# Structuring knowledge in a graph

Stokman, Frans N.; Vries, Pieter H. de

*Published in:*
EPRINTS-BOOK-TITLE



*Publication date:*
2000

# Chapter 11. Structuring Knowledge in a Graph

Frans N. Stokman and Pieter H. de Vries

## 1. Introduction

The idea that knowledge is a commodity that can be used by machines in various ways forms the entry to our discussion of knowledge based systems. We will limit ourselves to those systems that are in an active dialogue with their human user. As a consequence, the use and structuring of knowledge in autonomous systems, like robots or visual pattern recognizers, will not be discussed. In general, knowledge-based systems cannot be considered as models of human cognition. Rather, they should provide an extension of the user's cognitive abilities.

The emergence of knowledge-based systems showed that within a limited domain of expertise it is possible to construct computer programs that give intelligent advice to professional users. These systems, however, are still used on a rather small scale. Two obstacles seem to prevent a more widespread use. The first is the inconvenience that a knowledge based system is generally constructed in view of one type of task. However, one would like to use the same knowledge base for different purposes, e.g., a knowledge base in the medical domain should not only be applicable to diagnosis but also to other tasks such as therapy selection, explanation of results to the user, instruction to students, and information retrieval for planning new research. The second factor that stands in the way of more frequent use of knowledge-based systems is the scope of their expertise. Typically, the scope is limited to the knowledge of one expert. For more advanced applications, however, this will not be sufficient and the problem arises of under which conditions the representations of different experts can be integrated. In particular, the question arises as to what contribution a "new" expert can make to an already existing body of knowledge.

The research project on procedures and concepts for the construction and analysis of knowledge graphs of the Technology University of Twente and the University of Groningen aims to develop a system for the representation of scientific theories in (at least) medical and social sciences. The structuring of knowledge in a graph can be seen as the construction of a knowledge-based system integrating knowledge from different sources. In this paper we will give a short overview of knowledge-based systems, and subsequently will jus-

tify the representation of empirical scientific knowledge in a graph. We will then outline the procedures for structuring knowledge in a graph, and finally will elaborate the role of textual analysis in the knowledge-acquisition phase.

## 2. Exploitation and Structuring of Knowledge

The way in which knowledge is used may to a certain extent influence the form in which it is stored in a machine. A model specifying the organization of this storage is referred to as a representation of knowledge. The application of a knowledge representation for a specific task is referred to as exploitation. The process of obtaining and adapting such a representation will be called the structuring of knowledge.

A typical example of exploitation of knowledge is decision support in its various forms. An example can be drawn from medicine where we have a task, diagnosis, that is executed on a representation of a generic patient. The task requires an initial specification of a particular patient, i.e., his symptoms. On the basis of the inferences defined for the representation a specific model for that patient is computed. This kind of inference can be generalized to other forms of decision support. For instance, one not only wants to have a correct diagnosis for the abnormal behaviour of an object but also a description of the behaviour when certain courses of action are taken. Apart from decision support, information retrieval and instruction are important forms of exploiting knowledge. We will not discuss all varieties of exploitation here.

Knowledge-based systems are faced with the problem of acquiring their knowledge. The knowledge acquisition comprises both the collection of knowledge elements in a given domain and their integration in a common representation. The knowledge elements as they are collected do not in themselves suggest a valid representation. It is in the process of their integration that a representation emerges. Because of the role of active integration we will use the concept of "structuring of knowledge" rather than the more passive "acquisition of knowledge."

Learning is a typical form of structuring knowledge. Here, the question is how a representation is organized in a way that permits continuous updating and integration of knowledge. Of course this description of learning leaves many psychological aspects untouched. However, it is sufficient to separate exploitation from structuring of knowledge and we will outline some models for machine learning in the next section. Learning can also be seen as a form of structuring knowledge in which representations from different cognitive systems are compared and integrated. We can regard cognitive systems as a

set of experts (either human or artificial) that each have a representation of a common domain. A question that arises in this context is how the knowledge of a common domain is extracted from its carrier, i.e., from its text, and organized into an integrated representation. A second question concerns what effective procedures can give the relations between representations of different individuals. These and other questions will be dealt with in our discussion of knowledge graphs.

## 3. An Overview of Knowledge Representation

When talking of knowledge representation we can distinguish between procedural and declarative knowledge. This distinction was elaborated in philosophy when Ryle (1949) distinguished between "knowing how" and "knowing that." In the context of this discussion we can describe procedural knowledge as a set of prescriptions for actions. Generally, these prescriptions are referred to as situation-action rules. In a given system the set of rules, and therefore its knowledge, is tuned toward the task for which the system is designed. In a declarative representation, knowledge is not given as a set of rules but as a set of assertions about a certain subject. Conclusions can then be drawn from these assertions by inference methods. These inference methods can be based on formal systems like logic (e.g. the principle of resolution; Robinson, 1965), or, as in our case, graph theory. An important property of a declarative representation is that the same knowledge base can be used in different tasks. For each task an inference mechanism is defined that interprets the same set of assertions. Within a declarative approach to knowledge representation we face the problem of distinguishing among various types of relations that are specified in the assertions about a domain of interest. We will refer to conceptual knowledge as a declarative representation. Furthermore, the core of conceptual knowledge consists of explicitly defined types of relations between concepts. Various types can be distinguished, e.g., they can express definitional as well as empirical relations between concepts. For the selection of the types of relations it is difficult to give general criteria. We note here, however, that if the type of relation is left open an important distinction between a procedural and a conceptual representation vanishes. A relation then merely has the function of an association (possibly directed) between two concepts. Such a neutral relation can adequately be formalized by either a production rule, logical implication or associative link in a graph.

In Table 1 forms of knowledge are contrasted with uses of knowledge to form a four-field table of knowledge based activities. The table should not be considered as a classification of systems, as systems exist that unite activities

belonging to different fields in this table. The table only enables a systematic evaluation of systems in this vigorous field of research.

*Table 1.* Classification of knowledge-based activities

| | Form of knowledge | |
|---|---|---|
| Use of knowledge | Procedural knowledge | Conceptual knowledge |
| Exploitation | Decision-support and information retrieval on the basis of chaining of rules | Decision-support on the basis of detection of causality, information retrieval on the basis of definition relations |
| Structuring | Verification of rules<br>Induction of rules | Integration of definitions and causal models |

The procedural knowledge is generally represented as a set of *if-then* rules, such as the following:

*if* heart-attack *then* conclude high blood pressure
*if* x is a bird *then* conclude x has wings and x can sing

The *if* part of a rule contains a description of a situation that can be observed in the data-base of a rule-based system. The *then* part contains an action. In the example the action is a conclusion. Note that the nature of the inference made in a rule cannot be obtained from the rule itself. The first rule in the above example for instance, embodies an inference based on a causal relation: the high blood presssure is the cause of the heart-attack. In the second rule, however, the inference is of a purely definitional nature. The fact that rules are neutral with respect to the type of relation they specify between antecedent and consequent terms brought about the term shallow reasoning for this use of knowledge.

The exploitation of procedural knowledge is accomplished by the chaining of rules. Rules can be chained in a forward or backward direction. The former occurs when a situation provides enough evidence for executing an action. This is the case, for example, in a design-task like configuring a computer with its peripherals ( McDermott, 1982). In backward chaining the execution of the action has to provide evidence for a postulated situation. This kind of reasoning is found in a diagnosis task. Here we postulate a certain situation, i.e., a disease, and carry out an action to obtain evidence for it (Davis,

Buchanan & Shortliffe, 1977). The logical scheme behind both kinds of reasoning is identical.

