Franciszek Studnicki; Beata Polanowska; Ewa Stabrawa; Jarosław Mścisław Fall; Andrzej Łachwa[1]

AUTOMATED RESOLUTION OF REFERENCES OCCURRING IN LEGAL TEXTS

Originally published as "Zautomatyzowane rozwiązywanie odesłań występujących w tekstach prawnych," Studia Semiotyczne 13 (1983), 65–90. Translated by Lesław Kawalec.

1.INTRODUCTION

1.1. The Jagiellonian University ICT Faculty are involved in a research project called *ANAFORA*, whose main aim is to create a method of the automated resolution of anaphoric phrases — expressions used in formulating references occurring in the so-called primary legislation, that is Acts of Parliament, decrees, executive orders, etc. By the resolution of such expressions we mean operations that are about the identification of the cases of an expression occurring in a certain part of the text (divided into the so-called documents) and possibly finding all the referents of these expressions, that is, all documents containing information about the expression references.[2]

1.2. Work on the first part of the project, leading to the reconstruction of those semantic properties of the expressions whose knowledge is in some way important for the operation, was completed in 1981 and this is the part

[1]The people listed as co-authors of the paper make up a research team led by F. Studnicki. The names have been mentioned in order of joining the team (in 1978-1979).

[2]The implementation of the method described herein will consist in its application in an ICT system designed for document search because the smallest portion of the information that such a system can provide is one consisting of the whole document (the whole paper of the whole section of the primary legal act), it is convenient and, on account of the aforementioned property of the system, absolutely sufficient to treat whole documents rather than phrases they contain as the referents of the anaphoric expression (that is the parts of the text to which the expression refers).

that will be described in this study. The research team is now working on the second part, whose aim is to confront the results obtained when working on the first part with the empirical material — the corpus of Polish primary legislation published between 1944 and 1979, represented by a sample of 200 such acts. The material also includes the regulations covered by the six codifications performed in that period. The third part of the project, aiming at the implementation of the method, that is leading the operations that add up to make it to a condition where the operation is performed by a digital machine, is just a preliminary stage of preparation. Only some of the programs of the operations have been done by now.

2. THE SEMANTICS OF THE ANAPHORIC EXPRESSIONS

2.1 The subject matter of the project is not only the resolution of the anaphoric expressions used in the formulation of the references in which the addresses of the referents are given explicitly (numerical references) but also the resolution of the references where the referents are indicated only by reference to some specific semantic properties (semantic references). The taxonomy of the anaphoric expressions we have adopted also identifies the so-called deictic expressions, where the referents are not indicated by supplying their absolute addresses, that is the numbers that match them with the original legal texts but by reference to the position they occupy in the texts in relation to the position of the document that contains the anaphoric expression. Another category are the so-called *associative anaphors*. This refers to cases where a document refers to (a) document(s) that only precede(s) it in a implicit manner, for example by way of using some marked and characteristically positioned phrases, such as 'however', 'regardless', 'apart from', 'irrespective of', etc. The semantic role of such phrases is about making the reader sensitive to the fact that the contents of the document where the expression appears are to be contrasted to the contents of the document(s) that precede it in the text or that is related to the contents of such documents with some other particularly strong semantic connections.

2.2 It must be stressed that not all the expressions that provide a number or other markings that correspond to some textual units in the original texts can be regarded as anaphoric expressions of the kind we are interested in. The documents that contain such expressions are thought to be only those that cannot be otherwise interpreted without the knowledge of the contents of those text units the expressions refer to. Therefore the expressions that introduce or repeal some legal norms or expressions indicating the

regulations that form the legal basis for the enforcement of other norms will not be considered anaphoric expressions of the kind.

2.3 The taxonomy of anaphoric expressions used in the formulation of references in primary legislation texts suits the kinds of indications that typify the expressions. By indication we mean the way the referents of an anaphoric expression act, a way which characterizes the expression. We distinguish between 4 kinds of anaphoric expressions: A (expressions that explicitly specify the addresses of referents), D (deictic anaphoric expressions), N (associative anaphors) as well as S type (semantic, that is, indicating their referents by calling upon the content substance).

2.4 In analyzing all types of anaphoric expressions, one needs to distinguish between the properties inherent in the surface structure and their semantic properties. When it comes to the semantic structure of the anaphoric expressions, we assume that at a certain level of generalization it looks analogical for all these expressions. Differences surface only in an investigation conducted at the lower level of generalization. At such a level each of the expressions, two direct semantic components can be identified: (1) the anaphoric functor and (2) the argument of this functor. By 'anaphoric functor' we mean a semantic component limited to revealing the illocutionary status of the expression in which it is contained, and, in particular, that it indicates the expression being anaphoric, that is, an expression referring to some information included in textual passages that it more or less clearly indicates. 'The argument of the anaphoric functor' ought to be construed as a component which carries information that is needed to identify the referents of the anaphoric expression being investigated.

Within the argument of the anaphoric functor, two direct semantic components ought to be identified: (1) its standard and (2) its specification. The role of standard is about it bringing some information on the kind of textual units the anaphoric expression references. Because what is of interest is only the information concerning the kinds of units, and therefore this information can appear in the standards of various anaphoric expressions. In such cases it can happen , and indeed it does happen, that the phrases that represent the standards of all these expressions in the surface structure become formally identical. The role of detecting specific properties of the referents, that is, indicating those of their properties that distinguish them from among all the units of the text equipped in the generic properties indicated by the standard, is performed by the second direct semantic component of the argument — its specification. The semantic structure of any elementary anaphoric expression (an anaphoric expression with just one indication) is explained by Fig. 1.

