AN INTEGRATED ENVIRONMENT FOR LEXICAL ANALYSES[1]

C. CALIGARIS, A. CAPPELLI, M. N. CATARSI, L. MORETTI
Istituto di Linguistica Computazionale - CNR
Pisa, Italy

0. Introduction

This article describes a project whose aim is to specify tools to be integrated in an environment for lexical analyses. As a result, a prototype of a workbench can be created which provides a user with several modules possessing different functions, in order to approach a text from different viewpoints.

The prototype has been implemented on Macintosh.

Every module can be used autonomously; once integrated in the environment they realize a sort of network of tools interacting with one another.

Let us take a look at the single components of the system.

Firstly, the user has at his disposal tools for the processing of a text in order to obtain indexes, concordances, lemmatizations and various types of statistic analyses.

The prototype also supplies the representational tools for structuring knowledge.

A module containing an ontological reference scheme may be used to show a network of relationships between concepts or to suggest the description of single concepts.

The user is also given a further possibility: access, starting from any node in the ontological network, to a lexical archive indicating all the terms that describe a specific conceptual field, with their relative definitions.

In this way, the system helps in the interactive treatment of texts and makes it possible to analyze and to organize various types of information about a text.

The front-end and certain modules have been implemented by using HyperCard™. This has certain consequences on the interface to the global system, and on the structure and function of any single component.

In a hypermedia framework, a text is no more a sequence of words or sentences, as phenomenologically it appears to a user, but it is a virtual network of the associations implicit in it.

In this way, the substance of a text coincides with the set of its possible readings: its informative content is a magma of fragments whose sense is re-created in the path of each reading.

From a theoretical viewpoint, a hypertext denotes a non-linear writing whose structure is a set of nodes linked by arcs. Nodes contain informative contents, while arcs represent the possible associations between different informative contents, in accordance with the logic of the hypertext itself.

To sum up, the organization of the different knowledge sources within the system facilitates the behaviour of a human operator working on a text from different viewpoints by using the computational metaphor of hypertexts as a means of presentation of data: he can consult a library of electronic books, generate and consult lexical archives and indexes of frequencies, and contextualize words representing the knowledge of a text, while using knowledge sources of different types as a control and a guide. The global architecture of the system is shown in figure 1.

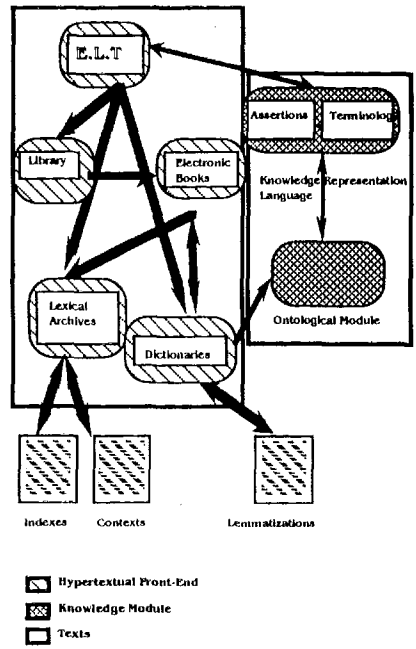

Figure 1

1. Lexical Treatment of Texts

The user has at his disposal certain tools by which he can build and consult several sources, each of which constitutes a sub-environment with its own specific tools. In particular, a library offers a set of texts to be treated by using a set of lexicographic tools (Elaborazione Lessicale Testi) (Moretti, 1991).

[1] This work was partially supported by Progetto Finalizzato Sistemi Informatici e Calcolo Parallelo of C.N.R.

The environment 'text' is composed of electronic books, and it allows the user to perform all classical operations of text processing with the text 'on line'. In particular, concordances can be obtained by choosing the length of the context, lists of frequencies or variants can be shown, and lemmatization can be performed interactively by using the lexical archive as a guide.
Hypermedia technology makes it possible to approach the text in several ways, since the fragments of a text can be linked in accordance to a possible reading criterion. In this sense, it is possible to match different critical editions or to follow the text in accordance with linguistic stylistic facts.

