REMARKS ON SEMANTIC INFORMATION DESCRIPTION BY NOUN PHRASES

G. Rahmstorf

November 1975 WP-75-153

INTRODUCTION

This paper summarizes some ideas about a new method for information retrieval and data description based on natural language features, which is applicable to both formatted and textual data.

The language's ability to express specific terms by noun phrases is particularly useful for describing information requests (queries) and data related to interdisciplinary fields, i.e. energy or environmental research. These fields are characterized by a fluctuating terminology, variety of different data and by information requests, which are difficult to predict.

Therefore, it is expected that the proposed information retrieval and data description method will be more efficient than the currently used methods, which are based on Boolean expressions.

To prove this a theoretical study and programs have to be accomplished. A schedule for such a task and an overview on the main subjects is given in Figure 1. Arrows in the flowchart mean "x is dependent on y."

The study is based on current project activities at IIASA, and the research proposal for 1976 on a data base (question-answering system) for the applied projects. To connect the study with IIASA's computer network project, special attention has been given to information retrieval in distributed data bases of a network.

The complexity of data bases require a step by step procedure, and it is necessary to define the system, mode of operation, application, and to commit resources in terms of machine, software, storage, manpower and data supply, before a data base can be implemented and maintained. Therefore, we suggest to start a technical subset (information retrieval and data description) and an application subset (energy data) of the broad problem field as a first step in the proposed direction.

A. Energy Data Analysis

The feasibility of methods in data storage and retrieval can only be studied by concrete applications. Therefore, we have to know what kind of data is needed by IIASA's projects, and

what the data problems of our specific research fields are. Each research field has different problems related to structure and usage of data, and it is necessary to choose a project as the application area for this investigation. We intend to concentrate our study on energy data, which is one of IIASA's most important research fields, because some requirements for energy data collection and analysis have already been discussed (Butrimenko and Häfele, 1975). Further close cooperation with the Energy Project is necessary to summarize our knowledge of present energy data base projects in the world, their structure, semantic scope, and availability of data. We should also specify the analysis of IIASA's energy data requirements in more detail.

B. Scientific Information and Communication

Obtaining information and accessing energy data is an example of the general problem of knowledge representation and scientific communication. Storing and retrieving data by use of computers or printed publications can be viewed as basic operations in a network system of international scientific communication. The structure and flow of this system should be analyzed. It is expected that a more intense use of computers to represent, retrieve, distribute, edit and process scientific results will enhance the productivity of the system and reduce the redundancy of results.

C. Functions of Information Systems

The term "information systems" is used for different computer applications, for example:

- Text retrieval systems
- Systems for formatted data
- question-answering systems
- systems for on-line measurements and control of environmental activities.

Therefore, it is necessary to identify IIASA's requirements by functional specifications. Questions to be discussed at this point are: what is the purpose of information systems? What are their general functional characteristics? What kind of operations are available for users of information systems?

D. Relations Between Data Aspects

An information system is not only characterized by functional specifications (C) but also by the data appearing in the data base and in the expressions of the communication language used in the dialogue between user and system.

Four aspects of data elements are distinguished:

- Name (identifier, label)
- Meaning (logical relations to other data elements)
- Syntactical structure (data type as integer, character string)
- Physical structure (location, signal pattern, time)

An analysis of data aspects is the theoretical base for a method called "access by meaning description" to be proposed for our further investigations.

E. Distributed Data Bases in Computer Networks

The Computer Science Project is involved with computer networks, because IIASA as an international institute cooperating with many national institutions and organizations is affected by problems of scientific information exchange. A computer network is a tool to enhance scientific communication by providing access to data bases of remote computer installations. Distributed data bases are a new research topic, and as there are only a few practical experiences a terminology has not been developed. Therefore, it is necessary to describe the problems of distributed data bases and to show, which problems can be solved by our proposal, as described under heading H.

F. Query Languages: Information Retrieval

Data bases in scientific computer networks differ in data structure, access method, query language, etc. depending on the hardware and software installed in the local computer centers. In order to come to a conclusion about standardization

or query language translatability currently used query languages for textual and formatted data have to be analyzed.

G. Inference, Problem Solving

Question-answering systems as discussed in artificial intelligence literature are information system with:

- An interface for (restricted) natural language;
- A highly structured data base (fact retrieval);
- Inference and problem solving capabilities.