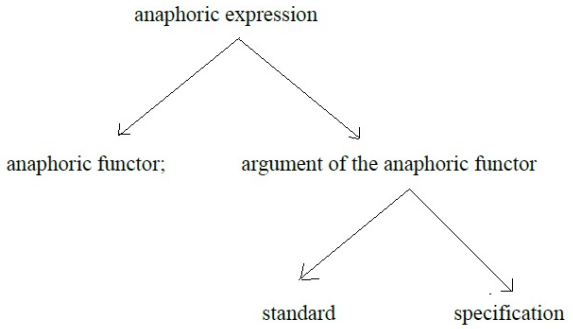

Fig. 1

2.5. The semantic structure of the anaphoric expression can be represented more or less completely in its surface structure. An incomplete representation may occur if some semantic components of such an expression are "nullified," that is, have no counterparts in this structure or when two or more of such components are jointly represented in the surface structure by one phrase that cannot be subdivided into two parts, each of which are representing one component. It often happens that the whole semantic structure of an anaphoric expression is represented in its surface structure by just one phrase, quite commonly made up of just one word. This is the case with the N-type anaphoric expressions (associative anaphors). In the case of incomplete representation, the components that have been "nullified" or that have shared representation with other components must be reconstructed by the addressee in the comprehension process. This can be done by making inferences based on information derived from extra-textual sources, such as the general or specialist knowledge on the part of the addressee. When the addressee is a machine, the inferences are made by means of some corresponding interpretative schemata, containing elements of general and specialist knowledge.

2.6. The process of comprehending a linguistic unit, such as a text or sentence, can only be treated as a process made up of a number stages that follow one another. If the process runs regularly, then with the end of each of the stages, the sense attributed to the text unit at the receiving end becomes more complete than it was at the previous stage. The process of understanding is usually "open-ended" because the recipient cannot reach a stage where nothing more could be added to make comprehension fuller.

Therefore the recipient must usually give up taking this process beyond a certain stage. It is usually around such a stage of the process where the sense which the recipient attributes to a linguistic unit is complete enough to satisfy the recipient's current need for information. The range of the information predominantly relies on the objective the information is needed to satisfy. Therefore every understanding can be regarded as instrumental. In cases where the information provided by a linguistic unit is necessary for an operation to be performed, we may speak of operational comprehension. The outcome of such comprehension processes can be thought to be satisfactory when it equips the recipient with information that is sufficient to perform an operation.

If the task of comprehending some linguistic units is left to a machine, the way the machine comprehends such units is always operational comprehension.

2.7. The operation of automated resolution of anaphoric expressions of the kind described above can be treated as a procedure composed of the following parts:

1. part 1: the identification of the anaphoric expression of a portion of a legal text (in a specific document, in particular);

2. part 2: the recognition of some semantic properties of the expression identified, leading to the generation of a formula that constitutes its generalized semantic representation;

3. part 3: using the formula that has been built in part 2 to select search procedures to be used in part 4;

4. part 4: the use of the search procedures selected in part 3 in the process of finding the referents of a given anaphoric expression, that is the documents the expression refers to.

Each of the parts covers one or more processes of comprehension. The process that obtains in part 1 is of a very general nature. It is limited to investigating one portion to ascertain whether the portion contains an anaphoric expression. The investigation is performed by way of reviewing the surface structure stratum of the unit of text, and a right document in particular, in search of some characteristic and characteristically positioned phrases, whose role is about making up the surface representation of the anaphoric functor. A positive result of this test launches a series of further

steps, which the subsequent parts of the operation described herein are made up of.

The second part describes a more complex comprehension process where a textual unit, document) identified in part 1 as one containing an anaphoric expression, is subject to two subsequent operations: the first is about identifying such component parts of the surface of a unit of text that can constitute the surface structure of an anaphoric expression. If the text unit (document) investigated contains more than one anaphoric expression, all these expressions must be identified. As we said before, it often happens that some semantic components of two or more anaphoric expressions (anaphoric functors of two or three such expressions)are represented in the surface structure of a language text jointly by one phrase only, sometimes one that consists of just one word. In such cases, all such expressions have in the surface structure of this unit a shared part, that is, one that belongs to each of those.

In the other operation, which part 2 consists of, each of the anaphoric expressions identified before is subject to testing aimed at revealing some generic semantic properties. What we mean is some properties that can be ascertained by means of an interpretative scheme used at this stage of the operation. All these operations are binary, that is, each of them can be either inherent in the anaphoric expression or may not. Using all the information supplied by the surface structure under investigation and by means of the interpretative scheme performing the operation described, the program generates a formula built on a language that we will call the language of semantic representation (JRS). These formulas will be called the formulas of semantic representation (RS formulas). Understandably, the formulas can only reconstruct some semantic properties of the anaphoric expressions investigated. These will be the semantic properties whose recognition is necessary at the right stage of the operation described here. It is easy to notice that the second part of the operation is a procedure leading to the transformation of the anaphoric expressions contained in the original text into their simplified and standardized counterparts in a language that is better suited for automated processing.

In part 3, the subject matter of comprehension are the RS-formulas generated in the second part. However, unlike in part 2, part 3 does not lead to generating linguistic units of some kind but to making a choice. In particular, a search procedure or a sequence of procedures ought to be chosen that would be best suited to searching for referents of the anaphoric expression equipped in generic properties that are reported by the right RS

formula.

There are some comprehension processes in part 4 done with the use of procedures launched by a choice made in part 3. The criteria on which the search in part 4 is conducted are based are much more specific than the ones used for the choice made in part 3. In particular, unlike what happens in part 3 (where the choice made by the program is dependent on rather few generic semantic properties), the search that will be made in part 4 must reckon with a practically unlimited diversity of peculiar semantic properties, postulated for referents by suitable anaphoric expressions.

2.8. The types of anaphoric expressions — A, D, N and S — will now be illustrated by means of some (fictitious) examples of legal acts that contain the anaphoric expressions.

Type A — anaphoric expressions that explicitly provide the addresses of referents.

Example 1

"#56. if the price should be paid in cash, CLAUSE 44 OF THE CIVIL CODE SHALL APPLY."

Example 2

"#15. If the perpetrator is under 16 years of age, the punishment PRESCRIBED IN CLAUSE 147 OF THE PENAL CODE shall be reduced by half."

Type D — deictic anaphoric expressions

Example 3

"#15. If the perpetrator is under 16 years of age, the punishment PRESCRIBED BY THE PREVIOUS CLAUSE shall be reduced by half."

Type N — associative anaphors

Example 4

"#15. HOWEVER, if the perpetrator of the crime is under 16, the punishment ought to be reduced by half."

