NICOLETTA CALZOLARI - LAURA PECCHIA - ANTONIO ZAMPOLLI *

WORKING ON THE ITALIAN MACHINE DICTIONARY: A SEMANTIC APPROACH

1. GENERAL FRAMEWORK

1.1. *Foreword.*

The work described by the two co-authors of this article is presented with a double objective: apart from giving specific details on a particular project they also wished to provide a concrete example of the type of research which has been made possible by the Italian Machine Dictionary (DMI).

The DMI is, in fact, one of the principal projects of the Linguistics Division (DL) of CNUCE. Other articles in the first volume of the Proceedings also refer to the DMI.[1] In this introduction I intend to indicate briefly how the DMI project, and, in particular, how the research described in the article has been inserted into the framework of the whole complex of activities of the DL and into our general conception of linguistic data processing (LDP).

As I have already stated in my introduction to these Proceedings,[2] it is my conviction that, at this moment, special attention should be taken in order to promote, both on the theoretical and on the practical level, systematic and ordered interaction among the many different LPD activities. In particular, this cooperation should be realized between those activities which focus on the construction of theoretical models and those focussing on the processing of large corpora of linguistic data. The activity of the DL, especially in recent years, has been increasingly directed towards this goal.

---

* A. Zampolli is the author of Part. 1., N. Calzolari and L. Pecchia are the authors of Part 2.

[1] See Vol. I, 1, pp. 257-262 and 297-301.

[2] See Vol. I, 1, pp. xx-xxi.

1.2. *Activities of the Linguistics Division* (DL).

For approximately 10 years all, or almost all, of the research projects in the different fields of the linguistic data processing in Italy have been worked out with the collaboration of the DL in the computational side of their work.[3]

In the field of lexicography, large corpora of texts have been processed in order to produce the lexical archives necessary to construct extensive historical language dictionaries (see, for example, the *Tesoro della lingua italiana delle origini* of the *Accademia della Crusca*), or dictionaries of " languages for special purposes " (e.g. the *Dizionario Giuridico* of the *Istituto per la Documentazione Giuridica*).[4]

In both modern and classical philological research, the computer is now used with increasing frequency in Italy in order to automate the customary and traditionally time consuming task of indexing texts and producing concordances from them (e.g. the project for the analysis of the corpus of *Grammatici Latini*, ed. Keil),[5] and also for a number of more specific, complex operations, such as the automatic comparison of different editions of the same text (e.g. the project for the ' contrastive concordances ' of *Orlando Furioso* of L. Ariosto).[6]

Literary criticism and the history of literature are also beginning to make use of similar procedures, employing, in particular, statistical

---

[3] For a more detailed description and the relative bibliography see ZAMPOLLI, 1973a, 1973b, 1977a. It is necessary to emphasize an important consequence of this fact. Firstly, almost all the projects underway in Italy in this sector adopt the standards introduced by the DL. In addition, an automatic library containing over 5000 texts in more than 20 languages has been established. This archive may be processed with general-purpose standardized programs because all the texts have been stored using the same scientific and technical criteria. Thus it is possible to perform some linguistic research operations which would otherwise be impossible. For example, one of our projects aims at constructing a new model of the quantitative aspects of the language, on the basis of the data provided by this archive. The earlier models have been falsified by the new quantitative data produced by the increasing number of text-processing projects underway in different countries. As a first step, we aim at identifying those linguistic facts which have a stable frequency in the texts of a language, those which have a frequency which is stable only within certain subsets of a language (literary genres, single authors, particular themes, etc.), those whose frequency does not show appreciable regularity. In a second stage, an attempt will be made to construct and verify quantitative models to describe the regularities actually found and to identify the contextual factors connected with such regularities.

[4] See A. DURO (1973), C. CIAMPI (1973) and F. DIMITRESCU (1973).

[5] See GRILLI and others (1978).