One advantage of highly structured data is that is allows application of inference rules to produce relevant data not explicitly stored in the data base. We understand that this is an important subject, but it seems appropriate to start with problems of knowledge representation, data description and information retrieval.

H. Proposal: Noun Phrases for Information Description

To solve the problem discussed in E and F we propose a data base catalogue that will contain information about the meaning and location of data files in the network. It is also possible to store additional information concerning other sspects of files, e.g. structure, name of record fields and query language to be used for the data base containing that particular file. It is planned to use (restricted) noun phrases of English and German to specify an information request, for example,

PRODUCTION OF NUCLEAR ENERGY IN FRANCE IN THE YEAR 1972

It is explained that noun phrases are a more precise tool than Boolean expressions currently used for describing an information request to a retrieval system, and therefore the precision and recall of the system will increase. The noun phrase query language is not restricted to a set of descriptors, but the system requires a lexicon containing the words of the language, which could be used in queries. The catalogue will answer queries by specifying locations of files with relevant contents.

The user may then access this file by using a query language which is necessary for the addressed data base system. This method avoids changing the existing data bases used in a network. It also avoids standardization and query language translations, but it is necessary to register data files of the network in a common catalogue. Because file entries in the catalogue are not only used to identify files, but also to describe data meaning, the structure of the catalogue is a semantic network (see subject K). The semantic network is language independent, and it can be used with noun phrases of English or other languages, because it is a concept network, and not a representation of syntactical structure of a specific language.

A working paper containing more details on our proposal is being prepared.

I. Formatted Data: Explanation and Nominalization

The catalogue mentioned in (H) will describe files of structured data, text data, programs or graphical information by concepts expressed as noun phrases. We prefer to use the relational data model as a base for the study on description of formatted data.

The meaning of data in a relation (set of data records) can be expressed by (1) a bracket notation, (2) a language sentence of verbal form, and (3) by noun phrases, for example:

(1) CONSUMPTION (REGION, OIL QUANTITY, TIME PERIOD)

(2) Region x consumes y barrels oil in the time period z

(3) CONSUMPTION OF Y BARRELS OIL IN THE PERIOD Z BY REGION X

The translation from (1) to (2) is data explanation by verbalization; translation from (2) to (3) is called nominalization.

In our proposal description of files by noun phrases has to be done intellectually by humans, but nevertheless a formalized translation is to be studied because it provides a tool

for relating text information retrieval and formatted data.

J. Text Data: Description by Titles

Describing text data blocks by elements of a thesaurus or some other means is called indexing. Large research activities in documentation and information retrieval are concerned with automatic indexing. We assume in our proposed system that every text information block is described by a title. No text comes without a title. Titling is the natural mode used by an author to describe the meaning of his text (scientific report, chapter in a book). Most of the titles used for scientific texts are noun phrases. But not every title given to publication is complete or exact enough for the purpose of a retrieval system. Therefore, some additional intellectual work on titles is necessary to adjust publication titles to the information retrieval requirements.

K. Knowledge Representation by Concept Networks

The meaning of a noun phrase describing a user's information request or a file (or text) of the information system is a concept. Every concept is logically defined by its relation to other concepts. These form a network, which is called a concept or semantic network.

The set of n-ary relations between concepts is open if the expressibility of the network shall be unrestricted. But we expect to find a limited number of relations that can be used to define all other relations. This approach has been used by the new linguistic theory called generative semantics. We also will take advantage of the experiences with semantic network projects described in the latest computer science literature. Essential attributes may be understood as those concepts, which are used in relations that define a concept. For example, the following relations contain essential attributes, because they define the concept reactor:

PRODUCING (REACTOR, ENERGY, NUCLEAR FISSION)

SUBCONCEPT (REACTOR, POWER PRODUCING DEVICE)

Accidental attributes contribute additional knowledge

about a concept, for example, the price of a nuclear reactor, length of time for developing, the mode, how it is controlled and the risks related to the operation of the reactor. The same network structure is used to store essential and accidental knowledge. Although it is difficult to map n-ary relations between concepts by a graphical network representation, the network seems to be the appropriate structure for concept relations.

L. Functional and Structural Design

A software package should have to be developed to prove the feasibility of noun phrases for information description. This software package should use energy data as an experimental base (see subject M).

As described earlier the noun phrase system that is used in a network (or even a single system) does not require changing existing data base systems. The user asks the noun phrase system about stored data and receives information about the name or address of a relevant file. He can then access this file and receive the required data.