A relatively simple method of structuring procedural knowledge is the inspection and updating of a chain of rules triggered by a particular problem presented to a rule-based system. The expert system MYCIN has such a complementary system for structuring rules, namely TEIRESIAS (Davis, 1982). In a dialogue the user can see what the consequences are of deleting or inserting a rule in the the knowledge base. A more complicated method of structuring procedural knowledge is the induction of situation-action rules given a set of examples. Winston (1975) describes a program that, on the basis of examples, identifies correctly the geometrical shape of an arch.

An example of the exploitation of conceptual knowledge is the application of inference procedures to semantic networks for the retrieval of information. In semantic networks many types and even sub-types of relations are distinguished (see Brachman, 1983). In every network, however, two types of relations play a central role: the relation indicating class membership and the relation giving a property of a class. An example of the former is the assertion "a canary is a bird," in which the concept of canary is linked to the class of birds. An example of the latter is the linking of the property "wings" to the concept of bird, i.e., the assertion "a bird has wings." These two types of relations are basic for the retrieval of information: when a particular property cannot be directly retrieved from a concept it is inherited from a concept higher in the class hierarchy. Retrieval from new generations of databases will draw heavily on a conceptual representation of knowledge (Riet, 1983). Conceptual networks can also be generalized to express relations between subnetworks, also referred to as partitions (Winograd, 1980).

Another form of exploitation of conceptual knowledge is causal reasoning. Various forms of causality can be detected, e.g., the so-called minimal cause (Vries Robbe, 1978). In the research on knowledge representation there is a tendency to refine the notion of causality. In Kuipers (1984) a distinction is made between the functional dependencies between quantities and the causal dependencies between the state-changes manifested by these quantities. For a further discussion of the refinement of the notion of causality in knowledge based systems the reader is referred to Vries Robbé and Vries (1985). The term deep reasoning refers to the inferences that are performed on a representation making explicit the relations giving the structure and function of a mechanism.

An example of a program structuring conceptual knowledge is EURISKO (Lenat, 1983). It contains a large network of concepts connected by various types of links such as generalization and specification links and so-called suggestion links. The program has several procedures of a heuristic nature

that explore the network. It has been shown that these procedures can integrate existing definitions of concepts to form new meaningful concepts. For instance, the concept of length was "discovered," starting from the elementary concepts of equality, list, and set. The comparison of representations of different individuals and their integration is another form of structuring conceptual knowledge. These topics will be treated in the discussion on knowledge graphs.

## 4. Representation of Scientific Knowledge in a Graph

As was stated in the introduction, the research group on procedures and concepts for the construction and analysis of knowledge graphs of the Technology University of Twente and the University of Groningen aims to develop a system for the representation of scientific theories in (at least) medical and social sciences. Theories in these sciences are empirically oriented, rather than deductive systems built upon a small number of premises as in, for example, physics. Scientific knowledge in the former sciences is oriented towards explanation and prediction of empirical phenomena by means of theories, in which covariations between classes of phenomena are ordered in a logically consistent and coherent system. The building stones of these theories are concepts of which at least some should be related to empirical phenomena. In order to test a scientific theory (in experimental designs as well as in non-experimental settings) two submodels can be distinguished: a *measurement model* that specifies the relation between manifest (experimental) behaviour and latent (theoretical) concepts; and a *structural* model that specifies the direction and type of association between the different concepts and that, as a consequence, specifies the structure of the phenomena to which these concepts refer.

As such the scientific process can be seen as a process in which, for a class of objects, relations between these objects and their properties are specified and estimated by means of structural models. These properties are related to behavior manifested by measurement models. The estimation of parameters is restricted to the set of variables that is considered in a structural model. Consequently, representation of these parameters with the object of using them in inferences in a knowledge-based system integrating several models would be misleading. Therefore, only the presence or absence of structural relations (possibly extended with a measure of their likelihood of existence) will be used in the designed inferential procedures. The user, however, is given the opportunity to specify all kinds of information for the relations he thinks useful.

Let us start the discussion on the chosen knowledge representation with the elements that constitute the basic parts of any measurement process. Meas-

urement can be seen as the process in which properties are instantiated for a given object (Pfanzagl, 1968; Krantz, Luce, Suppes & Tversky, 1971). Such an instantiation is denoted a value. *Objects*, *properties* and *values* are therefore essential building stones of empirically oriented scientific theories, and therefore of systems that represent scientific knowledge. Symbols and concepts (including subsets of natural language, logic, and mathematics) are the linguistic elements to denote these building stones. Objects, properties and values are related to concepts by a realization or projection of the linguistic concept or symbol into an empirical, "real-world," system. Objects, properties and values must have been defined on a linguistic level to enable one to speak about empirical elements. Within the empirical system a property is the result of the identification of an aspect of an object, and a value is produced by the measurement of a property.

Structural dependencies between properties of an object (i.e., the structural models) can be considered the core of scientific knowledge because they refer to relations between empirical phenomena that are corroborated in empirical research and are assumed to be more generally valid. Such structural dependencies can be represented by arcs between vertices in which the arc denotes a structural dependence relation (denoted CAU) and the vertices denote properties. Such a CAU relation is directed from "cause" to "effect." It should have at least a sign to denote whether the relation between the properties is positive of negative. Taking this perspective on causality, we follow Simon's definition of causality as "an asymmetrical relation among certain variables, or subsets of variables, in a self-contained structure" (Simon, 1977). This definition corresponds to the intuitive use of causality in scientific discussions, in contrast to definitions in terms of logical implications which have the counterintuitive implication that *A causes B* implies *not B causes not A*.

According to the definition of measurement, properties should be related to a generic object. To represent this, the relation *is part of* (denoted PAR) is introduced and represented in the knowledge base by an arc from property to object. A third type of relation - indispensable for purposes of integration and structuring of scientific knowledge from different sources - is used in the knowledge base to represent the relation *is a kind of* between two concepts (denoted AKO). This type of relation is introduced in order to represent that structural relations between properties of a certain class of objects can be considered a special case of those of another class of objects.

Values are not represented as vertices, but as information associated with vertices in their role as properties. In the current representation only presence and absence can be attached to a property if it is a dichotomy, or positive/neutral/negative when the property is a continuum. These values are not relevant in the structuring process, but only in the exploitation phase. In decision support for example, values can be assigned to certain properties by

the user. The consequences of this assignment are computed for other properties of an instantiation of an object on the basis of the signs of the CAU relations.

The above three relations, CAU, PAR, and AKO, are the three fundamental relations in the knowledge base, CAU being defined between properties, AKO between objects, and PAR relating properties to an object (see Figure 1). The three relations are represented by arcs simply because they are asymmetric relations. The direction of an arc does not prevent its use in the opposite direction in searching and inferential procedures such as path algebras. In the opposite direction, arcs represent respectively the relations *is caused by* (denoted CBY), *has as part* (denoted HAP), and *has as kind* (denoted HAK). The restriction of the system to the above mentioned types of arcs is not fundamental and the number of types of arc can be gradually extended. In the present stage of the project the restriction is warranted because of the very fundamental problems that are to be solved precisely in the field of integrating and structuring pieces of information and knowledge from different sources. This area is often avoided in artificial intelligence by restricting the knowledge represented in a system to that of one expert or source. Whether a vertex takes the role of object, property, or both, in the knowledge base is determined by its relations with other vertices. If a vertex is involved in AKO relations or is the head of a PAR relation, it takes the role of object; otherwise it takes the role of property.