[6] See SEGRE-ZAMPOLLI (1974).

processing as an auxiliary tool in the study of the style of individual authors, schools, or literary genres.[7] Linguistic statistics is also adopted in psycho-linguistic studies, for example to " measure " the linguistic alterations introduced by certain nosological categories.[8]

A combination of statistical processing and algorithms of the " pattern recognition " type are used in a heuristic way on traditional oral texts to identify clauses, formulae, and, in general, the various elements of the popular repertory.[9]

In all the above quoted types of projects, the electronic data processing essentially aims at organizing, in computer storage or in printed form, all the linguistic units of a certain level (words, syntagms, syntactical structures, etc.) occurring in a text, in order to enable a more efficient, rapid and economic retrieval of them. In other words, the processing basically consists in the following types of operations: to input, store, manipulate texts of different kinds (which may be considered as facts of *la parole*); to recognize and explicitly represent in the text the occurrence of linguistic units (phonemes, lemmas, affixes, syntagms, syntactical types, etc.: these units may be considered to be at the level of *la langue*); to execute some canonical operations (retrieval, ordering, counting, comparing, etc.) on such units, in batch or conversational form.

We also cooperate with some projects in the field of full-text information retrieval, which also uses lexicographical-type processing for documentary purposes, mainly on juridical and historical texts.

All the above mentioned activities make use of closely inter-related procedures which the DL has developed and put into operation with the collaboration of various Italian Universities and CNR Institutes.

More exactly, it could be said that the DL has realized, or is in the process of realizing, a certain number of basic processing " components " and that each of the procedures so far developed consists in the concatenation of some of these components.

The functions of each of these components are well-known within the LDP environment: the acquisition of texts in machine readable form; the production of the typical results of lexical analysis (different types of concordances, context-cards, etc.); the representation of the large variety of characters typical of the LDP; morphological analyses

---

[7] See A. ZAMPOLLI (1975).

[8] See CASTROGIOVANNI (1973).

[9] See CIRESE (1973).

and consultation of Machine Dictionaries (DMs); syntactical parsers; phonological transcription; etc.

I feel that the following three characteristics of these components should be emphasized.

a) They are conceived so as to be, as far as possible, generalized (i.e. applicable to all the texts processed at the DL, whatsoever their nature, language, or the purpose of the processing),[10] flexible (the user can activate, within the set of rules which constitute the " algorithmic linguistic knowledge " of the program, those rules which best respond to his particular needs),[11] and modular (the components must be inter-compatible and open to the inclusion of any eventual new components: the inter-compatibility is ensured by exchange-interfaces between the various components; these interfaces consist in a formalism which provides structures, organizations and codes for the representation of linguistic units both at the text and at the linguistic system level).

b) These components may be used – at least in principle – with the same basic functions both in lexicographical-philological type applications and in translation, documentation, question-answering, etc.[12]

---

[10] For example, the component proposed for the acquisition of texts in machine readable form performs the following functions: accepts, as input, texts in any natural language (as long as they can be transcribed alphabetically) of any period, or literary form or genre (scientific texts, recorded dialogs, protocols, interviews, novels, inventories, etc.); stores the texts on auxiliary memory; produces listings which reproduce the text as near as possible to its original form; supplies text editing facilities for checking and correction of eventual errors. At the basis of this component is an encoding system which is designed to represent all the different graphemes and graphic features which can appear in printed texts or can be inserted in them in the preediting stage.

[11] For example, the context of a word can be constructed and delimited by activating and ordering diversely a suitably chosen subset of the available rules from a general contextualisation algorithm (see ZAMPOLLI, 1971): to coincide the context with a structural unit (verse, strophe, etc.); to delimit the context exclusively on the basis of the punctuation immediately preceding or following it; to assign a specific portion of the syntactic structure as context, etc.

[12] In particular, at the beginning of the 60s, attempts were made to classify the different systems for LDP according to the so-called ' depth-parameter ' of the linguistic level of operation. Such classifications selected a certain " depth " level along this parameter, and drew in correspondence to this level the demarcation line between the uses of the computer in linguistics which merit the name computational linguistics (CL) and those which do not.

Our viewpoint is different. All computational systems functioning for linguistic researches or which operate on linguistic data belong to the CL. Besides, at least in principle, the majority of those systems, independently from the fact that they are considered either below or above an established demarcation line, have a number of components

c) They are, as far as possible, the result of studies which are both research and operationally oriented.

1.3. *The Italian Machine Dictionary* (DMI).