Type S — semantic anaphoric expressions

Example 5

"#86. If no tariffs are in force, THE REGULATIONS OF THE CIVIL CODE CONCERNING RETAIL SHALL APPLY."

In each of the examples above, the sequences of words create a surface structure of an anaphoric expression. The semantic components of the anaphoric expressions occurring in examples 1-5 are represented in their surface structure in ways that are explained in the following table:

In example no.	phrase representing the anaphoric functor	Argument represented by phrase	Standard represented by phrase	Specification represented by phrase
1	Shall apply	Clause 44 of the civil code	#	44 of the civil code
2	Prescribed in	Clause 47 of the penal code	#	147 of the penal code
3	Prescribed by	Preceding clause	clause	preceding
4	however	n/a	n/a	n/a
5	Shall apply	The regulations of the civil code concerning retail	regulations	of the civil code concerning retail

Fig. 2

Concerning anaphoric expressions, indicated in examples 1 and 2, the problem of surface structure representation of their semantic components is clear. In particular, each of the components has its own sufficient representation in the surface structure of the expression. Therefore the referents of the expressions can be identified by the sole use of the information contained in this expression, and thus without the information coming from other sources.

In the anaphoric expression contained in example 3, the semantic component which we have called specification is represented by the phrase "preceding," but it is a deictic phrase, which makes a specific sense in a specific deictic system. In the case under consideration, such a system is created (which does not always obtain) out of elements of a linguistic nature only. One can speak of such a system particularly because the D-type anaphoric expression is included in a linguistic unit called a clause, which is part of a linguistic unit of a higher kind — the text of a legislative act — which is, as far as its surface structure is concerned, a linear collection of clauses and thus one on which the relations of 'precedence' and 'succession' are well defined. Therefore, if we assume that the clause indicated in example 3 is an element of this collection and is not its first element (which follows from the number that has been attributed to it), we can assume that the phrase 'preceding' represents the specification of the anaphoric expression contained in the regulation in a way that is sufficient for the identification of the (only) referent of the expression.

When it comes to example 4, the issue of the surface representation of the semantic components of the anaphoric expression it contains is a little more complex. In particular, the presence of the anaphoric expression in #15 is signaled in its surface structure only by one (one-word) phrase 'however', placed at the beginning of the clause. The role of such phrases has been presented above in 2.1. In Fig. 2 the phrase 'however' was classified as the

surface representation of the anaphoric expression included in example 4. One could claim, however, that the phrase represents, in the surface stratum of example 4, not only the analytical functor of the expression but the whole expression, too.

Such a claim would be only partly justified, though, as what we learn straight from the phrase is limited to the information that example 4 includes an N-type anaphoric expression. No other information that might be used in the identification of the referents of the expression (information that is usually carried by phrases representing the other semantic components of an anaphoric expression) is not included in the surface structure of example 4. The information that is missing can only be retrieved by way of using reconstruction mechanisms, particularly those whose functioning is about getting information from some external sources (cf. 2.5). Hence it can be assumed that the role of the phrase "however," used in example 4 resembles ones that in other types of anaphoric expressions are played by some characteristic phrases that represent in the surface structure a semantic component that we have called above the anaphoric functor.

It must be emphasized that the indication in example 4 is not as unambiguous as the ones that occur in 2 and 3. In particular:

a) It is by no means clear whether in example 4 there are one or more referents indicated. Let us assume that the whole information included in example 4 is known to us but that the only thing we know about its context is that the example is part of a text made up of clauses arranged in a linear fashion along with the numbers attributed to them and that it is not the first of the articles that belong to the text. As things are, it cannot be ruled out that the part played by the phrase "however" used in clause 15 is about contrasting its content substance not only with the contents of one of the clauses preceding #15 of the text, but the contents of two or more such clauses, each of which prescribe a different punishment for a different crime. Therefore, it cannot be ruled out that the anaphoric expression included in example 4 does not indicate just one clause but more.

b) Whereas there can be little doubt that the anaphoric expression included in example 4 (in clause 15) draws upon some information included in a clause or some clauses placed in the text preceding clause 15, it is by no means clear whether the expression only draws upon the information included in the clause that directly precedes clause 15 in this text or, perhaps solely, the information from some more (but not too) remote clauses.

The particular way of indicating referents that is characteristic for N-type anaphoric expressions means that the kind of indication used there can

be treated as something in-between the indication that more or less clearly shows the place where the referents are located in the text (the indication found in the A and D-type anaphoric expressions) and the indication that is about making a reference to some semantic properties of the referents (occurring in the S-type anaphoric expressions).

As regards the (S-type) anaphoric expression included in example 5, there is no doubt that all the semantic components of the expression are sufficiently represented in its surface structure. The way the expressions of this kind indicate referents has been sketched in 2.1.

3. INTERPRETATIVE SCHEME FOR THE ANALYSIS OF ELEMENTARY EXPRESSIONS —

INTRODUCTION

3.1. We have already said that at a certain stage of the operation of the autonomous resolution of the anaphoric expressions discussed here, the task or recognizing some of the semantic properties of such expressions is given to a program equipped in a special frame-like interpretative scheme (IS).[3]

IS contains as its proper part a certain data structure that adopts a more or less complicated form depending on the degree of complexity of the anaphoric expression which is subject to research in a particular case. We are dealing with the simplest form of the anaphoric expression and, in consequence, with the simplest form of IS, where the IS is an elementary anaphoric expression. We use this term to denote anaphoric expressions where only one indication (of any type) appears. All the anaphoric expressions appearing in the examples above have been the examples of this kind.

3.2. The simplest version of IS (hereinafter 'a ladder') will be treated as an ordered pair:

$$[T; R],$$

where T means a sequence of eight numbered fields, hereinafter 'terminals', R being a set of rules of which the operation of filling the terminals (allocating values to them) is governed . In line with the rules belonging to R, the two left-most terminals (1 and 2) form a unit meant to inform us which of the four kinds of indications currently occurs in the anaphoric expression

[3]Schemes that have been introduced into ICT by M. Minsky (1975). The concept of [interpretative] frames was also developed by E. Charniak (1975).

being investigated. Each of the remaining terminals (3-8) informs us about whether there occurs a semantic value (different for each) in the anaphoric expression (only some selected semantic properties whose ascertainment is relevant for the right course of the third part of the operation). The rules belonging to the set R will be more precisely presented in 4. The arrangement of terminals making up the data structure is presented in Fig. 3 (cf. 4.1).