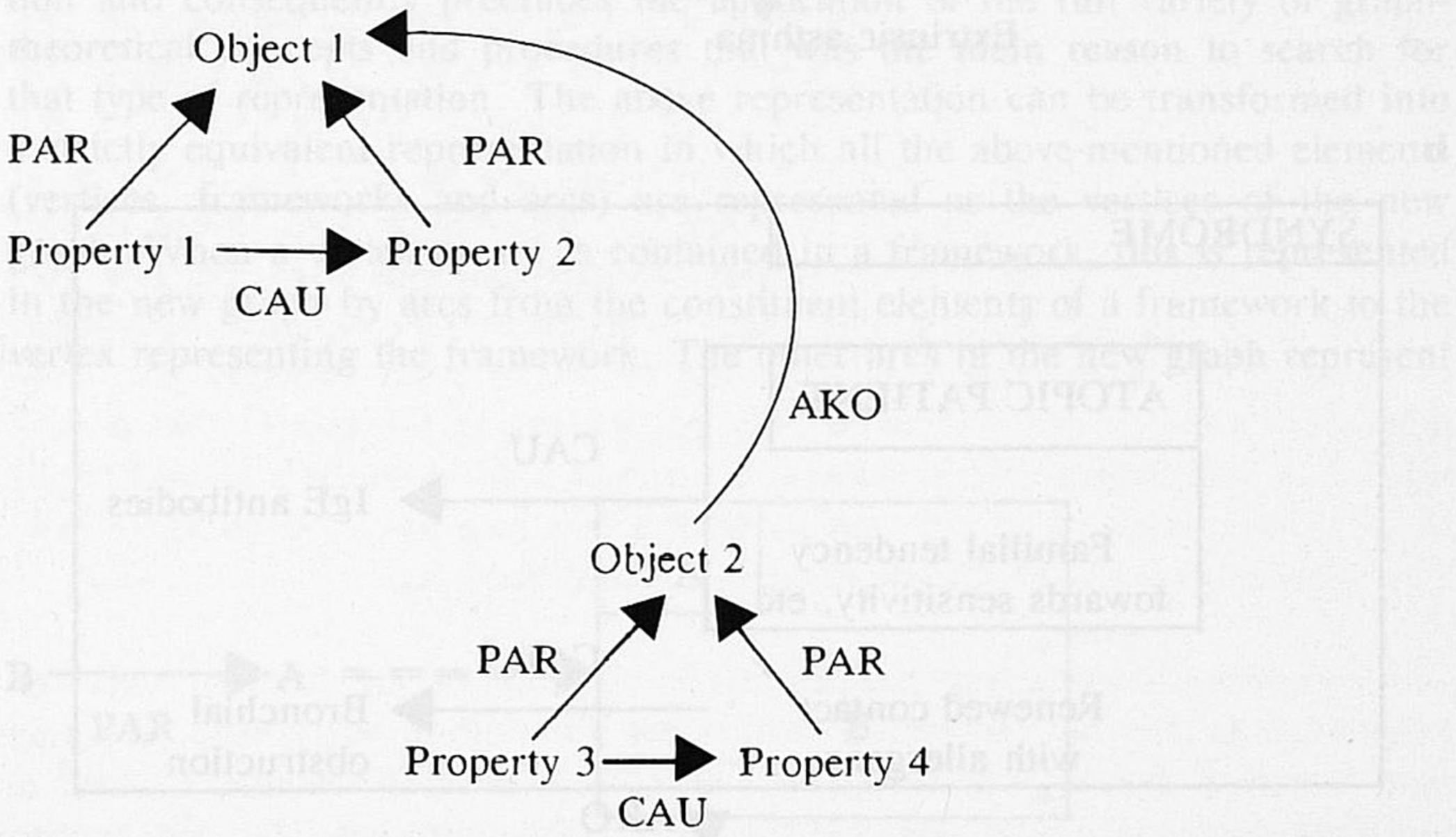



*Figure 1*. Types of relations in the knowledge base

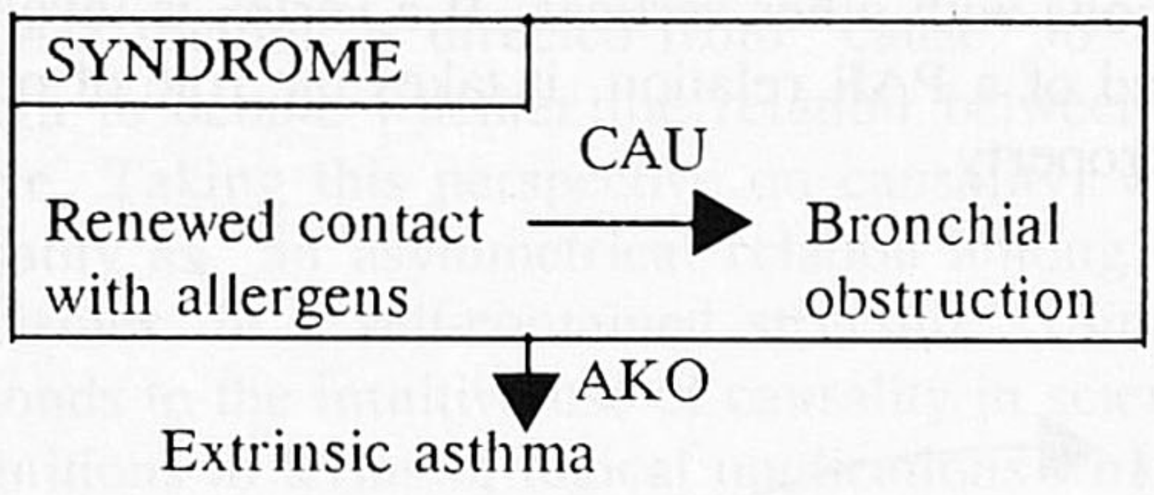
The above representation, however, is not yet sophisticated enough to represent scientific knowledge. Quite often concepts are introduced to denote relations between concepts and these are subsequently used as part of a relation with other concepts. An example is contained in the following sentence:

> The syndrome, in which renewed contact with allergens leads to bronchial obstruction, is an example of extrinsic asthma (Example a)

Such a relation on a relation might have been represented by an arc on an arc, but this was rejected because it can be considered as a special case of a more general phenomenon in which concepts are introduced to denote a whole process consisting of a set of relations and concepts. This can be illustrated with the following example:

> Atopic patients have a familial tendency towards a sensitivity for known inhalation allergens (like dust, molds and pollens). Hereby IgE antibodies are created. The renewed contact with the allergen leads to