The system consists of the noun phrase processor, concept network, file catalogue and dialogue processor as described in figure 2 . The noun phrase processor is related to one communication language, which in this case is English noun phrase. It encompasses syntax and semantic rules and a lexicon. The lexicon defines words (nouns, adjectives, prepositions) by syntactical predicates, and by pointers to the concept network. The pointers to the nodes of the network define the meaning (1) of a word, and the file catalogue contains names (or addresses) of all formatted and textual files available in the network. Each file is described by a pointer to a node of the concept network, which is the common meaning description for user and system. Another type of attribute in the file catalogue may specify how the file is structured and accessed.

The dialogue processor controls the communication process between user and system. It seems appropriate to separate it from the noun phrase processor, because it may be useful to change the communication languge (to German noun phrases or to formal expressions) without impact on the whole software package. The noun phrase system may be located in any node of the computer network, but the initial requests of the users coming from various terminals of the network must be guided to the noun phrase system by the network control programs. The concept network has to be represented by a physical data structure. If the concept network becomes larger the problems of access time and data organization on external storage devices become more important. On the other hand it is intended to use available data management services as much as possible to avoid extensive system programming. It is neceessary to analyze which language is more suitable to implement conceptual knowledge of the network.

M. Data Collection and Implementation

To study new query methods for textual and formatted data it is necessary to have experimental files, and as stated before energy data is proposed for various reasons.

By experimenting with a smaller set of energy data we can also study design, implementation, control, application and updating of larger energy data base systems. We are then prepared to start succeeding activities at IIASA or to establish a complete large energy data base. The intended experimental system requires the following types of stored information:

1. Files of formatted--mostly numerical data
   (e.g. energy resources, conversion, distribution, technological processes, environmental effects);

One possible source of data: Brookhaven National Laboratory

2. Files of textual data
   (e.g. bibliographic data and abstracts of scientific reports on energy problems relevant to the IIASA project)

Source of data: at IIASA available reports, IAEA Vienna (INIS)

3. Technical terms of energy research
   These terms have to be defined formally by other terms and relations (as part of the concept network);

Source of data: ÖNORM A7000 - A7009, technical glossaries

4. A set of words of conversational language
   These nouns and adjectives should be the most frequently used words of noun phrases describing concepts of energy research. They have to be defined formally.

Source of data: German and English dictionaries, Energy Research Report Titles

Formatted (1) and textual (2) energy files can be used by any kind of program independently of the development and completion of the noun phrase query system to be implemented on top of those files. The development of the lexicon and the concept network may take longer, but practical use of energy files by mathematical models, statistical analysis, etc can go on at the same time. It is expected that data collection, preparation and keypunching will be a very time consuming activity.

N. Semantics

Semantics is the relation between valid syntactical expressions of the communication language and concept network. Semantic rules associate a sub-structure of the network to a syntactical expression. The associated sub-structure of the concept network is understood here as the "meaning" of the syntactical expression.

The meaning of a noun phrase is always a concept. Complete sentences which represent a relation between concepts (a proposition) are not part of the noun phrase communication language. The simplest noun phrases are single words. The meaning of the words are defined in the lexicon by pointers to the concept network. It is more difficult to define the meaning of simple syntactical structures. We have to decide how the following fundamental syntax structures of noun phrases are represented in the concept network:

(ADJ NP)	Adjective attribute
(NP NP_G)	Genitive attribute
(NP1 PREP NP2)	Preposition attribute
(N1 N2)	Complex noun

The semantic analysis is complete if the type of relation is determined between the concept associated to the noun phrase and each concept associated to each element (word) of the noun phrase. If, for example, NP NP_G represents the meaning of the syntactical expression (NP NP_G) it is necessary to analyze what relations exist between:

NP	NP_G	and	NP
NP	NP_G	and	NP_G
NP		and	NP_G

These relations are dependent on the definitions of the concepts NP and NP_G.

O. Structure of the Lexicon

The lexicon is an ordered list of words of the communication language. Syntactical predicates necessary to analyze noun phrases are associated with every word, for example, word class (noun, adjective, preposition). The meaning of a word is defined by pointers to nodes of the concept network. There are several unsolved lexical problems, and the main one in our context is the semantic description of words which represent a relation (and not a concept).

P. Syntax of Noun Phrases

Syntactical structure of noun phrases can be described by context free production rules:

NP1 → N

NP1 → ADJ NP1

N → N N

NP1 → NP1 N_G

NP2 → NP1 PREP NP1

NP2 → NP2 PREP NP2

(N = noun; ADJ = adjective; PREP = preposition;
NP1, NP2 = noun phrases; NG = noun in genitive case).