Each of the terminals can alternatively take the values of 1 or 0. Regarding the terminals 3-8, this results in the corresponding semantic values being binary. The filled-in ladder forms the elementary formula of the language of semantic representation (2.7.). This formula is a simplified semantic representation of the anaphoric expression. The role of such formulas in the operation of the automated resolution of anaphoric expressions (cf. Ibid.). The rules governing the filling of the terminals making up the ladder are identical with the rules of creating the elementary formulas from the language of semantic representation (JRS).

It is self-evident that matching a given anaphoric expression with a JRS formula is tantamount to translating this expression into the language. On account of the special character of the JRS, only some semantic properties of the anaphoric expression thus translated "survive" the translation operation. They are those semantic properties in particular that the program reconstructs using the IS. As can be seen, the role of the scheme is above all about limiting the actions performed at a specific stage of the operation to those that reconstruct within the JRS (a language better suited for automated operations than a natural language) some selected semantic properties of the expressions. This is especially so concerning those semantic properties that have some significance for the appropriate selection of the referent finding procedures (cf. 2.7.). At the same time, however, the interpretative scheme described before is designed in such a way as to make the constructed JRS formulas be used for reconstructing all such properties.

3.3. JRS is an artificial language and has a very simple syntax, which will be described in more detail in 6. The semantics of JRS is a simplified equivalent of the syntax of a language in which the anaphoric expressions are built, and thus a simplified equivalent of a passage of the language used in primary legislation texts.

JRS was built in such a way as to be capable of serving as the language in which expressions will be built that will be an outcome of the process of translation described before. In these processes, the only sources of information that can be used by the translation program include the surface structure of anaphoric expression currently in translation and

the IS interpretative scheme. Trusting that the processes we are talking about will lead to a satisfactory reconstruction of the semantic properties of the expressions in question is based upon an assumption that is called correspondence presupposition. The assumption has it that the connections between the formal properties and semantic properties of the anaphoric expressions are strong enough to make a program that uses an IS capable of making the inferences on some semantic properties of the expressions to be investigated on the basis of their formal properties.

3.4. The IS has been classified as a frame-like scheme. It needs to be stressed, though, that its functioning differs in some ways from the way standard interpretative schemes of its kind operate. In particular, the functioning of such schemes is first and foremost about making programs capable of making the right inferences from the information supplied by a given information scheme concerning the environment and thus organizing their knowledge of the external world. Unlike this, the information that the program can get by using the IS does not concern the external world (i.e. the environment using this systemic scheme), but relates to some properties of the data stored by the system. As can be seen, the functioning of the IS is about providing assistance to the system that uses it in organizing the information concerning the contents of its memory.

1. USING THE "LADDER" TO RECONSTRUCT THE SEMANTIC PROPERTIES OF ANAPHORIC EXPRESSIONS

4.1. Introduction

The data structure herein called the "ladder" takes the form of eight consecutive fields (terminals). The first left-most two fields are meant to reconstruct information concerning the type of indication that occurs in the anaphoric expression being investigated. The fields make up a unit we will call an indicational area. As regards the next six fields (terminals), each of them is meant to contain information about the occurrence or non-occurrence in the anaphoric expression of some marked semantic property. The fields (terminals) are ordered in a way that is reconstructed in Figure 3.

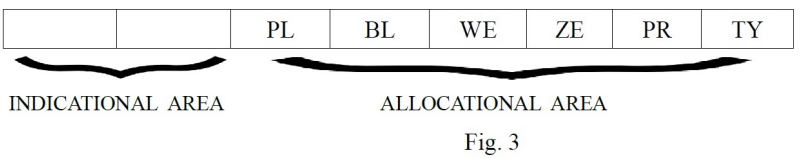

Fig. 3

Each of the markings — PL, BL, WE, ZE, PR, TY — is allocated to one semantic property, which is to be reported by the second area of the ladder to be called an 'allocational area'.

The semantic properties having the respective markings will be described below (4.8-4.13).

The analysis of the properties this paper describes is an abbreviated version of the analysis we made in the original account we made to report the first part of the ANAFORA research project. On account of the lack of space, we have to skip many of the details that have some importance for the implementation of the method we propose.

4.2. The role of the JRS formulas in the operation of the automated resolution of the anaphoric expressions in question has been presented above. The instrumental nature of this role means that the set of semantic properties reconstructed by these formulas can be treated as a collection of preliminary requirements imposed by the anaphoric expressions on the documents — candidates for the status of the referents of these expressions.

4.3 Indicational area (terminals 1 and 2)

The values taken by the terminals making up the indicational area are dependent upon the indicational type occurring in the currently instigated anaphoric expression. The types A, D, N and S are ascribed the values 11, 10, 01, 00 respectively.

The differences between the types have been outlined in 2.1. Some details will now be provided, but due to the limitations of space, they will not be fully developed.

4.4. Semantic value A

The property is about the anaphoric expression investigated indicating its referents by providing their internal addresses. These are the numeric or alphanumeric denotations allocated to these documents or multi-document blocs which also include the documents in the original text.

The addresses can appear in various forms. The differences between the forms will be about phrases that indicate the same addresses in the given anaphoric expressions possibly including different lexical units alternatively, various grammar words and differing configurations of words.

It often happens that the address that appears in the anaphoric expression is incomplete and thus ambiguous. We deal with such cases when a phrase may give the number of the clause and the section number but fails to make an explicit mention of the normative act that includes the text units. The disambiguation of such incomplete addresses must in such cases be made by special sub-procedures involved in the relevant procedures of

searching for referents.

4.5. Semantic property D

The outline of this property is given in in 2.1. D-type indications may occur in two variants. The first one of those will be called "direct deictic indication." It occurs when the anaphoric expressions refer to the documents that directly precede the document where the anaphoric expression is found or the document(s) that immediately follow in the text. All the examples of D-type anaphoric expressions presented so far were about this particular version.