**a**




**b**

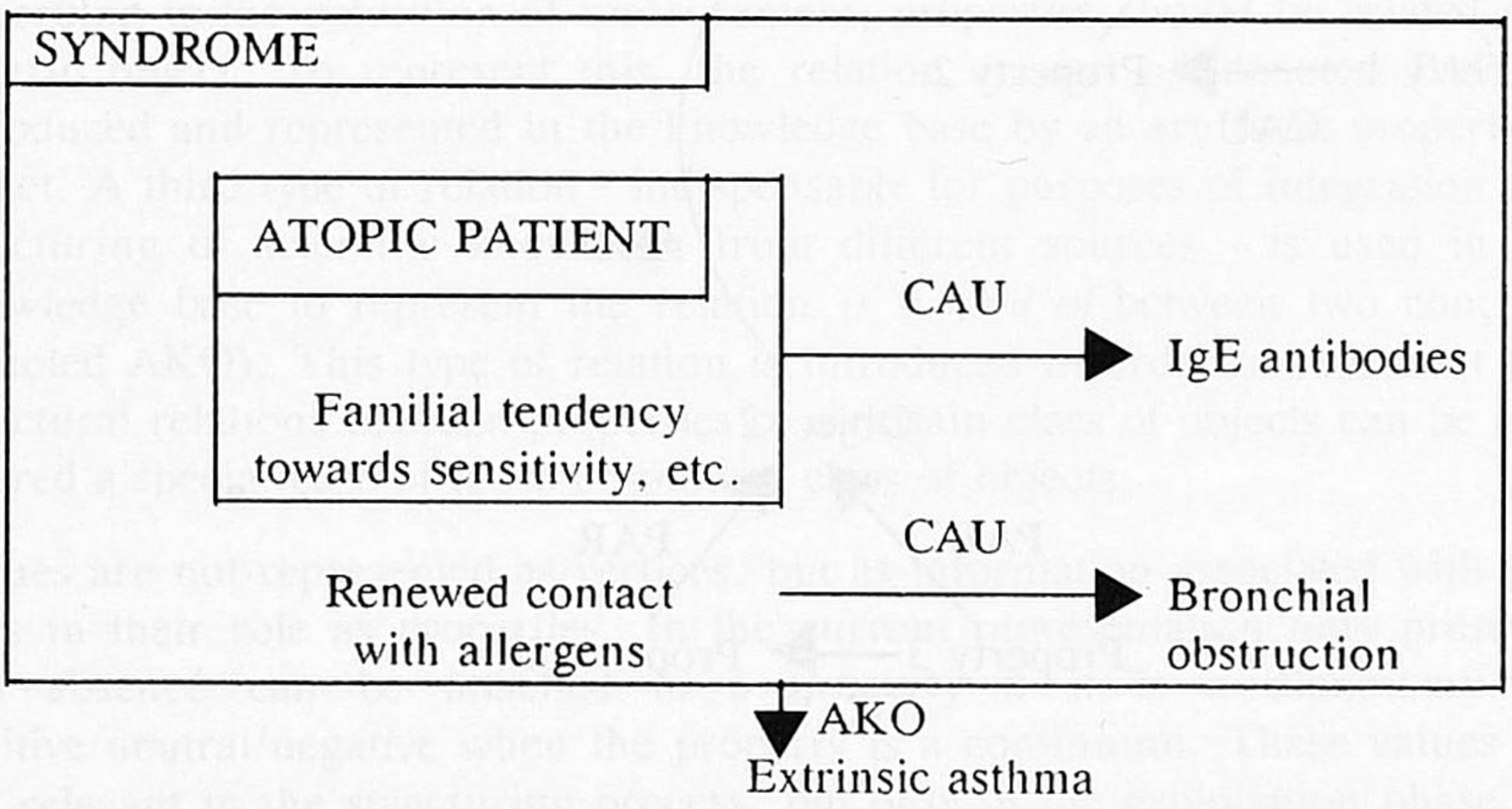



*Figure 2.* The representation of the text examples a and b using frameworks

a bronchial obstruction. This syndrome is an example of extrinsic asthma (Example b)

Therefore, a general solution was sought, and consisted of the introduction of a new primitive into the knowledge representation: the framework. A framework represents a whole set of concepts and relations. A new relation, the FPAR relation, represents the relations between a framework and its constituent concepts and relations. Figure 2 shows the representations of the above examples with frameworks. The main features of a procedure for text analysis that has been developed within the context of the present project are given later.

Frameworks and the FPAR relation make it possible to represent relations between whole processes. But the FPAR relation can also be seen as a generalization and an alternative representation of the PAR relation: all PAR relations are replaced by frameworks. The framework then represents an object and the vertices within the framework its properties (see Figure 3). When a knowledge graph is constructed for a certain class of objects, e.g., a class of patients in a medical application, the whole graph can be considered as a framework. A PAR relation connects the arcs and vertices of a graph (representing medical processes in this case) to a framework (in this example the patient). The introduction of the framework makes it possible to represent explicitly the class of objects considered.

A representation with frameworks is no longer a graph-theoretical representation and consequently precludes the application of the full variety of graph-theoretical concepts and procedures that was the main reason to search for that type of representation. The above representation can be transformed into a strictly equivalent representation in which all the above-mentioned elements (vertices, frameworks and arcs) are represented as the vertices of the new graph. When a vertex or arc is contained in a framework, this is represented in the new graph by arcs from the constituent elements of a framework to the vertex representing the framework. The other arcs in the new graph represent

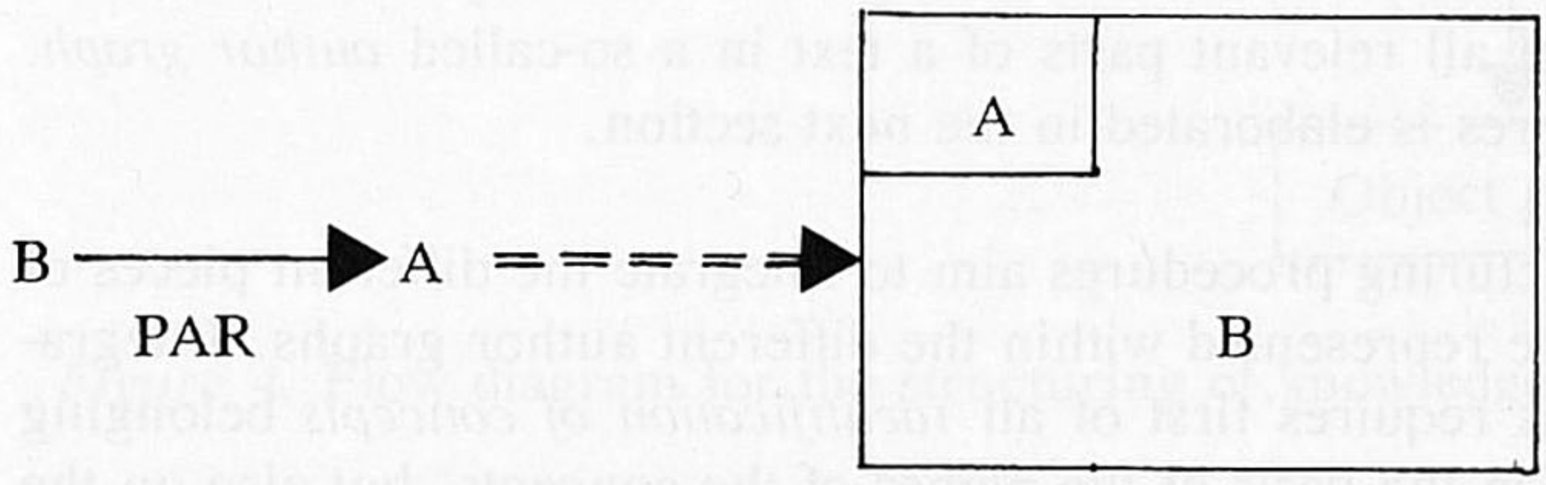



*Figure 3.* Alternative representations of the PAR-relation

the incidence relations between the vertices and arcs in the original representation. The chosen representation is known as the total graph.

