histoire du Bas Moyen-Age
Paris, France

Gian Piero Zarri

Centre National de la Recherche Scientifique
Laboratoire d'Informatique pour les Sciences de l'Homme
Paris, France

The work we describe here is a preliminary study concerning the automatic translation of natural language statements into the RESEDA semantic metalanguage. A first stage of the procedure consists in marking the "triggers", defined as lexical units which call upon one or more of the predicative patterns allowed for in the metalanguage. The predicative patterns obtained are then merged, and their case slots filled with the elements found in surface structure according to the predictions associated with the slots.

INTRODUCTION

The work that we intend to present here [1] is a preliminary study concerning the automatic translation of natural language statements - which describe the biographies of historical personages taken into account by the RESEDA system - into the RESEDA semantic metalanguage.

The RESEDA project itself concerns the creation and practical exploitation of a database containing the biographies of historical personages of the late Middle Ages in France. 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).

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 ; 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) ; to each predicate, one or more "modulators" may be attached. The modulator's function is to specify and delimit the semantic role of the predicate. Each predicate is accompanied by "case slots" which introduce the predicative 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. (Zarri *et al.* 1977).

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, realized manually by the analyst, gives rise to a first type of representation, formalized according to the categories of the RESEDA metalanguage ; a second automatic step results in the "internal" numeric code. The external "manual" coding of the three events just stated will be the following:

```
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
                         date1 : 23-october-1414
                         date2 :
                         bibl: Demurger1,273
```

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 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 analysed 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".

DESCRIPTION OF THE METHOD USED

In the field of the application of Artificial Intelligence techniques to natural language processing, from the very beginning, stress was put on the importance of semantic and pragmatic components. In this framework, creating a formal representation of the message carried by a surface expression is usually achieved by one of two methods.

The first, and most traditional, respects the usual progression of the three levels of analysis, morphological, syntactic and semantic whilst combining their results in a final interpretation : for discussion, see for example Winograd (1972), Woods (1973), Marcus (1979), etc.

Schank and Wilks, on the contrary, put forward the idea, which was subsequently taken up by many researchers, that a predominantly semantic analysis, with syntax relegated to a secondary role, was possible. The deep structure representation that is being created is thus used to make appropriate predictions about the logico-semantic function of the elements ; these expectations are progressively met during the examination of the surface structure representation, see Schank (1975), Wilks (1975), etc.

The hypothesis adopted for our project draws more from this second method, in that the structures of RESEDA's internal representation provide, beforehand, a very complete framework of the predictions which are to be a guide in scanning the text to be translated into the system's metalanguage.

To describe our approach, we will utilize the above example. The initial text in natural language is first (pre)processed to obtain its constituent structure. For this purpose, we have used the French surface grammar implemented in DEREDEC, a software package developed at the University of Québec at Montreal 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 provide 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.

Semantic parsing

The first stage of the general procedure consists of marking the "triggers", defined as lexical units which call 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 ; this is not without similarity with the "skimming" found in DeJong (1979a, 1979b).

However, we do not limit ourselves to a simple keyword approach. Certain lexical items are potential triggers, but their actual triggering in a given context depends on rules using both the morpho-syntactic analysis provided by DEREDEC and the socio-historical knowledge stored in the RESEDA system. These rules intervene at this stage to decide whether triggering should take place and to choose the predicative patterns. In the sentence given, belong to the list of potential triggers the verbal forms : "named", "lost", "granted" ; terms pertaining directly to the metalanguage : "office", synonymous with <post> in RESEDA, and its specification "provost" ; date elements : "september", "october". After applying the rules, the following patterns have been triggered :

```
was named   →  begin+(soc+)BE-AFFECTED-BY SUBJ <personage>-surface subject of the
                                                trigger
                                          OBJ   <post>-surface complement
                                          (SOURCE <personage>|<social-body>-sur-
                                                face 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

22 september 1413  →  date a

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

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

23 october 1414  →  date b

was granted  →  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
```

The second stage of this general procedure consists of examining the triggers belonging to the same morpho-syntactic environment. If there are several predicate triggers in the same environment, and if the predicates triggered are the same - which means that the predicates and case slots must be the same and that the modulators, dates and the space location information must be compatible - then it can be said that the triggers refer to the same situation. As a 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 the search for fillers in the surface expression.