The second version, called "indirect deictic indication" occurs when a D-type indication points to the referents of the anaphoric expression by appealing to their being included in a bigger text unit JT^i (such as a chapter) that precedes the text unit JT^iwhich the document containing the expression forms part of, or by appealing to the fact that these are included in a larger text unit JT^k that appears in the text right after the unit JT^j. Here is an example of a document where an anaphoric expression of the kind occurs:

Example 6

"#88. In the case of sale by auction, THE RULES INCLUDED IN THE PREVIOUS CHAPTER SHALL APPLY."

The main source of the difficulties arising in the resolution of D-type anaphoric expressions is that the indications they contain tend to be ambiguous. This is particularly true of cases where indications point to more than one element (such as referring to the preceding and following clauses without specifying the number involved, rather than the preceding of following clause). The difficulties increase when D-type anaphoric expressions do not refer to some standardized text units such as clauses or sections but to some creations that cannot be identified without the application of some semantic criteria. This occurs when the anaphoric expressions refer to the 'following principles', or 'rules contained' in a text unit (a chapter), etc. Such anaphoric expressions can only be resolved with the use of special procedures whose activity is about phrase disambiguation. Such procedures must form part of the procedures of searching for referents.

It is obvious that the main reason for the difficulties occurring in the resolution of D-type anaphoric expressions is that the phrases that represent the arguments of such expressions in the surface structure are often affected by ambiguity.

4.6. Semantic property N

We have noted the ambiguity of N-type anaphoric expressions (2.8.). It ought to be stressed that, unlike the anaphoric expressions of other kinds,

N-type anaphoric expressions are always ambiguous, not just sometimes. This is caused by the fact that the representations of the semantic components of such expressions in the surface structure are reduced to a minimum.

4.7. Semantic property S

As we said before, the property is that, in the expressions that have these, the referents are not indicated by making a reference to some specific location in the text but by referring to some specific semantic properties. The example of the S-type anaphoric expression has been presented above (2.8.).

4.8. The allocation area of the "ladder"

The second area of the ladder — the spaces from 3 to 8 — has been called the allocational area. Making use of this name is justified by the fact that all semantic properties reported by the terminals that make up the area concern the allocation of referents, that is, their location in the legal texts.

4.9. PL semantic property

This is about an anaphoric expression pointing to more than one referent. The anaphoric expression can be equipped by a third property in a number of ways. Concerning A-type anaphoric expressions, their plurality can be achieved by giving two or more referent addresses or by using a collective address, that is, an address subordinated to a certain multi-element bloc of documents (rather than one document) such as a chapter. Also, concerning D-type anaphoric expressions, plurality can be attained by using the generic name of the corresponding text unit (clause or section) in the plural.

Concerning N-type anaphoric expressions, on account of their confirmed ambiguity, it can never be out of the question that they are equipped in PL property. Therefore, it seems reasonable to treat all expressions of this kind as ones that have this property.

Burdening the program with examining whether the anaphoric expression being analyzed is equipped with a PL property is justified by the fact that when the expression proves not to have it, searching for its referents can be stopped after finding only one document that is a referent of this expression.

4.10 BL semantic property

The term 'bloc of documents' is understood as a non-empty set of documents comprising one or more of such documents, occurring one after another in the original text.

The anaphoric expression is equipped in BL semantic property when it refers to a bloc of documents. A reference to a bloc of documents can occur either when the anaphoric expression exchanges the external addresses

of all the documents that are the elements of this bloc or by the occurring of the so-called collective address in the expression: the address of a text unit (chapter) making up the document bloc. All D-type anaphoric expressions are treated as equipped in BL semantic properties for the following reasons:

a) if the anaphoric expression has no PL semantic properties, that is, when it refers to just one document, then one is equipped in BL property only because a single document is by definition a bloc.

b) if a D-type anaphoric expression has PL properties, such as when it generally refers to the preceding or following clause, then any range of such a reference is questionable, there are no obvious reasons to accept that the documents thus indicated are separated from one another with documents that this indication does not refer to.

The same concerns N-type anaphoric expressions, which we will also treat as equipped in the property.

4.11.WE semantic property

These are expressions that refer to documents that form part of the same legislative act (Act of Parliament) which also includes the document that has the anaphoric expression. It must be stressed that the anaphoric expression can refer to documents that form part of the normative act and the documents contained in other normative acts. In such cases we have to deal with an anaphoric expression equipped in both WE property and the semantic property ZE (see below 4.12.).

It is obvious that all the anaphoric expressions of the types D and N are equipped in the semantic property WE.

4.12. ZE semantic property

This is a property of the anaphoric expressions that refer to documents other than those contained in the normative act whose part is the document that contains the expression. It has already been said that an anaphoric expression can be equipped in both WE and ZE (4.11.).

4.13. PR and TY semantic properties

TY anaphoric expression is one that refers to documents included in the same normative act but ones that directly precede or follow the document where the expression is found. PR is about something to the contrary. The same anaphoric expression can at the same time be equipped in a semantic property PR and TY.

4.14. The JRS formula, created by filling in all the terminals of the ladder with suitable values ought to be treated as a result of the transformation of the anaphoric expression into its equivalent in JRS — language of semantic representation — an outcome of a translation into JRS. As we said

before, on account of particular properties of this language, these formulas reflect only some of the semantic properties of the anaphoric expressions subjected to translation. Therefore, the information which the formulas will carry is just a simplified counterpart of the information contained in the expressions being translated. Despite this, the process of filling the terminals of the ladder is at the same time one of translation and one whose correctness is completely independent from the reality the anaphoric expressions subjected to translation refer to. The correctness of the process is totally dependent on whether the RS formula attained as a result of this correctly reconstructs the semantic properties of the anaphoric expressions in translation, and the ones covered by the IS interpretative scheme in particular. (cf. 3.). Therefore, in cases where the word 'clause' or 'principle' used in the anaphoric expression in the plural, the ladder terminal PL takes the value of 1. The terminal's taking this value ought to be treated as regular even when it has no referents in the corresponding text(s).