2. Knowledge Representation Language

The knowledge representation language is a member of the family of hybrid systems, and is made up of a terminological component and an assertional one, although certain characteristics make it more similar to classical KL-One (Cappelli et al., 1983; Brachman & Schmoltze, 1985; Nebel, 1989).
The terminological part may be used for the definition of generic concepts, representing classes of objects, while the assertional part is used for the definition of individual concepts, representing single objects.
The structures of the terminological part serve to specify the properties of the generic concept that we are defining. The principle of inheritance applies among the concepts of the network. The sub-concept inherits the properties of the superconcept, even if these are not expressly declared. Furthermore it is possible to indicate, by means of other generic concepts, the relationships that exist between the properties of the generic concept that we are defining: these relationships are known as *structural descriptions*.
The syntax of the terminology is shown in the following

<terminology> ::= <generic declarations> ;
<role declarations> ;
<paraindividual declarations> ;
<generic declarations> ::=
(<generic identifier> = <generic>)*
<role declarations> ::= (<role identifier> = <role>)*
<paraindividual declarations> ::=
(<paraindividual identifier> = <paraindividual>)*
<generic> ::= <generic identifier> |
thing |
(**primC** <index>) |
(**and** <generic> <generic>) |
(**or** <generic> <generic>) |
(**all** <role> <generic>) |
(**atleast** <number> <role>) |
(**atmost** <number> <role>)
(**sd** <paraindividual> <generic>)
<role> ::= <role identifier> |
(**primR** <generic> <name>)
<paraindividual> ::= <paraindividual identifier> |
(paraindividual <generic> <name>+)
<generic identifier> ::= stringa di caratteri
<role identifier> ::= stringa di caratteri
<paraindividual identifier> ::= stringa di caratteri
<name> ::= stringa di caratteri

The structures of the assertional part serve to define individual concepts by specifying the values assumed by the properties of the corresponding generic concept.
The language is based on an intensional semantics, formally specified in Mazzeranghi (1991), and its constructors are interpreted on a universe of structured objects. In other words, the denotation of a generic concept is represented by its properties.
It is thus suitable to account for complex processes involving properties of objects which are specific to the linguistic analysis of a text and, in particular, to the structuring of lexical knowledge.
The expressive power of the language has been further increased in order to account for other conceptual facts, such as recursive definitions (*father/mother*) or definitions expressed by procedures (*length*, *addition*, *subtraction*) (Mazzeranghi, 1991).
As an example, the partial definition of the concept *football-team* is shown in the following:

football-team = (and *team*
(all member *football-player*)
(atleast 11 *member*))

that is to say, a football-team is a type of team whose members are football-players who are at least 11. The denotation of football-team is the following:

$$\mathcal{I}(\textit{football-team}) = \mathrm{PROD}(t_1([T_1]_{min_1}^{max_1}),\ldots,t_n([T_n]_{min_n}^{max_n}),\underline{member}([P]_{11}^{nil}))$$

where:

PROD denotes the Cartesian product,

$[A_n]_{min_n}^{max_n}$ denotes the lists of elements belonging to A,

whose length is between min and max
(if max=nil then there is no upper bound to the length of the lists),
member, which is the name of the role member, acts as a type constructor,
$t_1,\ldots,t_n$ are the names of the properties inherited by team,
$T_1,\ldots,T_n$ are the value-restrictions of the properties inherited by team,
$min_1,max_1,\ldots,min_n,max_n$ are the number-restrictions of the properties inherited by team,
P is the denotation of football-player.

The denotation of football-team is graphically represented in figure 2 (where circles represent denotations of generic concepts and squares represent denotations of roles).
The language can be used to interrogate the ontological module, which can give information about both the syntax and the semantics of the definition of a concept, which in turn can be transferred into the body of a programme specified in terms of the language itself.

3. Ontological module

The ontological module serves to guide the user in the acquisition and structuring of knowledge by suggesting

hypotheses about the description of concepts and their possible relationships.

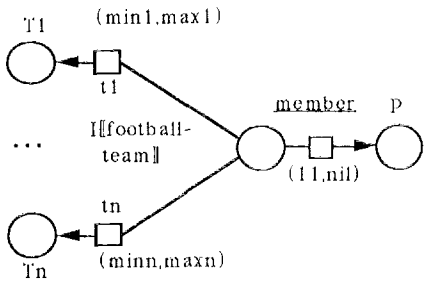

Figure 2

At present, it contains a collection of two hundred concepts organised into the form of a semantic network, with which it is possible to classify a vast portion of reality.
This leads to a taxonomy which serves as an ontological reference guide, suggesting the map of possible relationships between concepts and the most plausible elements of their structure.