A graph-theoretical representation has major advantages above other representations because of the large number of graph-theoretical procedures and concepts that can be applied meaningfully to graphs representing (scientific) knowledge. These procedures and concepts concern such aspects as features of the overall structure of the graph, the (relative) centrality of arcs and vertices within a graph, and the detection of subgraphs with certain characteristics. These aspects may be less important for systems that are restricted to the representation of the knowledge of one expert or (scientific) scource, but they are highly important in the light of the aims of the present project that is oriented towards comparing, structuring, and integrating scientific knowledge obtained from different sources.

## 5. Procedures for Structuring Knowledge in a Graph

The flow diagram given in Figure 4 specifies the system for the manipulation of knowledge graphs. It contains the basic classes of procedures for structuring and integrating knowledge into an integrated graph and the class of procedures to use the integrated knowledge base for exploitation. To represent the main knowledge in a particular field of science the first basic problem consists of the selection of sources from which the knowledge can best be collected. Several recommendations might well be proposed by the system, but no explicit procedures will be developed to support the decisions of the researcher in this respect. Formally, however, it can be seen as a first step in the structuring of knowledge in an integrated knowledge base. More important for us are the other four classes of structuring procedures, because our project aims to develop well-defined procedures to support knowledge engineering in those respects. Extraction of the relevant concepts and relations from a text is the first major class of structuring procedures that is considered in the project. This extraction through *textual analysis* should result in a representation of all relevant parts of a text in a so-called *author graph*. This class of procedures is elaborated in the next section.

Three classes of structuring procedures aim to integrate the different pieces of knowledge as they are represented within the different author graphs. Integration of author graphs requires first of all *identification of concepts* belonging to different graphs. On the basis of the names of the concepts, but also on the basis of structural equivalence of concepts and subsets of concepts in the graphs, concepts that are identical should be identified. In the context of con-

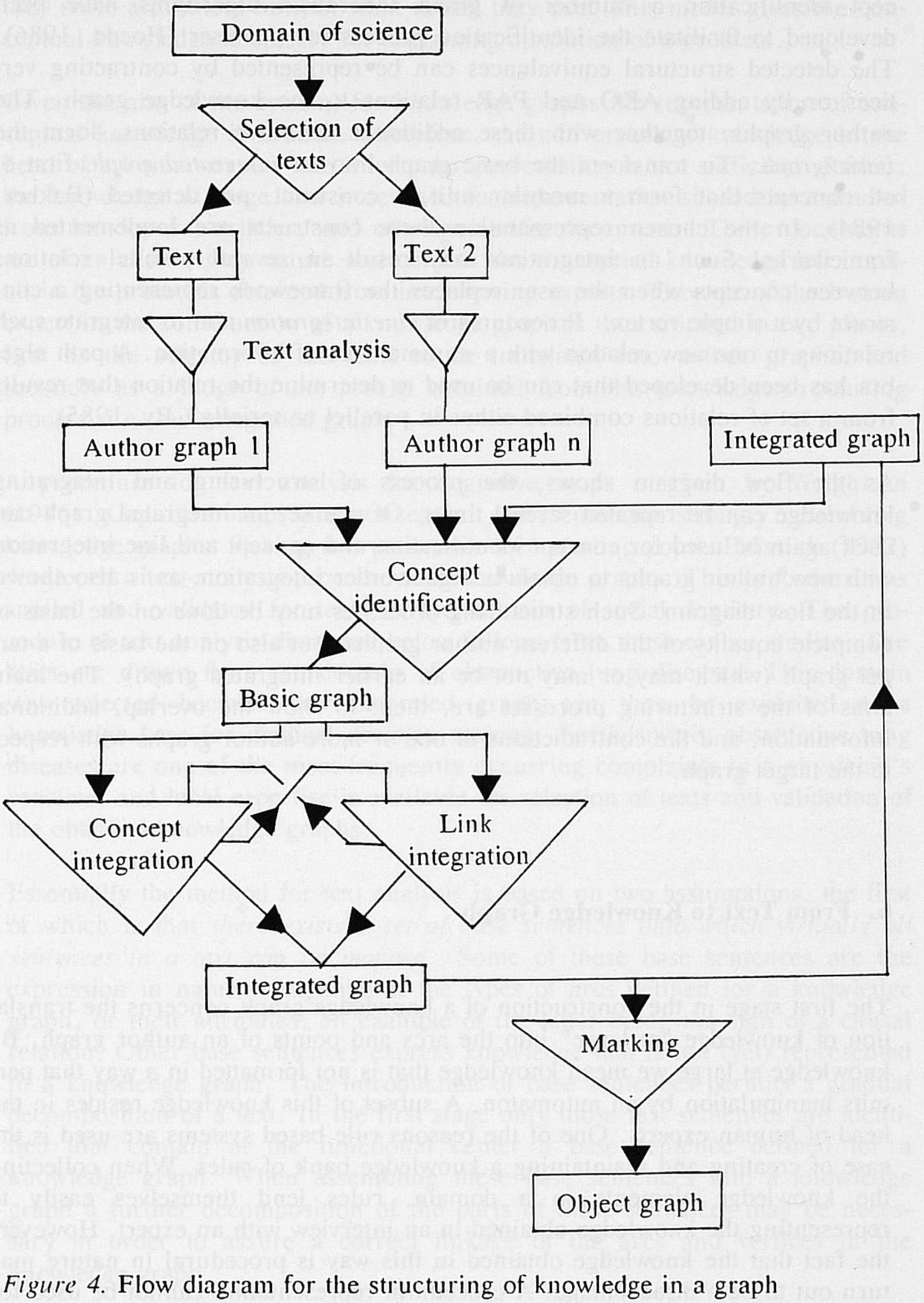



*Figure 4.* Flow diagram for the structuring of knowledge in a graph

cept identification a number of graph theoretical algorithms have been developed to facilitate the identification process for the user (Hoede, 1986). The detected structural equivalences can be represented by contracting vertices or by adding AKO and PAR relations to the knowledge graph. The author graphs, together with these additional definitive relations, form the *basic graph*. To transform the basic graph into an *integrated graph*, first of all concepts that form a modular unit, a construct, are detected (Bakker, 1984). In the chosen representation these constructs are implemented as frameworks. Such an integration may result in several parallel relations between concepts when the user replaces the framework representing a construct by a simple vertex. Procedures of *line integration* aim to integrate such relations to one new relation with a minimal loss of information. A path algebra has been developed that can be used to determine the relation that results from a set of relations combined either in parallel or serially ( By, 1985).

As the flow diagram shows, the process of structuring and integrating knowledge can be repeated several times. Of course, an integrated graph can itself again be used for concept identification and concept and line integration with new author graphs to obtain a higher-order integration, as is also shown in the flow diagram. Such structuring processes may be done on the basis of complete equality of the different author graphs, but also on the basis of a target graph (which may or may not be an earlier integrated graph). The main aims of the structuring processes are, then, to show the overlap, additional information, and the contradictions of one or more author graphs with respect to the target graph.

