

THE RESEARCH PROJECT "ANAPHORA" (IN ITS PRESENT STATE OF
ADVANCEMENT)

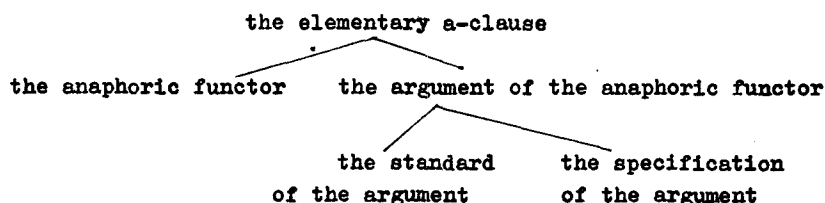
F. Studnicki, B. Polanowska, E. Stabrawa, J. M. Fall,
A. Zachwa

Institute of Computer Science, Jagellonian University
Cracow, Poland

1. The aim of the project is to work out a method of resolving automatically the anaphoric clauses of a certain class, in particular those used in formulating interdocumentary cross-references in primary legal texts (statutory texts). By resolving an anaphoric clause of that class we mean the searching out possibly all of its referends. The implementation of the planned method should enable the users of the full text legal data banks to obtain in search operations, apart from the documents satisfying the requirements defined in the usual queries, also such documents to which the former explicitly, or even implicitly refer. The project has been planned as one composed of three parts. A report on the results of part I was presented at the Fifth ICCH Conference in Ann Arbor in May 1981. The present text aims at showing the main outlines of the approach applied in part II. To make some aspects of that part clear, however, certain references must be made to part I. Part III is, as yet, at the stage of preliminary discussions.

2. The general approach applied in the whole of the project is of a semantic kind. It has been assumed, in particular, that at a certain level of generalization all elementary anaphoric clauses of the above class (let us call them the a-clauses) have in spite of the diversity of their types, an

analogous semantic structure, which can be represented by the following diagram:



Consider the following fictitious example of a legal provision in which an elementary a-clause is inherent:
"Art. 56. In cases when the price is to be paid in cash article 44 of the civil code should be applied."

In art. 56 the a-functor is represented in the surface structure by the phrase "should be applied", the argument by the phrase "article 44", the standard by the phrase "article" and the specification by the phrase "44". The role of the a-functor is confined to signaling the fact that the clause in question has the illocutionary status of an anaphoric utterance, while that of the argument (and its immediate semantic constituents) consists in carrying information relevant to identifying the referends of that clause. Four types of the elementary a-clauses, in particular the type A (the explicitly addressing), the type D (the deictic), the type R (the implicitly referring) and the type S (the semantic) are distinguished. The distinction corresponds to four types of indication met in the clauses in question. By indication we mean the way in which referends of the a-clause are referred to by it.

3. The operation of automated resolving of an a-clause can be conceived of as composed of four stages. Stage 1 consists in identifying an a-clause of this kind within a definite document (article, paragraph,...), mostly by recognizing

the phrase representing its a-functor. At the stage 2 certain of the semantic properties of the analyzed a-clause, relevant for the selection of the most appropriate search procedures, are identified by the program. This stage results in the generation of a formula which is a generalized semantic representation of the actually analyzed a-clause. Such a formula (the SR-formula) is built in a specific language of semantic representation (the SR-language), which is a language with a drastically reduced vocabulary and a very simple syntax outlined in part I. At the stage 2 only the semantic properties accounted for by the specific frame-like interpretation scheme IS are taken into consideration. Stage 3 consists in utilizing the SR-formulas, generated at the stage 2, in automated selecting the search procedures to be employed at the stage 4. The selection is made from among a set of such procedures supplied by the corresponding program. At the stage 4 the selected procedures are employed in the process of searching out the referends of the actually analyzed anaphoric clause (i.e. the documents to which it explicitly or implicitly refers).

4. The simplest version of IS, to be used in interpreting the elementary a-clauses, (i.e. the a-clauses in which only a single indication is inherent), can be conceived of as an ordered pair $[T ; R]$, where T stands for a data structure (called "the ladder") composed of 8 subsequent fields (terminals) and R for a set of rules by which the operation of filling out of definite terminals is governed. According to the rules R two terminals located to the furthest left, are destined to carry information on the type of indication inherent in the actually analyzed elementary a-clause. The remaining terminals (3-8) account each for a definite semantic property of such a clause.

5. By composed a-clauses are meant those in which more than a single indication are inherent. Such clauses are

semantically represented by the composed formulas of the SR-language, in particular by a number of filled out "ladders", connected by the use of certain connectives of the classical calculi.

6. The empirical investigations which form the subject of part II were carried out on a representative sample of the Polish statutory texts of the years 1944 - 1979. The research aimed at reconstructing all possible ways in which the semantic properties of all kinds of a-clauses may be represented in the original texts. Such a reconstruction was indispensable to building the algorithms of transforming the "natural" a-clauses into the corresponding SR-formulas, as well as to the building the possibly ~~most~~ effective procedures of searching out the referends of the analyzed a-clauses. The research resulted in forming: a) the lists of words which occur in the phrases representing in the surface structure the corresponding semantic constituents of the a-clauses of all types, b) the lists of words which occur in the phrases reflecting definite semantic properties of such clauses, and c) lists of grammars reconstructing the empirically observed syntactic connections between those words. Such a way of presenting the results of the empirical investigations inherent in part II seems most suitable for constructing the aforementioned algorithms and procedures.

7. Part III of the project is concerned with the ways of implementing of the planned method. Only a few of the corresponding algorithms and procedures have already been worked out by the authors.