3.1. Ontological Theories
Many theories have been proposed about ontological descriptions of concepts (Smith & Medin, 1981).
In the classical model, concepts are described by using necessary and sufficient conditions. In other models, proposed by psychologists, descriptive elements are partitioned into *properties* and *dimensions*, the former being labels assuming binary truth values, while the latter only numerical values. In certain cases, descriptive elements are related to their *definiendum* on the basis of probabilistic parameters or fuzzy logic.
A taxonomy of part-whole relations has also been proposed (Winston et al., 1987; Frederking & Gehrke, 1988) where properties are classified into six types (*component/integral object, member/collection, portion/mass, stuff/object, feature/activity, place/area*) and deductions can be performed according to certain principles which govern the relation between the *definiendum* and its descriptive parts, such as, for instance, transitivity.
Ontologists have proposed global models on the basis of types of concepts and of their properties. The world is then partitioned into *substances* and *accidents* and certain classical notions are defined, such as *genus*, *eidos* , etc. (Simons, 1983).
Körner (1970) defines a categorial framework as a whole where epistemological, logical and ontological aspects are intertwined.
Keil (1989) introduces a division into three general types of concepts: *natural*, *nominal* and *artifact*, and describes criteria for their individuation and description.
Knowledge-based systems using large knowledge bases organized on the basis of ontological principles have been proposed in Artificial Intelligence (Nirenburg & Monarch, 1987; Lenat & Guha, 1990; Onyshkevych & Nirenburg, 1991)
Briefly, efforts have been devoted to finding out criteria for structuring the world by individuating both general types of concepts imposing general constraints on subtypes and types of properties which are pertinent to specific types of concepts. In particular, the logic has been investigated which governs the relationship between a *definiendum* and its *definiens*, even if so far results are far from being definitive.

3.2. Ontological classification
To be epistemologically adequate, an ontology must include i) a taxonomy of concepts with their descriptions, ii) classification and individuation principles associated to concepts.

3.2.1. Taxonomy
As regards the construction of the taxonomy, certain options have been adopted, with the aim of accounting for aspects of the inner nature of concepts and guaranteeing a consistent method of acquisition of knowledge and, consequently, a plausible level of inferential power.
At the top of the taxonomy, as "pure ontological" *summa genera*, the distinction into: *natural* (apple, lion), *nominal* (mayor), and *artifact kinds* (car, chair) has been drawn.
Natural kinds are those existing in nature and are described by natural sciences; they "refer to classes of things that occur in the world independently of human activities" (Keil, 1989 p.25). Artifacts are elements intentionally built to perform a specific function. Nominal kinds are more abstract entities which consist of a description (mayor) which can be applied to instances belonging to different kinds.
This distinction between ontological kinds is relevant in order to structure the universe into chunks of knowledge which are homogeneous from an inferential point of view. Let us introduce an example in order to clarify the structure of the map.
The nominal kind "mayor" can be applied to a person who is a human being - a natural kind -, and it denotes a temporary status of such a human being. To be no longer a mayor does not imply the negation of the existence of an individual, while to negate the essence as a human being does. This classification obviously has effects on the ontological existence of objects (Wiggins, 1980; Keil, 1989). From the point of view of the topological structure of the map, this phenomenon creates a complex chunk of knowledge, as shown in figure 3.
Only a correct disposition of the concepts involved guarantees the right instantiation of individuals, thus allowing true inferences.

3.2.2. Descriptions of concepts
In describing a concept, certain inherent properties are expressed. To be something means sharing certain types of descriptive parts with a set of other concepts. The description of a single concept has to express the properties on the basis of which it can be differentiated and individuated.

In the ontological map, certain types of properties are associated with a concept which, as a whole, constitutes a guiding reference scheme for the description of all its dependent subconcepts.