The DMI has also been realized in accordance with these criteria. It has been conceived and is used as a means for semi-automatic lemmatisation, i.e. for the recognition of the occurrences of the various units of the Italian lexical system within a text. It is used in lexicographical, statistical, philological text processing and is utilized in full-text information retrieval systems in order to identify in the documents all the different forms which belong to the same lemma of a specific form appearing in the " question " asked by the user. It will be used to associate to the words from a text the information requested by syntactical and semantical parsers (morpho-syntactical categories, syntactical " valences ", semantical markers, etc.).[13]

In the lemmatisation stage, the DMI can be adapted by the user to obtain lexical analyses at different levels of complexity. We think of the definitions of a lexical unit (lemma) as a set of pertinent features (morphological. syntactical, graphical, etc.). Different inflected forms

---

in common. For example, a procedure for lexical analysis necessitates: the acquisition in machine readable form and the computer printing of a variety of texts and graphemes; a morphological analyzer and the consultation of a DM for semiautomatic lemmatization; syntactic and semantic parsers for homograph disambiguation. An automatic translation system requires all these features (in addition to the transfer and generation components).

[13] Of course, we have considered whether it would be possible and convenient to compile a DM without having first defined in detail the components which will use the linguistic information contained in it. As an example, let us consider the choice and the formalization of grammatical information (morpho-syntactical categories, valences, specification of possible constructs, etc.) to be coded in the dictionary as " input " of a syntactic parser. Obviously, this depends on the grammatical model and the strategy used by the parser. This does not necessarily mean, however, that once a DM has been compiled with specifically chosen grammatical information, it is necessary to substitute the grammatical part of the DM if the grammatical model should change. Although there are a number of different opinions on this important point, our experience has suggested that, eventually, it will be necessary to extend and complete the already existing information rather than substituting it. In the majority of cases, independently of the definition of their theoretical status, the basic syntactical properties of a lexical unit may be formulated in a neutral way with respect to the model and systems which use them. This affirmation can be largely verified, at least for models within the same " scientific paradigm ", e.g. the generative-transformational ones. Nevertheless, there is perhaps enough evidence to assert that the basic information, at the morpho-syntactical level, is still, to a large extent, valid, even when considering other paradigms such as the so-called " artificial intelligence paradigm ".

of a text are considered to belong to the same lemma if and only if they have in common all the pertinent features which identify a lemma, distinguishing it from all other lemmas. We have constructed an inventory of features which may be used in the definition of a lexical unit. Such an inventory is based upon a survey of the features used both in lexicographic practice and in linguistic theories. Each entry of the DMI is associated with the set of all the possible features of the inventory which may be used in its definition. The user is allowed to disactivate those features which he does not wish to utilize: for example, the differences between nominal and verbal use of participles or those between adjectival pronouns and pronouns, etc. Obviously, if some distinctions are neutralized, the number of lexical units which constitute the DMI, as defined by the user, and very often the number of possible homographs, are reduced. In other words, if we consider the DMI a concrete representation of the Italian lexical system, in which the lexical units are defined using all the features proposed by the different lexicological and lexicographical traditions, the user can modify the structure of this system and the inventory of its lexical units in accordance with his specific linguistic requirements (Zampolli, 1973a). In this perspective, the DMI is used not only as a tool for text processing but also as an object of studies and research in itself.

While in studies at the level of *la parole* the object is given immediately for the LDP in the form of corpora of texts, the object in studies on *la langue* must be specifically constructed. An example which can be given is the first step in a research on the functional load of the phonological oppositions of a phonematic system. This step consists in the inventory of the minimal pairs existing in the lexicon for each opposition and therefore it presupposes the existence of an inventory of all the different forms of the studied language in phonological transcription. The burden of creating an inventory of this type and dimension, and the complexity of the operations required in order to discover and count all the minimal pairs are such that all those tasks are impossible without a computer. Another example could be a study on the " rendement " of the different suffixes, which requires an inventory of all the words in which each suffix appears.

In order to make research work of this type possible, the DMI has been conceived diversely from most of the other DMS in existence. These have usually been realized exclusively as components in translation procedures, information retrieval systems, etc. Such DMS, almost always, include only a limited number of lexical items.

The DMI has a structure and dimensions that allow us to consider it as an exhaustive, automatically processable representation of the lexical component of the Italian linguistic system. The DMI is, therefore, intended as an instrument for research studies at the level of *la langue* where exhaustive inventories, data and observations are necessary.

1.4. *Theoretical background.*

The research project described below by N. Calzolari and L. Pecchia is an example of how the DMI can be used in this direction.