Thus, the first three triggers of the example, recognized as relevant to the same environment, are combined in the following formula :

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

The units of the surface expression corresponding to the predictions of the pattern obtained are then retrieved and standardized according to RESEDA's categories imposed by the pattern (André Marchant : André-Marchant, personage ; provost : provost, post ; King's Council : king's-council, social-body, etc.). Eventually, we obtain plane 1 in André Marchant's biography.

The example we have shown illustrates a particularly simple case, in which it is not necessary to establish links between the planes 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 one 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.

Heuristic rules

The process we have sketched so far requires a corpus of heuristic rules, to solve ambiguities which are left aside by the prediction system - which cannot go beyond the capabilities of RESEDA's predicative patterns.

We shall say just a few words about the heuristic rules designed to solve cases of anaphora (as in our first example, "he", "this office", "who").

In the approach that we propose, marks of anaphora are identified during the general analysis procedure with unassumed predictions, triggering the appropriate heuristic rules. The actual solving, after validation of the marks, brings into play a number of criteria from simple pairing off and morphological agreement to more subtle criteria, like contextual proximity, persistance of theme, etc. Thus, morphological agreement and contextual proximity are used to replace "who" by "Tanguy du Châtel" in our first example ; persistance of theme enables us to make up for the missing date of Tanguy du Châtel's posting by date b in the list of triggers.

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, 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 (Sidner 1978, 1979).

The interest for our 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 our 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.

The approach used by Klappholz and Lockman depends on the hypothesis that there is a strong link between coreference and the cohesive links of a discourse. These links, when marked progressively in the text, become the indices of a structure of the discourse, organized as a tree structure and created dynamically (Lockman 1978). These cohesive links (effect, cause, syllogism, exemplification, etc.) are very similar to the logical connections between planes in RESEDA (causality, finality, condition, etc.).

CONCLUSION

The study that we describe here is intended to automatically attain a representation of fundamental underlying semantic relationships corresponding to a French surface expression. These results 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, direct encoding of natural language documents in a factual database.

FOOTNOTES

[1] This research is jointly financed by the "Agence de l'Informatique - A.D.I." and the "Centre National de la Recherche Scientifique - C.N.R.S.". The project leader is Gian Piero Zarri.

REFERENCES

[1] Boguraev, B.K. and Sparck Jones, Karen, A Natural Language Analyser for Database Access, Information Technology : Research and Development 1 (1982) 23-39.

[2] DeJong, G., Skimming Stories in Real Time, Ph.D. Thesis, Yale University Computer Science Department, New Haven (1979).

[3] DeJong, G., Prediction and Substantiation : A New Approach to Natural Language Processing, Cognitive Science 3 (1979) 251-273.

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

[5] Marcus, M., A Theory of Syntactic Recognition for Natural Language (MIT Press, Cambridge, 1979).

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

[7] Une grammaire DEREDEC des structures de surface du français, appliquée à l'analyse de contenu des textes, Université du Québec à Montréal (1980).

[8] Schank, R.C., ed., Conceptual Information Processing (North-Holland, Amsterdam, 1975).

[9] Sidner, Candace L., The Use of Focus as a Tool for Disambiguation of Definite Noun Phrases, in: Waltz, D.L. (ed.), Theoretical Issues in Natural Language Processing - 2 (ACM, New York, 1978).

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

[11]Wilks, Y., A Preferential, Pattern-Seeking Semantics for Natural Language Inference, Artificial Intelligence 6 (1975) 53-74.

[12] Winograd, T., Understanding Natural Language (Academic Press, New York, 1972).

[13] Woods, W.A., An Experimental Parsing System for Transition Network Grammars, in: Rustin, R. (ed.), Natural Language Processing (Algorithmics Press, New York, 1973).

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

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

[16] Zarri, G.P., Ornato, Monique, King, Margaret, Zwiebel, Anne, Zarri-Baldi, Lucia, Projet RESEDA/0 : Rapport Final, Equipe Recherche Humanisme Français, Paris (1977).