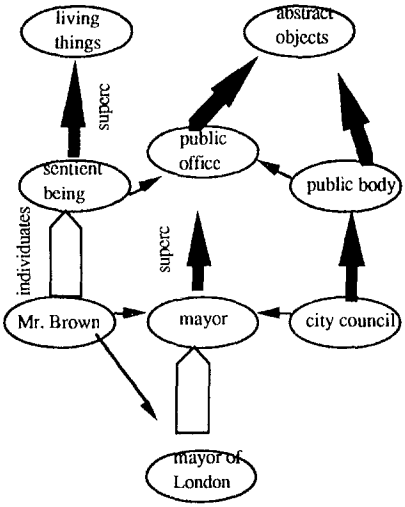

Figure 3

As an example, the concept "container" is associated to a set containing the following types of properties: *content*, *stuff*, *shape*, *function*, and *component*.
It is worth noting that these last are types of properties to which specific values can be associated in the description of each single subconcept or individual.
On the basis of these types, a set of constraints can be specified, such as, for instance:
the property 'Stuff' follows the part-whole taxonomic model as shown in Winston et al. (1987) and Frederking & Gerhke, (1988);
the property 'Content' is organized on the basis of the "place/area" model, where the following transitivity principle is valid: if in(x,y) and in(z,x) then in(z,y);
'Shape' in certain cases refers to the shape of one of the components of a container, which may coincide with the shape of the whole;
'Component' also follows the part-whole model;
'Contextual use' is to be intended as a social and not a functional use, the latter being the specific use of containing something.
To sum up, every type of property is interpreted through a specific set of rules. In this way, a sort of infinite lattice structure is realized where different axiomatic systems of knowledge coexist (see figure 4), each of which has its own interpreter and interacts with the others (Woods, 1990).

3.2.3. Individuation principles
The map has been created by using the knowledge representation language previously described, which supports classification and individuation principles.
The calculus of the properties of a concept makes it possible to build concepts using constructs, such as, for instance, and, or, not, applied to roles of concepts, or to compare concepts, or to classify concepts on the basis of their whole structures.
Furthermore, the knowledge representation language has acquired more "ontological" adequacy by the insertions of global ontological rules concerning the number of properties a concept can possess, such as for instance:
- if two concepts each have only one property and the properties belong to the same type, then the properties cannot have the same value;
- no value can appear more than once in the description of a concept, etc..
These rules act as integrity constraints in the creation of

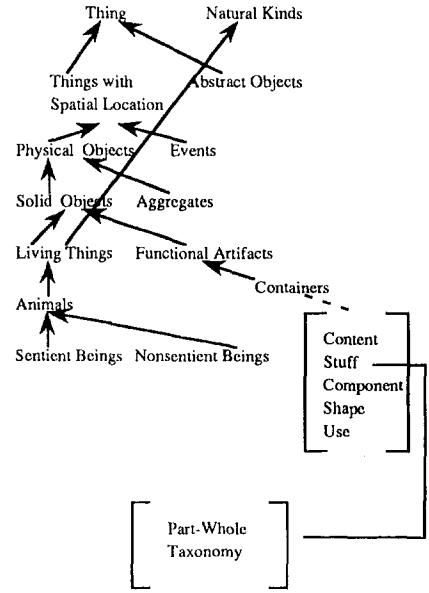

Figure 4

concepts and control both the syntax and the semantics of the knowledge base beeing created. In other words, the result has been achieved of specifying a sort of "style checker" guiding in the manipulation of knowledge.
Furthermore, procedures of any kind can be associated to concepts for their interpretation (Ihooks).
In this way the knowledge representation system realizes *de facto* an object-oriented system.
In our system it is possible to specify an assertional

language which makes it possible to introduce an individual concept into a programming language, like any other data type. For instance, an individual concept is passed to a function as a parameter; once verified that this individual is an instance of a generic concept, or of one of its subconcepts, the function will be executed.

4. Lexical archive

The lexical archive contains a set of lemmas to which with the following information is associated: i) a set of forms with morphological categories; ii) etimology; iii) phonological transcription; iv) definitions in form of text.

Every type of information can be used for retrieving data inside the lexical archive. In order to retrieve conceptual knowledge, which can be extracted from definitions, many possibilities are given. By applying the ELT tools, which make it possible to contextualize portions of texts, the visualisation of the definition of a word can be obtained, or the immediate super-ordinates of the word, or the entire conceptual hierarchy implicit in the whole archive can be retrieved, or parts of definitions in order to find out differences or commonalities can be compared.

4.1. *Linking ontology and lexical items*

Concepts in the ontology are linked to lexical terms of the lexical archive and, vice-versa, from any lexical entry in the archive, the ontological module can be accessed. This is done by using a set of *entry points* which correspond to specific elements in a definition.

