

FROM HISTORY TO COMPUTER SCIENCE : A FORMALIZATION OF THE INFERENTIAL
PROCESSES OF AN HISTORIAN

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

The object of the present paper is to discuss the results of the work carried out to define as strictly as possible the formal system of representation of the inference procedures ("hypotheses") characteristic of the RESEDA project. RESEDA is a question-answering system which operates on a data base containing biographical information concerning characters involved in the medieval history of France.

1. INTRODUCTION

1.1 The RESEDA project aims to set up an advanced question-answering system operating upon a biographical data base. The information housed within this base concerns characters involved in the medieval history of France and their socio-historical background. Strictly formulated questions are presented to the system; responses, given in the same format, are obtained by direct-search through the data base or by inference procedures. The objectives of the first RESEDA/0 step of the project* were to construct a general framework for the system and to verify the adequacy of descriptive metalanguage chosen for dealing with this special sort of data. A new phase RESEDA/1 has just recently commenced. It strives towards two goals : a) making the system fully operational, in progressing beyond the phase of demonstration tests to approximate a routine practical use of the data; b) examining in theoretical and practical detail the possibility of a partially automated construction of the inductive "hypotheses" which support the system's inference operations.

1.2 In preparing for the latter, the RESEDA team has had to restudy in depth the work already accomplished on hypotheses in the first phase of the project. It was in effect necessary to establish on a more strictly defined basis, the formal conventions under which the hypotheses are drawn up, and to define these conventions down to the smallest detail. This paper relates

to the results of this thorough work. With regard to the metalanguage used to represent biographical data, discussed in the IJCAI/77 report [3], I will give no more information than is strictly necessary for the full understanding of my talk; the interested reader is referred to recent papers on this subject [3,6,7].

2. AN EXAMPLE OF AN HYPOTHESIS

2.1 Fig. 1 shows the formal representation of one of the most simple hypotheses in the RESEDA system, the "hypothesis of a gift to a religious community".

2.1.1 Every hypothesis -inference procedure which can be interpreted in KRL "servant-triggers" terms, see [1]- consists of two separate parts, the "premiss" and the "condition".

2.1.1.1 The premiss provides the general framework capable of being adapted to the formal representation of any question, the answer to which can be found in the hypothesis under examination. If we use the distinction introduced by Smith [5] between "abstraction" and "generalization", we could say that the formal expression of the premiss is obtained by the abstraction of specific formulations, in RESEDA's metalanguage, of different questions which can be associated with the hypothesis.

2.1.1.2 The condition is formed of a sequence of theoretical schemata ("condition schemata") of "planes", linked by "and", "or", "not". The term "plane" [4] is reserved for expressing the representation in the data base, in terms of the metalanguage, of each elementary "episode", each "fact" -expressed in natural language- which make up the fundamental stages for a biography. For an example, refer to the episode from Pierre l'Orfèvre biography further below : "It is known

* RESEDA/0 was supported by a grant from the "Délégation Générale à la Recherche Scientifique et Technique" (CNRS-DGRST Contract n° 75.7.0456); the new RESEDA/1 phase is financed by the "Institut de Recherche d'Informatique et d'Automatique" (CNRS-IRIA Contract n° 78.206).

Hypothesis of a gift to a religious community

premiss : α

a) $/-/-/k/-/-/$
 BE-AFFECTED-BY SUBJ $\underline{1}$
 OBJ $\langle \text{good} \rangle$
 MODAL $\langle \text{acquisition-by-the-donor's-initiative} \rangle$
 SOURCE $\underline{V1}$

restrictions on the variables of the premiss schemata :

$\underline{k}$ = $\langle \text{geographic-localisation} \rangle$

$\underline{1}$ = $\langle \text{religious-community} \rangle$

condition :

$(A \vee B \vee (C \wedge D) \vee E \vee (F \wedge D))$

A) $/-/-/k/-/-/$
 BE-PRESENT SUBJ $\underline{V1}$

B) $/-/-/-/k/-/-/$
 BE-AFFECTED-BY SUBJ $\underline{V1}$
 OBJ $\underline{q}$

C) $/-/-/-/k/-/-/$
 BE-AFFECTED-BY SUBJ $\underline{V2}$
 OBJ $\underline{r}$

D) $(\underline{V1} \text{ COORD1 } \underline{V2}) \text{ SPECIF } \langle \text{kinship} \rangle$