Relative clauses are not allowed here within noun phrases. As a result of a syntax analysis the system associates with a given noun phrase a valid string of symbols of the set x

X =(N,ADJ,PREP,NG)

Each symbol has to be recognized as an element of the lexicon.

Q. Design of an IIASA Question-Answering System

A conference on question-answering systems was held at IIASA in June 1975. The working group of this conference proposed to develop a question-answering system at IIASA, which would not be restricted to noun phrases and information retrieval. But it is necessary to limit research goals for the next time period to the described subjects, which are the necessary steps towards a more complex question-answering system. Inference and problem solving should be based on information in an expressive concept network. Results of the proposed data base study may influence the design of a more powerful future IIASA question-answering system.

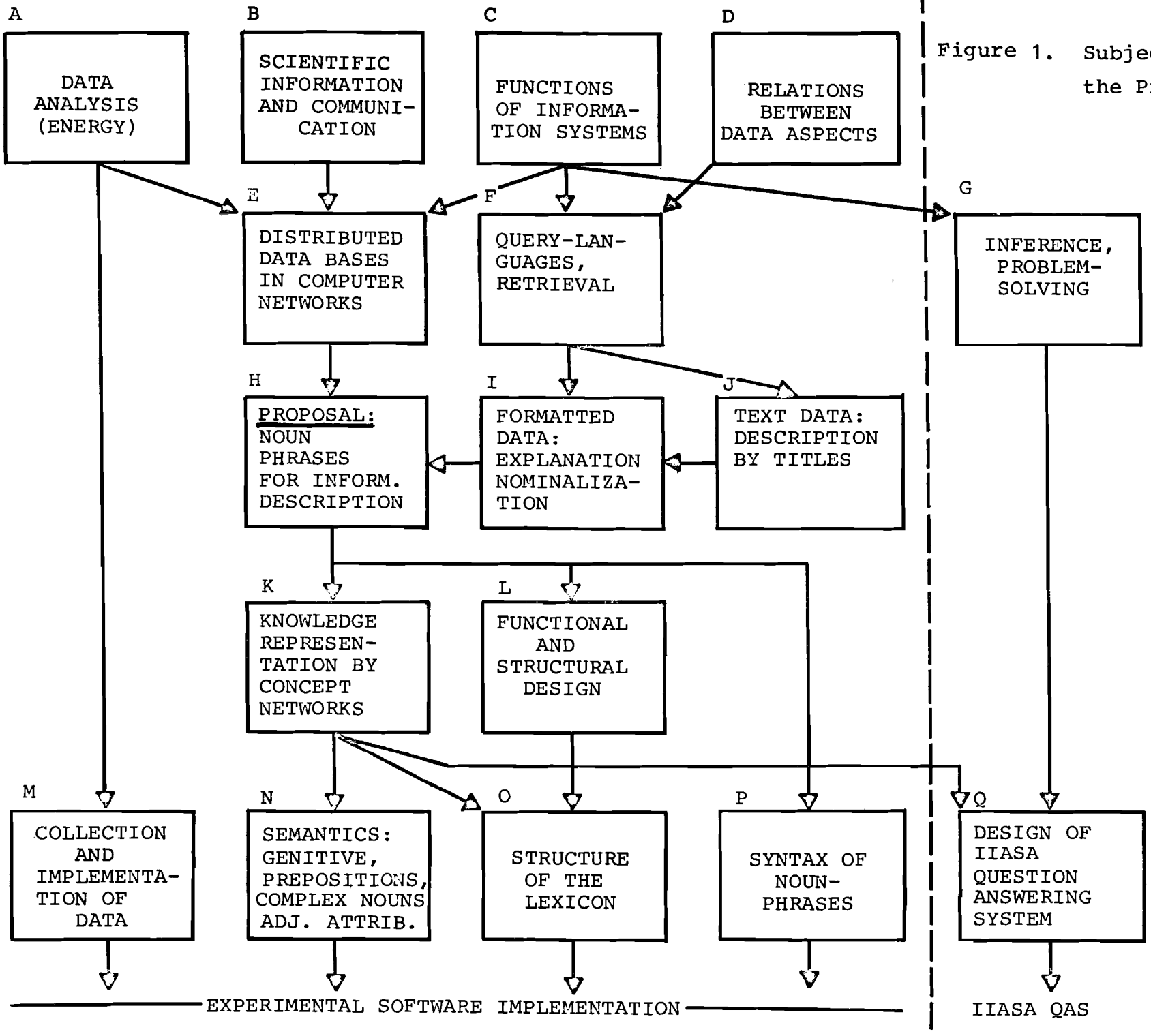

Figure 1. Subjects Related to the Proposed Method (H

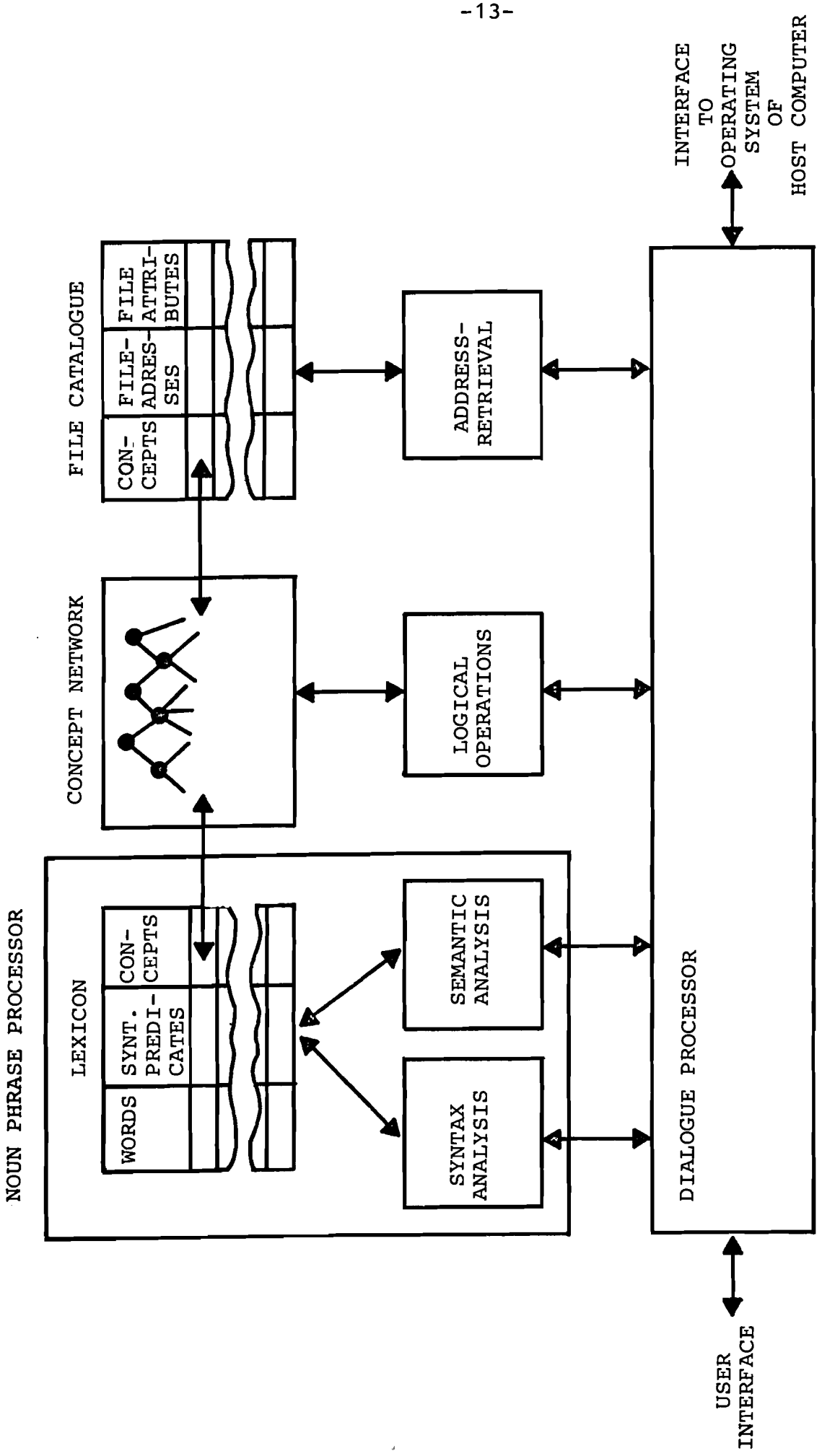

Figure 2. Overall structure of a noun phrase system.