The actual situation of linguistic theory is that of constant change and development. Not only are the traditional models being continuously modified but some researchers affirm also that the debate is now between theories which belong to different scientific " paradigms ". Examples usually quoted are the number of different generative-transformational schools (interpretive semantics, generative semantics, etc.), relational grammar, cognitive semantics. In this situation, some researchers present the following alternatives: whether the scope of the research work conducted in LDP must, of necessity, be directed towards a specific linguistic theory, or whether LDP can produce results which can be utilized by different linguistic schools.

For the sake of simplicity we will examine certain examples from the syntax field. A clear example of LDP activity directed at a specific linguistic theory is, in my opinion, offered by the so-called ' grammar testers ', i.e. those computational systems which apply a lexicon and a grammar for automatic sentence generation.[14]

These systems, at least in the intention of their creators, constitute a concrete and precise specification at the computational level of a determined linguistic theory; the grammar is considered as a program used to produce sentences; the algorithms which interpret the rules are considered as a part of the meta-theory; the production of concrete sentences serves to verify the coherence of the rules, the completeness and lack of contradiction of the formal apparatus and to indicate, practically, the extension of the subset of language generated by the grammar.

Evidently, these systems are intentionally strictly connected with

---

[14] This is not the place to enter into a discussion on the complex and well-known problem of the relation and the differences between " generation " as an abstract calculus of all the possible grammatical objects and the automatic " production " of concrete sentences.

the corresponding linguistic theory, constituting, it could be said, the computational " transcription " of it. The rapid evolution of the theories, the models, the formal apparatus require a continual updating of the corresponding computational systems, which does not seem very easy to realize, at least in practice. Furthermore, the generative-transformational schools whose theories are usually incorporated in these systems have so far only described isolated regions of the linguistic structure, aiming at verifying the adequacy of descriptive methods rather than at describing coherently and exhaustively a language. As a consequence of this, anyone wishing to use the results of their researches in a computational system would face a set of isolated observations distributed in different regions of a language, not systematically linked to each other, but divided by so far unexplored regions.

On the other hand, however, the analytical methods produced by the generative-transformational theories have revealed a very efficient heuristic power, and have considerably increased the precision and subtlety of the observations. The number of new phenomena that have been revealed has grown notably in the last 20 years.

In front of this situation, the behaviour of LDP researchers may range between two alternatives.

The first position is usually characterized as the rise of a linguistic " computational paradigm ", which is distinct from, if not directly in contrast with, the generative-transformational paradigm, and tends to assume the computational aspect among the principal characteristics of a linguistic theory. The conviction is expressed that the primary " focus " of linguistic research must be shifted from the description of the competence as formal abstract mechanisms towards the simulation-like studies of the processes which underlie the production and the comprehension of the utterances. The " natural language understanding " computational system could constitute a powerful experimental and heuristic tool for the study of the complexity and the constraints of these processes, making it easier to emphasize the mechanisms of interaction between the components which are involved in these processes.

The scantiness of the results obtained so far (some of the devotees of this approach have likened the situation to that of medieval alchemy as opposed to modern chemistry) makes it impossible to formulate even a summary judgement. Nevertheless, it is quite clear that this type of research is limited, and will be probably limited, at least for some time in the future, to the consideration of extremely limited language subsets.

The second position seems to prefer, in the actual situation of linguistic theory, a systematic examination of the data to an immediate construction of a formal global model. Obviously, the use of abstractions or notions (e.g. those of transformation or of componential analysis) whose theoretical state may vary depending on the global evolution of the theory itself, but which have been seen to be experimental devices of extraordinary efficiency in the analysis, is not rejected. The complex formal mechanisms proposed by the generative-transformational school is not implemented into a computational system as a representation of a " language theory " but some of their characteristics (form of the rules, relationship between the rules, etc.) are utilized to store, handle and organize the data accumulated in the inductive moment of the research.

LDP essentially offers two complementary contributions to this approach. Firstly, it supplies techniques which permit the automatic handling of the data. Secondly, LDP studies algorithms which permit the data to be structured conveniently, organizing them so that their regularity, diversity, correlations, etc., can be evidenced without it being necessary to make this organisation dependent on the " a priori " choice of a general global theoretical model.

The inventories of linguistic units recorded in " machine readable form " must be considered within this framework and, in particular, those lexical inventories in which each lexical unit is supplied with an explicit, suitably coded, representation of its linguistic behaviour should be considered.