## 6. From Text to Knowledge Graph

The first stage in the construction of a knowledge graph concerns the translation of knowledge "at large" into the arcs and points of an author graph. By knowledge at large we mean knowledge that is not formatted in a way that permits manipulation by an automaton. A subset of this knowledge resides in the head of human experts. One of the reasons rule based systems are used is the ease of creating and maintaining a knowledge bank of rules. When collecting the knowledge elements in a domain, rules lend themselves easily to representing the knowledge obtained in an interview with an expert. However, the fact that the knowledge obtained in this way is procedural in nature may turn out to be a disadvantage. A procedural representation cannot be used for generating adequate explanations, as was outlined in the overview of knowledge representation. This inadequacy follows directly from the way in which the representation was constructed. Often, experts can only say how

they solve a problem but not what model they use. This model gives the conceptual knowledge necessary for explaining a solution of a problem.

Given the aim of the system, for knowledge graphs the representation of conceptual knowledge is crucial. Therefore, the conversion of knowledge in the head of a human expert into a formal representation is not sufficient. Rather, the knowledge at large that is selected for representation should be extracted from handbooks, articles, and other scientific documents. These documents are of an explanatory nature and contain the conceptual knowledge left implicit in the protocols obtained from interviewing experts. The role of the human expert in the construction of a knowledge graph is thus situated in two places, first as the selector of the texts to be translated into author graphs, and secondly as a judge of the results obtained from the knowledge-structuring process, i.e., an integrated graph.

A procedure for text analysis for cognitive maps is given by Wrightson (1976). Cognitive maps are conceptual representations strongly resembling knowledge graphs. Taking this procedure as a starting point, Buissink (1982) developed a text-analysis procedure for knowledge graphs. This procedure was applied and validated for texts in the domain of social dentistry. The procedure of text analysis discussed below follows from these earlier studies. The texts are drawn from the domain of obstructive lung diseases. This domain was selected because the integrated graph can then be exploited as a knowledge base for medical decision making. Furthermore, obstructive lung diseases are one of the most frequently occurring complaints in a physician's practice, and local expertise is available for selection of texts and validation of the obtained knowledge graphs.

Essentially the method for text analysis is based on two assumptions, the first of which is that *there exists a set of base sentences onto which virtually all sentences in a text can be mapped.* Some of these base sentences are the expression in natural language of the types of arcs defined for a knowledge graph, or their attributes, an example of the latter being the sign of a causal relation. Other base sentences express knowledge that is not (yet) represented in a knowledge graph. The introduction of base sentences permits a gradual decomposition of a text. In the first stage only those text sentences are identified that contain at the functional center a base sentence defined for a knowledge graph. When assembling these base sentences into a knowledge graph a further decomposition of the parts of these sentences may be necessary in order to assure a correct linkage of the arcs and vertices in the knowledge graph.

The second assumption in the method of text analysis concerns the information conveyed by base sentences, and is that *The object for which the knowledge graph is constructed is defined.* The object, (see the overview) acts

as a context in which base sentences are evaluated. We assume that in a scientific text at least two objects are described. the object investigated and the object conducting the investigation, i.e., one or more scientists. The description of the former gives a structural model of the observed object, which the scientist claims is, at least to some extent, independent of his observation. A scientist does not claim such an objective status for the sentences describing the second object, i.e, him- or herself. These sentences express the motives that led to the investigation of the first object. Although these sentences might contain base sentences expressing a type of arc defined for a knowledge graph they must be skipped because they are not a part of the structural model of the predefined object of investigation.

The decomposition of a text sentence into a set of base sentences is accomplished by means of a so-called *dependency grammar*. A dependency grammar is based on the concepts of argument, operator, and modifier. We speak of dependency because in this grammar an argument is defined as dependent on an operator and an operator as dependent on a modifier. The grammar gives an instrument for determining the functional center of a sentence, i.e., the main operator in a sentence. Dependency grammars have been analysed by Harris (1982) for their application to natural language. Their implementation in a computer program is described by Sager (1981) .

The concepts involved in a dependency grammar can be defined as follows:

*Arguments* are words on which other words do not depend. Typically, these words are (non-relational) nouns denoting "static" objects like box, chair, etc.

*Operators* are words of which other words depend as arguments. Examples of operators are verbs, adjectives, and relational nouns (e.g., father of, example of). Operators differ in the number of arguments they bind. In a knowledge graph all relations, including frameworks, are modelled as binary relations. For this reason we assume that the maximum number of arguments bound by an operator is two. Operators that may seem to bind more arguments can be handled by adding modifiers to a two-placed operator.

*Modifiers* are words on which an operator or other modifier is dependent. Adverbs are examples of one-placed modifiers. Modifiers can also be two-placed, in which case they relate an object to an action, or an action to another action. Prepositions are examples of the former. Examples of the latter are conjunctions like *while* and *because*.

The concepts of argument, operator, and modifier can be applied recursively to decompose a sentence. This decomposition is recursive because the arguments dependent on an operator can be considered as constituents that can again be decomposed into operators, arguments, or modifiers. An example of