For similar reasons, filling the specific ladder terminal with value 0 only means that the anaphoric expression being investigated contains nothing that could indicate that the expression is equipped in the semantic property the terminal reports. Some deviations from these rules (pertaining to N-type anaphoric expressions) have been presented in 2.8. Importantly, JRS is the same-level language as the one in which the anaphoric expressions are built, not a metalanguage. Therefore we assume that RS formulas represent corresponding anaphoric expressions. Rather than the formulas describe the expressions in JRS.

On account of some analytical interdependency obtaining between semantic properties acknowledged in IS, all the reconstructions obtained as a result of the application of this scheme are to a degree redundant. It is obvious that in cases where the indicational area of the ladder takes the value of 01, the terminal TY adopts the value 1, etc.

5. SEMANTICS OF COMPLEX ANAPHORIC EXPRESSIONS

5.1. The distinction we are making between elementary and complex anaphoric expressions is based on semantic criteria. Therefore, the complexity we mean when making these distinctions is a semantic complexity reflected only more or less clearly by the formal properties of such expressions, that is, their surface structure.

We have said that (4.1.) indications that occur in anaphoric expressions of any type may be interpreted as requirements imposed on these expressions by documents that are candidates for being referents. Therefore,

from a pragmatic point of view, each of the requirements can be linked to an anaphoric expression and in particular with one that expresses such a requirement. The requirements expressed by elementary anaphoric expressions will be called elementary requirements; the ones expressed by complex anaphoric expressions — complex requirements.

5.2. That a document satisfies a requirement imposed by the anaphoric expression is not always enough to secure the status of a referent. It often happens that the document acquires the status only when it satisfies some requirements imposed by two or more anaphoric expressions, particularly by two or more elementary or complex anaphoric expressions interrelated with each other with special relationships and jointly making up an anaphoric expression (see below 5.5. — 5.9.).

Saying that an anaphoric expression is a complex anaphoric expression is tantamount to saying that the expression imposes on the documents in question more than one (elementary or complex) requirement. However, such a statement fails to provide information on what relationships obtain, in this case, between these requirements.

5.3. The term "c-component" of a complex anaphoric expression means an elementary or complex anaphoric expression which is its part. A 'direct c-component' of a complex anaphoric expression is such c-component which is not a c-component of any c-component of such an expression. All other components of complex anaphoric expressions will be called their indirect c-components.

The 'first order anaphoric expression' is such an elementary or complex anaphoric expression which is not a c-component of any other anaphoric expression.

The term 'model' of (elementary or complex) anaphoric expression in a given A file will be construed as a non-empty set of documents belonging to the A file which fulfills the description contained in the argument of such an expression. The term 'semi-model' of an (elementary or complex) anaphoric expression will be any non-empty subset of its model.

When a first-order anaphoric expression has its model in file A, we will call such a model a reference of this model in file A, with the documents belonging to this model called referents of this expression in file A.

5.4. The relationships that can obtain between the requirements imposed by a complex anaphoric expression on documents that are candidates to the status of referents will be presented by means of the following four examples:

Example 7

"#81. if the price should be paid in cash, CLAUSE 44 OF THE CIVIL CODE AND THE REGULATIONS OF THIS CODE CONCERNING THE PAYMENT IN FOREIGN CURRENCIES SHALL APPLY."

Example 8

"#44. If the perpetrator is under 16 years of age, the punishment PRESCRIBED IN CHAPTER 6 SHALL APPLY AS LONG AS THEY CONCERN THE ADMINISTRATION OF PUNISHMENT."

Example 9

"#42. If no tariffs are in force, THE REGULATIONS OF THE PRECEDING CLAUSE AND THE REGULATIONS OF CHAPTER 9 ON RAILWAY TRANSPORT SHALL APPLY."

Example 10

"#41. If no tariffs are in force, THE REGULATIONS OF THE PRECEDING CLAUSE TRANSPORT, EXCEPT THOSE PERTAINING TO ROAD TRANSPORT SHALL APPLY."

5.5. A first-order anaphoric expression in example 7 (say Zza-7) has 2 c-components. Both the components are elementary anaphoric expressions (Eza-7.1, Eza-7.2). Information about the relationship obtaining between the requirements expressed by Eza-7.1 and Eza-7.2 is included in the phrase that occurs in the 'and' phrase occurring in the surface structure of Zza-7 between the phrases representing the arguments of Eza-7.1 and Eza-7.2 on the shaping of the model Zza-7 in (real or hypothetical) file A. In particular, it will turn out that the two requirements are independent of each other in the sense that if Zza-7 has a model in file A, then both the model Eza-7.1 and the model Eza-7.2. are independent sub-models of Zza-7 in the file.

The expression Zza-7is a first order anaphoric expression. Therefore, if this expression has a model in file A, the model is its reference in this file, that is, a collection of all its referents in file A. The expression Zza-7 has no other c-components except the components Eza-7.1 and Eza-7.2. So, if Zza-7 has a reference in file A, the reference is the sum of its sub-models Eza-7.1 and Eza-7.2. Hence, the document that fulfills the requirement expressed by Eza-7.1 (identical with "clause 4 of the Civil Code") has a status of a referent of Zza-7 irrespective of whether it also fulfills the requirement imposed by Eza-7.2 (irrespective of whether it concerns "payments in foreign currencies") and *vice versa.*

Such a relationship between the requirements expressed by the c-components of a complex anaphoric expression, whose particular case is a relation established by the expression Zza-7, that is, a relationship where each of the requirements expressed by the c-components of the complex

anaphoric expression is determined by a specific sub-model of the expression in the file, will be called a relation of independence.

5.6. The complex anaphoric expression in example 8 (Zza-8) is also a first-order anaphoric expression that has two c-components which are elementary anaphoric expressions (say, Eza-8.1 and Eza-8.2). Each of the components expresses a certain requirement imposed by Zza-8 on the documents-candidates to the status of its referents. However, the relation between the two requirements sets it apart from the one in Eza-7. The information on the kind of relation is included in the phrase 'as long as', which occurs in the surface structure of Zza-8 between the phrases representing the arguments of the elementary anaphoric expressions Eza-8.1 and Eza-8.2. The nature of the relationship will become manifest when we take into account the influence exerted by Eza-8.1 and Eza-8.2 on the shaping of the model Zza-8 in the (real or hypothetical) file A. In particular, I will predict that none of the direct c-components of the complex anaphoric expression Zza-8 in question determines on its own any sub model of the expression in file A. Therefore neither the fulfillment of the requirement expressed by Eza-8.1 nor the fulfillment of the requirement expressed by Eza-8.2 equips the corresponding documents in the status of referents of the complex anaphoric expression Zza-8. This status can only be enjoyed by the documents that at the same time fulfill the requirement expressed by Eza8.1 and Eza-8.2 (the documents included in chapter 6 and concerning the administration of punishment).