Certain concepts of the ontological network are associated with a list of *operators* which map the concept in significant words inside definitions. As an example, the concept of "human being" can be mapped onto the operators 'person', 'who' which realize the concept of "human being" in the lexical archive. Accessing the lexical archive starting from the ontological module, lexical tools are triggered which make use of the list of the operators as searching criteria. In this way the explicit organization of knowledge of the ontological module is virtually linked to the organization which is implicit in the lexical archive.

5. Conclusions

To sum up, we may say that we are trying to create an environment composed of various tools, integrated together, which allows the treatment of a text, and to facilitate the construction and the use of knowledge bases, created from the text itself, for a human operator.

The construction of each single module and its integration within the global system has been carried out taking into account the philosophy of knowledge-based systems and hypertexts.

The latter represent a good tool for the presentation of data, thus allowing 'personal' readings of them: once they are integrated with knowledge-based tools, the global expressive power of the system substantially increases, since data can be abstractly manipulated.

Knowledge representation tools make it possible to build specific theories of the world; by using these tools with the control of an ontological reference schema, any user can realize his own theory of the world in a continuous comparison with a 'standard' organization of knowledge. The specific theory is then able to increase the modalities of searching through data stored in different modules, since it acts as an intelligent interface to data. For instance, it can be used as a filter in searching in the lexical archive, thus overcoming the low degree of expressiveness of its stored information. In this way, a more flexible interaction with any module can be obtained.

References

Brachman R. J., Schmolze J. G., An overview of the KL-ONE Knowledge Representation System, *Cognitive Science* 9 (1985).

Cappelli A., Moretti L., Vinchesi C., KL-Conc: a Language for Interacting with a SI-Nets, in Proceedings of the 8th-IJCAI Conference, Los Altos: Kaufmann, 1983.

Frederking R. E., Gehrke M., Resolving Anaphoric References in a DRT-based Dialogue System, in H.Trost (ed.), *4 Osterreichische Artificial-Intelligence- Tagung*, Springer, 1988, 94-103.

Keil F. C., *Semantic and conceptual development, Cambridge* (Ma.): Harvard University Press, 1979.

Keil F. C., *Concepts, Kinds, and Cognitive Development*, Cambridge: MIT Press, 1989.

Körner S., *Categorial Frameworks*, Oxford: Blackwell, 1970

Lenat D. B., Guha R. V., *Building Large Knowledge-Based Systems, Representation and Inference in the Cyc Project*, Reading (Ma.): Addison-Wesley, 1990.

Mazzeranghi D., Una Semantica Intensionale per un Linguaggio di Rappresentazione della Conoscenza, ILC-KRS-1991-3, Pisa: Ist. di Linguistica Computazionale, 1991.

Moretti L., Text Processing in un Ambiente Ipertestuale, in Atti del Corso Seminariale "Nuove Tecnologie e Beni Culturali" a cura dell'Accademia di Studi Mediterranei di Agrigento, 1991.

Nebel B., *Reasoning and Revision in Hybrid Representation Systems*, Berlin, 1990.

Nirenburg S., Monarch I., The role of Ontology in Concept Acquisition for Knowledge-Based Systems, Carnegie- Mellon University, Pittsburgh, PA, 1987.

Onyshkevych B. A., Nirenburg S., Lexicon, Ontology and Text Meaning, in J. Pustejovsky & S. Bergler (eds.), *Lexical Semantics and Knowledge Representation*, Berkeley (Ca.), 1991.

Simons P., A Lesniewskian Language for the Nominalistic Theory of Substance and Accident, *Topoi* 2 (1983), 99-109.

Smith E. E., Medin D. L., *Categories and Concepts*, Cambridge (Ma.): Harvard Univ. Press, 1981.

Wiggins D., *Sameness and Substance*, Oxford: Basil Blackwell, 1980.

Winston M. E., Chaffin R., Herrmann D., A Taxonomy of Part-Whole Relations, *Cognitive Science* 11 (1987), 417-444.

Woods W. A., Understanding Subsumption and Taxonomy: A Framework for Progress, TR-19-90, Harvard Univ. Center for Research in Computing Technology, Aiken Computation Laboratory, Cambridge (Ma.), 1990.