E) $/-/-/k/-/-/$
 soc+BE-AFFECTED-BY SUBJ $\underline{1}$
 OBJ $\underline{V1}$ SPECIF $\langle \text{member-of-an-organization} \rangle$

F) $/-/-/k/-/-/$
 soc+BE-AFFECTED-BY SUBJ $\underline{1}$
 OBJ $\underline{V2}$ SPECIF $\langle \text{member-of-an-organization} \rangle$

restrictions on the variables of the condition schemata :

$q \wedge r = \text{vault } V \langle \text{mark-of-identification} \rangle$

figure 1

that around 1360 Pierre l'Orfèvre was a member of the chapter of Senlis cathedral, as their dean", which will give rise to the plane 3 in fig. 3. Translation from episode to plane is done by separating out spatio-temporal data plus bibliographic references (the "thematic") from the description proper of the episode. In the "description part" of plane 3, BE-AFFECTED-BY is the "predicate" used in the translation of the episode; SUBJ, OBJ, are the "main cases" while SPECIF simply introduces a degree of precision with respect to the argument JEAN-L'ORFEVRE which fills the slot OBJ. "soc" is a "modulator" which enables us to specify the fact that the "state" defined by the predicate is inserted into a socio-professional context; "const" carries the meaning of "verified fact".

Returning now to the condition, the information which it enables us to find in the system's data base provides a possible interpretation of the fact (episode) represented by the specific plane to which the question has been reduced. The formal expression of the condition is obtained, by abstraction, from the coding relating to the planes which are shown to be capable of fulfilling an *a priori* undefined number of specific actualizations of the premiss.

2.1.2 With regards now to the formal details of the representation of fig. 1, let me take the premiss first. The symbols $\underline{k}$, $\underline{1}$ and $\underline{V1}$ which appear in the "scheme of premiss", α , are "variables of the premiss schemata". In order to be bound, these variables must satisfy the "restrictions" connected with them (the restriction on $\underline{V1}$ is shown implicitly: variables of the type Vn can only be satisfied by the name of one of the personages recognized by the system).

Going on to the condition: the symbol " $\wedge$ " which joins the two "condition schemata" C, D (F, D) has its usual meaning of "C and D" ("F and D"); "A $\vee$ B" has the meaning of "A or B". The condition must therefore be interpreted in this way: an explanation of the episode represented by the premiss filled with elements drawn from the question, can be given by the planes of the data base which can be adapted to the "condition scheme" A, or to scheme B, or to two schemata C and D -which must in this case be satisfied together- or to scheme E, or to two schemata F and D.

2.2 Imagine now a question like "Why did Jean l'Orfèvre at some unspecified time, but certainly before August 1412, the time of his death, give a latin bible by legacy to the Chapter of Senlis cathedral?"; the question would be coded as in fig. 2. A reply obtained by direct-match would show that it is possible to retrace in the data base a plane -containing at least all the information distinctly specified in the question- which should already be *explicitly* connected by a "pointer-correlator" [7: 16-17] of the type CAUSE to a second plane, or a succession of planes, which would so match with the variable x .

/avant-août-1412/b//Senlis/b/Senlis/b/
 BE-AFFECTED-BY SUBJ cathedral-chapter
 OBJ bible SPECIF latin
 MODAL legacy
 SOURCE JEAN-L'ORFEVRE
 CAUSE1 ?x

figure 2

Let us now imagine that this explicit link does not exist; RESEDA will have to resort to the system of hypotheses to reply to the question

asked.

The predicate present in the question (in our case BE-AFFECTED-BY) identifies the set of inference procedures which must be examined in order to see if there exists at least one hypothesis whose premiss can *a priori* match the type of question proposed. When this test has been passed, a real match between the question and the premiss is required. In our case, and referring back to fig. 1 and 2, the result obtained will be that from this point onwards the variables *k*, *l* and *V1* will be connected to "Senlis" (the place associated with the "subject" of the attribution in the question thematic), to "cathedral-chapter", and to "Jean l'Orfèvre".