The relationship between the requirements expressed by the components of the complex anaphoric expression, whose special case is the relationship established by Zza-8, that is, the relation where the model of a complex anaphoric expression in file A is a multiplication of the models of all its components will be called the relation of positive coordination. The notion of negative coordination will be explained below (5.9.).

5.7. The relationship of positive coordination can of course also obtain more than two requirements expressed by the c-components of the anaphoric expression. The requirements bound together by this relationship will be called positively correlated requirements. The positively coordinated requirement systems will be pairs, threes or n-s of the (elementary of complex) requirements interrelated with the relationship of positive coordination; the systems of positively coordinated anaphoric expressions — the pairs, threes of n-s of (elementary or complex) anaphoric expressions expressing such requirements. The systems of positively coordinated anaphoric expressions are, of course, also anaphoric expressions — complex anaphoric expressions. Similarly, the systems of positively coordinated requirements are in them-

selves requirements — complex requirements, to be more specific. Notably, not all complex anaphoric expressions and not all complex requirements are at the same time such systems because the anaphoric expressions that constitute the c-components of a complex anaphoric expression may, as we know, express requirements that are independent of one another in the sense described in 5.5.

Elementary or complex requirements, whose fulfillment in itself guarantees a document a status of referent of a given anaphoric expression, will be called an independent requirement. A requirement expressed by an anaphoric expression of the first order is always such a requirement. A requirement expressed by an anaphoric expression that is a c-component of another (complex) anaphoric expression can either be an independent requirement or a dependent one depending on whether the model of anaphoric expression expressing this requirement in file A is or is not a sub-model of the complex anaphoric expression in this file.

5.8. A first-order anaphoric expression contained in example 9 (say, Zza-9) is different from the complex anaphoric expressions Zza-7 and Zza-8 in that the its second direct c-component (unlike the corresponding components of the expressions Zza-7 and Zza-8) is a complex rather than elementary anaphoric expression which has its own c-components (which are of course indirect c-components of the Zza-9).Therefore the semantic analysis of the expression Zza-9 must be performed subsequently on two planes: at the level of its direct c-components and at the level of its indirect components. It will have the form of a bottom-up analysis, which means that it will start at the lower and finish at the higher level.

The indirect components of a complex anaphoric expression Zza-9 (say, Zza-9.1 and Zza-9.2) are both elementary anaphoric expressions. Together they form a two-element system of anaphoric expressions that are positively coordinated. Therefore, if a complex anaphoric expression, whose direct c-components are these expressions, has a model in this file A, the model is a multiplication of the models of anaphoric expressions Eza 9.2.1 and Eza 9.2.2 in the file. At the same time direct c-components of a complex anaphoric expression Zza-9 (say, Eza-9.1 and Eza-9.2) are independent from each other. Therefore if a complex anaphoric expression Zza-9 has a model in file A, such a model is a sum of the models of anaphoric expressions Eza-9.1 and Eza-9.2 in this file and, in consequence, the sum of the model Eza-9.1 and the set that is a multiplication of the models Eza-9.2.1 and Eza-9.2.2 in the file.

5.9. A complex first-order anaphoric expression in example 10 (say,

Zza-10) is a particular case of an anaphoric expression whose direct components are bound with the relationship of negative coordination. This term is supposed to mean a relationship between two component parts of the complex anaphoric expression Zza with which, if the Zza has a model in (real or hypothetical) file A, the model is a difference between the models of the c-components.

We have said before that the information concerning the nature of the relation obtaining between the c-components of the anaphoric expression is usually contained in some peculiar and peculiarly positioned phrases occurring in the surface structure of this expression. In the expression Zza-7, this information was included in phrase 'and' whereas in Zza-98 in the phrase 'as long as'. These phrases undoubtedly play a role in anaphoric expressions that is analogical to that which in the classic propositional calculus are played by the conjunctions of alternative and conjunction. In the complex anaphoric expression Zza-10 a characteristic phrase 'except those' appears, located in the surface structure of this expression between the phrases representing the arguments of its two direct c-components. The role which this phrase performs in the expression Zza-10 can be treated as analogical to the role which in the formula $p \wedge (\tilde{}q)$ is played by the sequence of the symbols $\wedge\ \tilde{}$. The result of using in Zza-10 the phrase 'except those' is that if Zza-10 has a model in a given file A, the model is a difference between the models of its first and second direct component, that is the difference between the set of all documents contained 'in the preceding chapter (that is in the chapter that precedes the original text of the chapter, whose clause 42 is part of) and the set of all documents 'pertaining to road transport'.

We are dealing with the relationship of negative coordination where the complex anaphoric expression imposes onto the documents that the description contained in the argument of one of its c-components should fulfill a negative requirement. We understand this requirement to be one that excludes from a set of documents one of its non-empty subsets.

Negative requirements can be coordinated in a positive or negative manner. It can happen that a negative requirement is limited by another requirement of the same kind, in particular by the requirement that removes a non-empty subset of documents from the action of the exclusion.

5.10. With the automated resolution of such complex first-order anaphoric expressions, where there are negative requirements, some losses can be incurred due to the documents thus excluded from the references of these expressions possibly — on top of referring to such topics whose exclusion was intended — referring to other topics which, unlike the former,

can be relevant in the light of the current need for information, felt by the users of the system. This is, however, caused by the fact that in the systems whose functioning consists of searching for documents, the smallest portion of the information that can be provided by the system and at the same time the smallest portion of the information that can be excluded from the set of documents attained by way of a given search operation, is the portion constituting the whole document. Concerning positive requirements, the undesirable consequence of this fact is that search precision is diminishing. Such loss, however, is by no means as painful as those that can result from the fact that the documents subject to the aforementioned exclusion refer to topics that are relevant to users. This fact can sometimes be a reason for a substantial decrease in the completeness of a search. Such losses, however, can happen only when the negative requirements are expressed by S-type anaphoric expressions (4.2. and 4.7.).