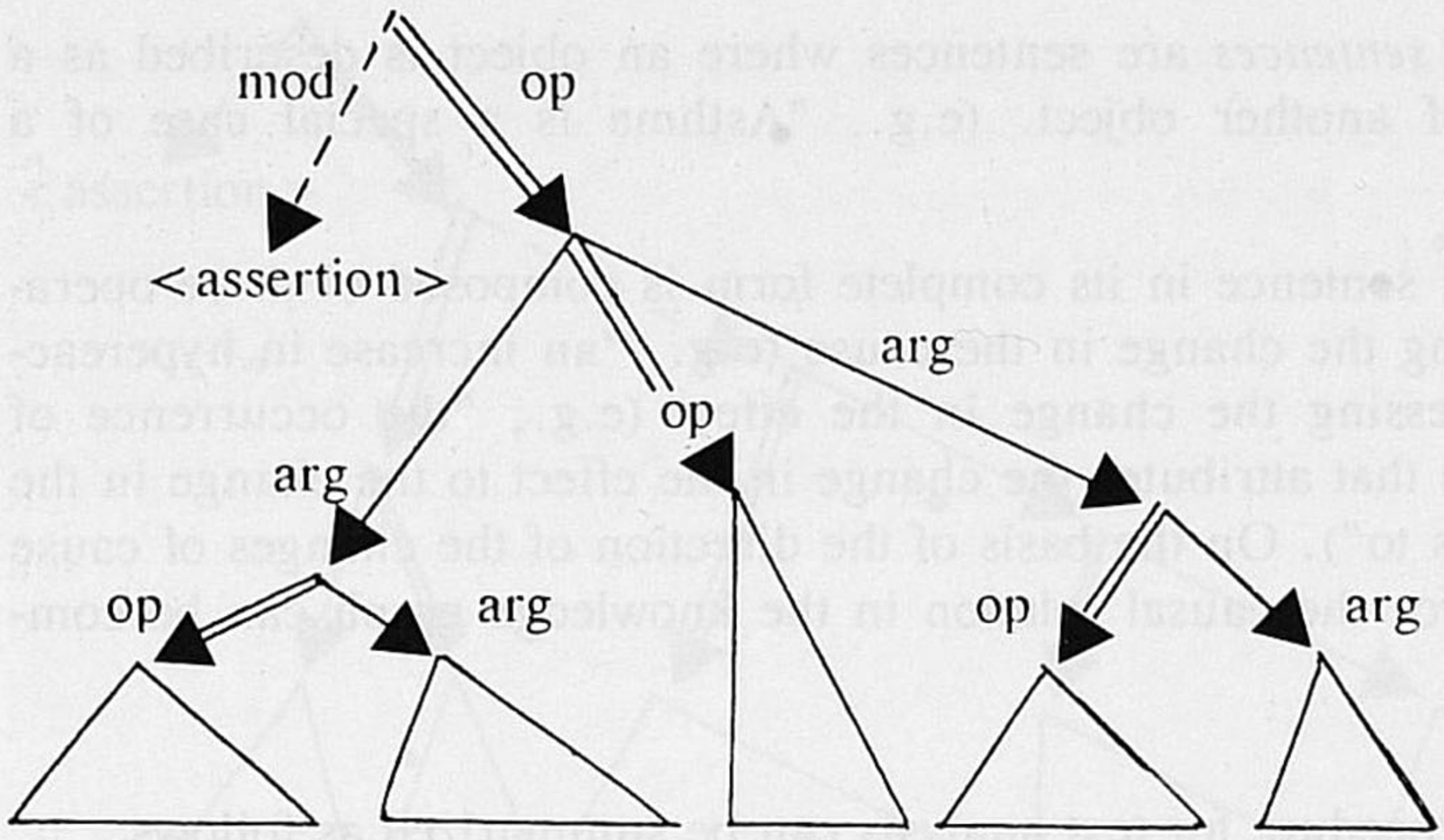



*Figure 5.* Decomposition of a sentence in arguments (*arg*), operators (*op*) and modifiers (*mod*)

a recursive definition of a sentence is given in Figure 5. Note from Figure 5 that the fact that the sentence in this example is an assertion is represented as a modifier of the central operator. In a knowledge graph only sentences that are assertions are represented. Sometimes an assertion is itself the main operator, as in a sentence like "We emphasize that bronchial obstruction is an important characteristic of asthma." In these cases the message, i.e., that which is asserted (which is the second argument of the central operator), is further decomposed. Dependency grammars permit an operator to be not explicitly stated in a sentence because it does not contribute to the informativeness of that sentence. For instance, it is in general not informative to state explicitly that a sentence is an assertion. Likewise, some operators that are indicated by the verb *to be* are sometimes left out due to their noninformativeness. Examples of such a form of operator reduction are sentences like "Asthma is a disease." Reintroducing the full operator into this sentence gives "Asthma is an example of a disease".

In order to decompose a text adequately we have extensively defined the operators that stand for the types of relations in a knowledge graph. The sentences giving these operators can be described as follows:

*Causal sentences* are sentences where a change in one property is described as structurally dependent of that in another property (e.g., "An increase in hyperreactivity leads to the occurrence of asthma")

*Composition sentences* are sentences that identify a property of an object. (e.g., "Obstruction of the bronchus is a property of asthma")

*Generalization sentences* are sentences where an object is described as a special case of another object. (e.g., "Asthma is a special case of a disease")

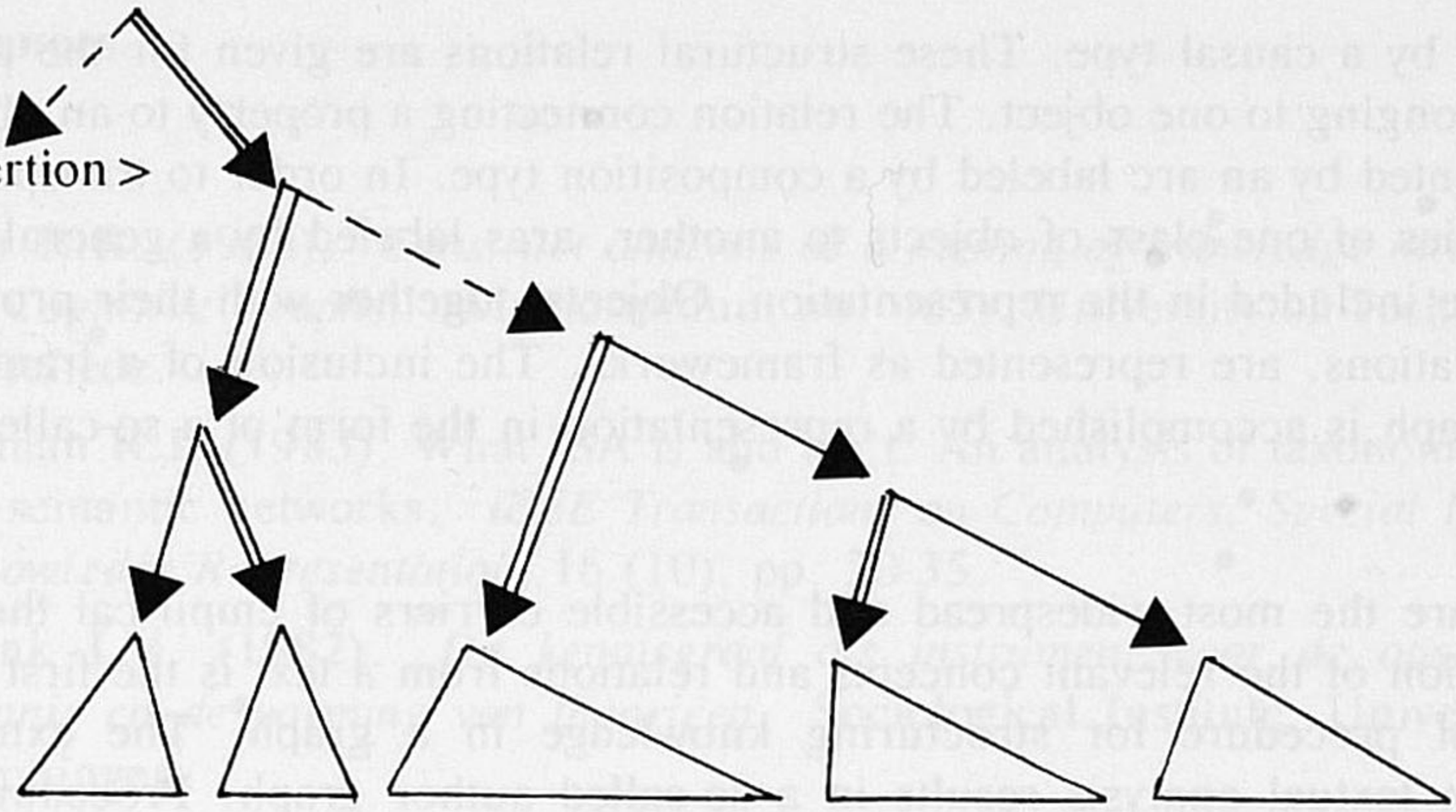
Note that a causal sentence in its complete form is composed of three operators, one expressing the change in the cause (e.g., "an increase in hypereactivity"), one expressing the change in the effect (e.g., "the occurrence of asthma"), and one that attributes the change in the effect to the change in the cause (e.g., "leads to"). On the basis of the direction of the changes of cause and effect a sign for the causal relation in the knowledge graph can be computed.