Research on the data base of planes which can satisfy the six condition schemata begins with an examination of the "volume" associated with "Jean l'Orfèvre" (volumes are permanent files -a file for each personage recognized by the system- containing "labels" identifying the planes in which the personage in question appears), in order to see if it is possible to retrace the planes of the form A, B, D and E -the condition schemata in which the (bound) variable *V1* is present. In our case, it can be assumed that the search will first retrace the plane 1 in fig. 3: "Jean l'Orfèvre lived in Senlis (location of the subject) for a period of which the starting date (code "-") is not known but which ended in August 1412".

```
1) /-/août-1412//Senlis/b/b/b//Bibl:Bozzolo/
    BE-PRESENT      SUBJ   PIERRE-L'ORFEVRE
2) (JEAN-L'ORFEVRE COORD1 PIERRE-L'ORFEVRE)
    SPECIF kinship
3) /vers-1360/b//Senlis/Senlis/b/b/Bibl:Bozzolo/
    const+soc+BE-AFFECTED-BY
    SUBJ cathedral-chapter
    OBJ  JEAN-L'ORFEVRE
    SPECIF dean
```

figure 3

Secondly, we will have a series of "relation planes" (special planes without thematics) of the form D, which provide the personages *V2* who are relatives of Jean l'Orfèvre. A couple of planes which satisfy ($F \wedge D$) are planes 2 and 3 of fig. 3: "Around 1360 Pierre l'Orfèvre, who is some non-specified relative to Jean l'Orfèvre, *V1*, was dean (specific term of the lexical tree <member...>) of the chapter of Senlis cathedral". The planes 1, 2 and 3 together form the reply to the question in fig. 2; the possible reasons for the gift to Senlis cathedral are therefore that Jean l'Orfèvre was an inhabitant of this town and that his family had special relations with the cathedral, given that one of his relatives had held an important position within the chapter.

3. CONCLUSION

Detailed analysis of the hypothesis "gift", proves clearly the *ad hoc* nature of RESEDA's inference corpus. However, this *ad hoc*ness does not constitute a limit on the practical utility of the system. The degree of complexity and subtlety which can be reached by the inference procedures is sufficient to cover a very large number of possible deductions. However, it cannot be denied that, at the present, RESEDA is a system containing a limited number of non-evolving structures (hypotheses), whose capacity to enrich the system by finding new relations between characters or the situations in which these characters are involved, is restricted to a closed set of pre-determined contexts.

Thus the new phase RESEDA/1 of the project intends to use a semi-automatic technique for the formation of the hypotheses. Such a mechanism should help to alleviate the handicap mentioned above, at least in the sense of facilitating a rapid and continual development of the inference corpus, if only at the quantitative level. The procedure involved should be the classical one of learning by examples, see [2]. The intended mechanism should be capable of utilizing the syntactic-semantic properties of RESEDA's formal system, in order to create the framework enabling the association of the two poles, question and answer, of the new hypothesis which is to be constructed.

REFERENCES

- [1] Bobrow, D.G., Winograd, T. "An Overview of KRL, a Knowledge Representation Language." *Cognitive Science*. 1:1 (1977) 3-36.
- [2] Hedrick, C.L. "Learning Production Systems from Examples." *Artificial Intelligence*. 7:1 (1976) 21-49.
- [3] King, M., Ornato, M., Zarri, G.P., Zarri-Baldi, L., Zwiebel, A. "Ghosts in the Machine: An AI Treatment of Medieval History." In *Proc. IJCAI-77*. MIT, Cambridge, Mass., August, 1977, pp. 873-879.
- [4] Quillian, M.R. "Semantic Memory." In *Semantic Information Processing*, Minsky, M. ed. Cambridge Mass.:The MIT Press, 1968, pp. 216-270.
- [5] Smith, B. "How is a Knowledge Representation System like a Piano?" *Artificial Intelligence Laboratory*. AI memo 497, MIT, Cambridge, Mass., 1978.
- [6] Zarri, G.P. "Sur le traitement automatique de données biographiques médiévales: le projet RESEDA." In *Computing in the Humanities*, Lusignan, S., North, J.S. eds. Waterloo, Ontario: University Press, 1977, pp. 151-161.
- [7] Zarri, G.P., Ornato, M., King, M., Zwiebel, A., Zarri-Baldi, L. "Projet RESEDA/0: Rapport final." *Equipe Recherche Humanisme Français*. CNRS, Paris, December, 1977.