5.11. The term 'direct c-component' of a complex anaphoric expression is understood to be such a c-component of this kind of expression which is not a c-component of any of its c-components (cf. 5.3.). A question arises of what criteria are authoritative enough to establish whether a given anaphoric expression Za^1 is simply a c-component of another complex anaphoric expression Za^2, which is a direct component part of a complex anaphoric expression Zza, or the expression Za^1 has the status of a direct c-component of the complex anaphoric expression Zza.

Consider the following example:

Example 11:

"#89. When no tariffs are in force, RULES CONCERNING RAIL TRANSPORT THAT ARE INCLUDED IN THE PRECEDING CHAPTER, CLAUSES 16 AND 17 OF THIS CODE AND ALSO THE REGULATIONS PERTAINING TO PAYMENT IN FOREIGN CURRENCIES SHALL APPLY."

Let us call the anaphoric expression included in example 11 Zza-11. Reading the example no. 11 we are inclined to treat the phrase "clauses 16 and 17" as representing the argument of the anaphoric expression that imposes on the documents-candidates to the status of the referents of a complex anaphoric expression Zza-11 an independent requirement, that is, a requirement that guarantees the documents to the status of the referents of the expression. According to this, we are not inclined to treat the phrase as one representing the anaphoric expression that is a c-component of any of the direct c-components of the anaphoric expression Zza-11. Let us, however, ponder the question of why the phrase should be treated like this. We claim

that such treatment is justified by the fact that if Zza-11 has its model in the (real or hypothetical) file A, then the anaphoric expression whose argument is represented by the phrase mentioned independently continues (that is without any contribution from the c-components of the anaphoric expression Zza-11) the sub-model of the expression in the file.

5.12. It has already been said (5.9.) that the information concerning the nature of the relationship that may bind the c-components of the complex anaphoric expression is usually found in some characteristic phrases placed between the phrases that represent in the surface structure the arguments of the c-components. The role performed by such phrases will be compared to the role played in the traditional propositional calculus by some logical constants. This, however, does not always occur. In particular, it may happen that the relationships in question are expressed otherwise. The diversity of linguistics means that what the legislator may use for this purpose depends on the peculiarities of the languages used in the texts. In primary legislation texts written in Polish, the most economical way of expressing the relationships of independence (cf. 5.5.) is placing the phrases that represent the arguments of two or more independent c-components of the complex anaphoric expression directly one after another and separating them with a comma. This is illustrated by the solution presented in example 11. In this example, this was what was done to the phrases: "RULES CONCERNING RAIL TRANSPORT THAT ARE INCLUDED IN THE PRECEDING CHAPTER, CLAUSES 16 AND 17 OF THIS CODE." However, the phrase "the regulations pertaining to payment in foreign currencies" was appended to the preceding phrases by inserting between this one and the preceding phrases the phrase "an also," that is by the application of a method that has been applied before.

6. THE LANGUAGE OF SEMANTIC REPRESENTATION OF ANAPHORIC EXPRESSIONS (JRS)

6.1. A JRS PASSAGE — THE EXCERPT USED IN THE CONSTRUCTION OF FORMULAS representing elementary anaphoric expressions — was presented in 4. It is easy to notice that the construction of the RS formula representing the elementary anaphoric expression is about filling, in a way that is compatible with the IS rules, all the terminals of the data structure herein called the 'ladder'. In JRS, complex anaphoric expressions are represented by expressions we have called complex RS formulas, that is, by such RS formulas that contain as their components other (elementary or complex) RS formulas interconnected by the following symbols: $\vee$, $\wedge$,

and ˜. The semantic role of these symbols is about supplying information concerning the relationships obtaining between the requirements expressed by some c-components of the complex anaphoric expressions. In particular, the relationship of independence (cf. 5.5.) is represented in JRS by the symbol ∨, the relation of positive coordination (cf. 5.7.) by the symbol ∧, and the relation of negative coordination (5.9) by the symbols ∧ and ˜ used in a way transcribed by the syntax of the language. This can be represented using Backus-Naur notation as follows:

<number> : : = 0 | 1

<elementary formula> : : = <number> . . . .<number>

8

<formula> : : = <elementary formula> |
(<formula> ∨ <formula>) |
(<formula> ∧ <formula>) |
(<formula> ∧ (˜<formula>)) |
((˜<formula>) ∧ <formula>)

6.2. These are 11 elementary and complex examples of RS formulas representing specific anaphoric expressions, and in particular the expressions included in the above examples 1 to 11 (cf. 2.8, 4.5, 5.4, and 5.11.):

1. 11010100 (if the document given in example 1 is not a document that forms part of the Civil Code)

2. 11011010 (if the document given in example 2 is a document that forms part of the Penal Code)

3. 10011001

4. 01111001

5. 00100100 (if the document given in example 5 is not a document that forms part of the Civil Code)

6. 10111001

7. (11010100 ∨ 00101000) (if the document given in example 7 is not a document that forms part of the Civil Code)

8. (11111001 ∧ 00101000) (if chapter 6 precedes clause 44)

9. (10011001 ∨ (11111010 ∧ 00101000)) (if clause 42 precedes chapter 9)

10. (10111001 ∧ (˜00101000))

11. ((00100000 ∧ 10111001) ∨ 11111001) ∨ 00100000)

6.3. It needs to be reminded that the role of RS formulas in the operation of the automated resolution of anaphoric expressions is about controlling the selection of the procedures of searching for the referents of the expressions of the kind currently being analyzed. These procedures have been described in detail in the third part of our report, dedicated to implementation issues.

Bibliography

1. Charniak, Eugene (1975) "Organization and Inference in a Frame-like System of Common Sense Knowledge." *TINLAP '75: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, 42-51. Stroudsburg: Association for Computational Linguistics.

2. Minsky, Marvin (1975) "A Framework for Representing Knowledge." In: *The Psychology of Computer Vision*, Patrick Henry Winston (ed.), 211-277. New York: Mc Graw-Hill.