The steps in the procedure for text analysis can be summarized as follows:

1. Search for operators, arguments and modifiers in sentence
2. Identify message in assertion
3. Classify main operator in message as a relation-type
4. Identify corresponding arguments as nodes or frameworks
5. Classify main operator in modifier as a relation type
6. Identify corresponding arguments as nodes or frameworks

It is worth noting that in a message both the operator and its modifier have to be inspected. The reason is that the causal relation can occur as a combination of an operator and a modifier. Generally, the causation is expressed by the main operator in the message and the changes of the properties are contained in its arguments. An example of such a sentence was shown in figure 5. However, it is also possible that the main operator in the message expresses a change and that its modifier reflects the causation (see Figure 6).

The method of text analysis described here is still under development. With respect to the decomposition of sentences into arguments, operators, and modifiers, we have examined to what extent automatic procedures can support the detection of central operators that express a type of relation defined for a knowledge graph. Furthermore, reliability studies will be undertaken to estimate the agreement among several lay human analysers of a text. These studies will show the variability that occurs due to text phenomena that are hard to formalize. Basically these phenomena concern two issues, namely, the identification of the appropriate context (i.e., the object of investigation), and the correct resolution of references occuring in sentences describing this object. Both issues still are an important bottleneck for an automatic analysis of text. The results of reliability studies will show to what extent a knowledge graph correctly represents the central assertions in a text.




Asthma occurs as a consequence of an increase in hyperreactivity

*Figure 6.* Causal sentence with main operator of the message reflecting change

## 7. Summary

In the preceding discussion we have elaborated the view that knowledge is a commodity that can exploited in various ways. Examples of such exploitations are decision making, information retrieval, and instruction. In order to exploit a knowledge base to its full extent it is necessary to seperate the knowledge in a certain area of interest from the task-specific inferences defined on it. In an overview of knowledge-based systems we have observed that such a seperation of domain and control knowledge is realized in systems based on a conceptual knowledge representation. Procedural representations, by contrast, are fixed with respect to a predefined task. Here, control and domain knowledge are merged into a knowledge base of rules. An important restriction of most knowledge-based systems is that their knowledge is typically limited to that of one expert. In general, few procedures are provided to evaluate the effect of "new" knowledge on an existing knowledge base.

The central issue in the discussion has been how conceptual knowledge from different sources can be obtained and organized in a coherent representation. Within the domain of theories from medicine and the social sciences a representation in the form of a knowledge graph is given. Given the aim of integrating knowledge from different sources, a small set of types of relations is taken as a starting point. These types are a minimal requirement for representing the knowledge embodied in theories in medicine and the social sciences. The structural relations specified in a theory are represented by arcs

labeled by a causal type. These structural relations are given for the properties belonging to one object. The relation connecting a property to an object is represented by an arc labeled by a composition type. In order to transpose the properties of one class of objects to another, arcs labeled by a generalization type are included in the representation. Objects, together with their properties and relations, are represented as frameworks. The inclusion of a framework in a graph is accomplished by a representation in the form of a so-called total graph.

Texts are the most widespread and accessible carriers of empirical theories. Extraction of the relevant concepts and relations from a text is the first major class of procedure for structuring knowledge in a graph. The extraction through textual analysis results in a so-called author graph. Procedures for concept identification transform a set of author graphs into a basic graph. Within a basic graph, concepts as well as relations can be integrated. In order to reason with different types of relations a path algebra was defined. The structuring of knowledge in a graph is a cyclic process that can take place on the basis of complete equality among author graphs, but also on the basis of a target graph.

For the extraction of an author graph from a text a dependency grammar is defined. According to this grammar, a sentence in a text is parsed into a structure of operators, arguments, and modifiers. Only sentences expressing an assertion are accepted as building blocks for a knowledge graph. In the asserted message the main operator and the operator in the modifier are provisionally classified as one of the three defined relation types. Their arguments coincide with vertices or frameworks, that may appear in the role of object or property (or both). A dependency grammar permits scanning of only the surface of a text. The outcome of the text analysis, i.e., the author graph, is only an approximative model of the knowledge in a text. It is sufficient, however, with respect to the structuring of knowledge obtained from different sources, such as the empirical domains mentioned above.

*Acknowledgements*

The research group on knowledge graphs consists of C. Hoede, R.R. Bakker, H.J. Smit (Twente University of Technology), F.N. Stokman, and P.H. de Vries (University of Groningen). The research is funded by the Dutch Organization for Scientific Research under grant no. 40-029. The research is done in cooperation with the MEDES project of the University Hospital Groningen under the direction P.F. de Vries Robbé. We thank Kees Hoede for comments on this chapter.

## References


Bakker R.R. (1984). *Construct analysis as a method of knowledge integration of Cognitive Graphs.* (Memorandum no. 463), University of Technology, Enschede.

Brachman R.J. (1983). What ISA is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Transactions on Computers, Special Issue on Knowledge Representation*, 16 (10), pp. 30-35.

Buissink J.V. (1982). *De kennisgraaf als instrument voor de opslag van kennis en de vorming van theoriëen.* Sociological Institute, University of Groningen.

By R. de (1985). *Semantische aspekten van paden in kennisgrafen.* Univ. of Technology, Enschede.

Davis R. (1982). Teiresias: Applications of meta-level knowledge. In: *Knowledge-based Systems in Artificial Intelligence.* R. Davis & D.B. Lenat (Ed.), McGraw-Hill, New York. pp. 227-490.

Davis R., Buchanan B. & Shortliffe E. (1977). Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence*, 8, pp. 15-45.

Harris Z. (1982). *A grammar of English on mathematical principles.* Wiley, New York.

Hoede C. (1986). *Similarity in knowledge graphs.* Department of Applied Mathematics, Twente University of Technology, Enschede.

Krantz D.H., Luce R.D., Suppes P. & Tversky A. (1971). *Foundations of measurement: Vol. I Additive and polynomial representations*, Academic Press, New York.

Kuipers B. (1984). Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence*, 19, pp. 39-88.

Lenat D.B. (1983). EURISKO: a program that learns new heuristics and domain concepts. *Artificial Intelligence*, 21, pp. 61-98.

McDermott J. (1982). R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 19, pp. 39-88.

Pfanzagl, J. (1968). *Theory of measurement.* Physica-Verlag, Wurzburg.

Riet R.P. van (1983). Knowledge bases - de databanken van de toekomst. *Informatie*, 25, pp. 16-23.

Robinson J.A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM*, 12 (1), p. 23.

Ryle G. (1949). *The Concept of Mind.* Hutchinson, London.

Sager N. (1981). *Natural language information processing: A computer grammar of english and its applications.* Addison-Wesley, Reading, MA.

Simon H.A. (1977). *Models of discovery and other topics in the methods of science.* Reidel, Dordrecht.

Vries Robbé P.F. de (1978). *Medische besluitvorming: een aanzet tot formele besluitvorming.* Dissertation. University of Groningen.

Vries Robbé P.F. de & Vries P.H. de (1985). Epistemology of medical expert systems. In: *Medical decision making, diagnostic strategies and expert systems.* J.H. v. Bemmel, S. Grèmy & J. Zvárová, (Eds.), North-Holland, Amsterdam. pp. 89-94.

Winograd T. (1980). Extended inference modes in reasoning by computer systems. *Artificial Intelligence*, 13, pp. 5-26.

Winston P.H. (1975). Learning structural descriptions from examples. In: *The Psychology of Computer Vision.* P. Winston (Ed.). McGraw-Hill, New York. pp. 157-209.

Wrightson M. (1976). The documentary coding method. In: *The Structure of Decision.* R. Axelrod (Ed.). Princeton University Press, Princeton. pp. 291-332