In addition, the use of a lexical inventory would facilitate the definition of the degree of exhaustivity of the descriptions and the evaluation of the extension of the phenomena studied. (The term ' extension ' must be here understood obviously not as frequency of appearance in texts but as frequency of appearance in the system).[15]

At the same time, it seems that the time has come to systematize and put at the public disposal the linguistic data accumulated in machine

---

[15] The information is often represented by binary matrixes in which a line corresponds to a lexical unit, a column to a specific linguistic property. This organization obviously facilitates the identification of identical or similar configurations, the verification of the coherence between the contents of the interelated columns, etc. (see JOSSELSON, 1969). The work of M. GROSS (1975) and his group in the construction of a grammatical lexicon of French certainly constitutes the most important example. Furthermore, the role which the lexicon and its description have assumed within the most recent developments of the generative-transformational school (Bresnan, etc.) should not be neglected.

readable form (texts, dictionaries, descriptions, rules, etc.) and the computational tools (software packages, integrated systems, mid level and high level languages for LDP, etc.) produced in different institutes of different countries in different ways, but on the basis of similar methodological assumptions and of a general common sum of knowledge.

It is within this framework, and not only for applied and operational purposes, that since 1968 (ZAMPOLLI, 1968), I have promoted the construction of the DMI as one of the principle projects of the newly constituted DL.

The project described in the following pages by N. Calzolari and L. Pecchia is an original development in the field of semantics along these general planning lines.

2. TOWARDS A FORMALIZATION OF LEXICAL DEFINITIONS

2.1. *Preliminary steps.*

This part of the article describes an attempt to formalize all the noun-definitions in the Italian Machine Dictionary (DMI). The definitions recorded in the DMI were taken from the *Zingarelli* Dictionary (1970) after having undergone a first process of normalization and shortening. Part of the normalization process was to classify the Zingarelli definitions into 9 different types and to mark each of these with a particular code.

The main types of definitions are:

1) the relational (coded as 1), which is composed of *a*) a fixed part representing a function, and *b*) a variable part, the basis;

2) the synonymous (coded as 2), which is made up of one or more single words which are referred to for an explanation of the meaning considered;

3) the one per ' genus et differentia ' (coded as 3), which is made up of *a*) a fixed word considered as a classifier (the ' generic part ' of the definition), and *b*) a descriptive or predicative phrase of the classifier (the ' specific part ' of the definition).

The framework of our research is typical of componential analysis, according to which even that which appears to be " a list of basic irregularities " (BLOOMFIELD, 1933, p. 162), i.e. the lexicon, could become a well-structured and therefore formalizable set, in other words,

a system. We were first given the idea for the analysis by the theory of componential analysis, but we have attempted to expand its field of application which, up to now, has been dedicated only to well structured sets, as shown in the work of componential anthropologists, (the domains of words of kinship), or to lemmas isolated from the rest of the lexicon (the well known example of Katz: ' bachelor '). Our intention has been that of extending the application of this theory to all nouns of the Italian lexicon. We are helped in this by the great quantity of material at our disposal in the DMI. As we are well aware of the limitations of componential analysis, we have used it only as a tool, not as an end, in achieving our purpose.

From the entire corpus of lemmas and definitions in the DMI we have excluded those lemmas and definitions which are marked as archaic or rare. We have analyzed, up to now, all those definitions classed under codes 3 and 5, i.e. those with one generic and one specific part. These are the most numerous groups of definitions. After this selection had been made, the total number of lexical items on which we are actually working is 28,873, among which 20,453 are monosemic and 8,420 polysemic; the total amount of their definitions is 44,051. We have worked on this corpus of lemmas and definitions using programs and checks of different kinds, working in two main directions which will be discussed later in more detail. Firstly, we have extracted a considerable number of markers which would be assigned to the highest possible number of lemmas. Secondly, we have started an analysis of prepositions, of prepositional groups and of other syntagms which can be considered as grammatical in a very generic sense. These syntagms have been chosen because they satisfy, simultaneously, the following two criteria:

a) that of occurring with a high frequency in the definitions;

b) that of showing well defined semantic relations existing between noun and noun, or between verb and noun, or between noun and proposition.

2.2. *Markers.*

In the first phase of our work, the aim was to extract a certain number of ' markers ', starting mainly from the definitions; in other words, working in an inductive way. We obtained the first basic working elements from a control of the frequency-list of the forms found

in the corpus of noun-definitions. This list helped us to make a first purely provisional inventory of lemmas which might be used as ' markers '.

Then, by looking up the concordances of these definitions, we were able to test the validity of these basic elements. In fact we have ascertained that the most frequent lemmas in the set of noun-definitions, (i.e., the lexical entries which will be most probably used as ' semantic markers ') almost always appear in the context in a generic sense and in the first position, only occasionally assuming a specific sense in different positions. The fact that, as expected, with the exclusion of syntactic words such as prepositions, conjunctions, articles, etc., the highest frequency-indexes pertain to the grammatical category of nouns has also been relevant.

We shall use the name ' markers ' to refer to these most frequent lemmas: but there is a difference between our ' markers ' and the markers referred to in componential analysis; although our markers function as markers usually do, i.e., they describe a meaning or part of it, they remain essentially lemmas. It is thus not necessary to use a metalanguage different from the language which is being described; the elements of the lexicon can be given a metalinguistic function.

These markers have been grouped into lists on the basis of different semantic criteria such as synonymy, antonymy, etc. We have also made a distinction between markers behaving as one-place predicates and markers behaving as two- or *n*-place predicates.

A first group of 450 semantic markers was extracted and matched by a program with the generic part of all the definitions. We have verified that 22,146 definitions out of 47,291 were covered, in their generic part, by these markers. This first part of the work is described in more detail in CALZOLARI, MORETTI (1976).

In the prosecution of the work, through further additions or substitutions of semantic markers which were either provided by literature on this subject, or resulted from our intuition, or by other successive analyses on the corpus, we have covered 40,135 definitions with 407 markers.

We have ascertained that, in almost every case, the generic part of the definitions of the DMI (and therefore of the Zingarelli) gives the word whose level is immediately higher with respect to that of the defined lemma (considering a hierarchical classification moving from the more specific to the more general, i.e. from a greater to a smaller intension). This homogeneity in the definitions justifies the validity

of the method we have adopted to refer all the lemmas back to the markers.

In practice, for the lemmas not covered by markers after the first procedure of matching, i.e. for those lemmas which are defined, in their generic part, by words which are too specific to be used as markers, we have established some chains which refer back to more and more generic words until at least one marker is reached. In order to construct these chains we used a procedure to convert all the lemmas into numbers: this seemed to be the simplest way to keep in main storage the great quantity of data we had to work with. Using a program which works on these numbers, we have simulated a path for each lemma. This path starts from the lemma itself, and the program examines the generic part of the definition of the lemma. The program checks if this generic part is included in the list of markers, and in any case examines this generic part itself as a lemma to be defined and looks for its definition; the procedure continues in this way until the generic part of a definition is found to be a marker without any other more generic marker above it. By this procedure, 91% of the definitions have been reconducted to the markers, i.e. 40,135 out of 44,051.

By means of these chains, we have given the noun-dictionary a resemblance to a tree-structure. This tree-structure has been formed using the definitions of the DMI for almost all the lemmas; the hierarchical structure we have given to the markers has, on the contrary, been partly taken from the definitions, and partly imposed by us according to the traditional rules of class inclusion.

* 400	*ACCORDO* 1 accord	→	*ARMONIA* harmony	→	*CONCORDANZA* concordance	→	*RELAZIONE* relation	→	*QUALITÀ* quality
* 200	*ACCORDO* 2 accord	→	*UNIONE* union	→	*ATTO* act				
* 200	*ACCRESCIMENTO* increase	→	*SVILUPPO* development	→	*ATTO* act				
* 800	*BARIBAL*	→	*ORSO* bear	→	*MAMMIFERO* mammal	→	*CLASSE* class	→	*GRUPPO* group →
		→	*INSIEME* set	→	*TOTALITÀ* totality	→	*QUANTITÀ* quantity	→	*ENTITÀ* entity
* 031	*BARIO*	→	*ELEMENTO* element	→	*PARTE* part	→	*PEZZO* piece	→	*PARTE* part
* 051	*DUCA* duke	→	*TITOLO* title	→	*NOME* name	→	*VOCABOLO* item	→	*PAROLA* word →
		→	*TERMINE* term	→	*PAROLA* word				

Fig. 1. *Examples of chains from lemmas to markers.*

In setting these chains (see Fig 1), we discovered that some chains of *definienda* and *definientia* are circular, e.g. *PARTE* is defined in the DMI as *PEZZO*, and *PEZZO* as *PARTE* (see also CALZOLARI, 1977). In the example given in Fig. 1, the asterisk indicates the presence of at least one marker in the chain; the first number indicates the length of the chain; the second the length of the chain if it is circular; the third the distance between the two identical lemmas in the circular chain.

It has been possible, using these chains, to assemble the entire dictionary around some essential cores of more inclusive meanings. These cores are the tops of the trees, and from there thick branches lead off to the more particular and specific levels of the lexicon. The final data concerning the number and depth of the chains are shown in Table 1.

TABLE 1.

Number of definitions: 44,051

Number of definitions which lead to a marker: 40,135

length	chains	circular chains
1	6960	474
2	4495	2734
3	7960	3576
4	5375	3659
5	2153	2029
6	1165	1601
7	422	576
8	181	306
9	42	244
10	7	83
11	0	6
12	0	3
total	28760	15291

Moreover, for every marker (see Fig. 2), we have counted the number of times it occurs in all the chains (second column), and the number of times it appears in the chains which stop at the first marker (third column). In both of these cases, we have computed separately the occurrences of the marker at all the levels (1st, 2nd, 3rd, etc.; the 1st column indicates the levels).

	depth	occurrence of the marker	occurrence as 1st marker
ESSERE	1	0	0
(being)	2	93	90
	3	135	43
	4	33	10
	5	4	4
	total	265	147
ANIMALE	1	1	1
(animal)	2	42	40
	3	139	53
	4	11	2
	5	1	1
	total	194	97
MAMMIFERO	1	0	0
(mammal)	2	126	124
	3	241	52
	4	32	8
	5	6	0
	total	405	184

Fig. 2. *Examples of computations on markers' occurrences.*

2.3. *Definition-Structures.*

As far as the structure of the definitions is concerned, we wanted to start the analysis again from the definitions themselves (not trying to test some preconceived structures), with a careful checking of the corpus of definitions.

We have extracted prepositions, and prepositional or grammatical syntagms, on the basis of a frequency-criterion, placing together under the term 'locution' or 'prepositional syntagm' (even if this term is not a very exact one) expressions of this kind:

a forma di (in the form of);
dal colore (of colour);
provvisto di (provided by);
munito di (furnished with);
in contrasto con (in opposition to);
consistente in (consisting in);
simile a (similar to);
originario di (originating from);
che serve per (which serves for/as); etc.

These phrases which we will call, arbitrarily, ' prepositional syntagms ' have been divided into various categories. This subdivision was made possible through an introspective examination of the associations of analogous meanings. The criterion was the individualization of the recurring semantic functions which have a similar meaning, even though these functions have been expressed lexically and/or syntactically in a completely different way.

One example of such grouped functions is the category *SCOPO* (aim), for which we have individualized the following set of lexicalizations (when necessary, with relative flection):

tendente a (tending to);
diretto a (aimed at);
volto a (directed to);
con lo scopo di (with the purpose of);
a scopo di (for the purpose of);
che ha lo scopo di (which has the purpose of);
che mira a (which aims at);
chi mira a (who aims at);
mirante a (aiming at);
rivolto a (turned to);
per conseguimento di (for achieving); etc.

We have grouped these lists of prepositions and prepositional syntagms into files on the basis of their affinity of meaning. This has been possible through the analysis of the functions and of the different possibilities of their expression, following inductive and deductive methods.

The validity of these associations of meaning, made intuitively, was afterwards verified empirically: various procedures for the extraction of the definitions in which each function appears, provided the material to be analyzed for this checking. For instance, in the analysis of various relations, such as those we called *ATTITUDINE* (aptitude), *COLORE* (colour), *FORMA* (form), *CONTENUTO* (content), *ORIGINE* (origin), *SCOPO* (aim), *USO* (use), *SOMIGLIANZA* (similarity), *COMPOSTO* (composed of), *MUNITO* (furnished with), *RELATIVO A* (relative to), the check of all the definitions in which elements of the corresponding lists appear has shown the validity (about 80-90%) of our groupings made on the basis of our intuition. In addition, from this careful examination of different groups of definitions, we obtained some data which made it possible for us to formulate some interesting considerations.

We have observed, for instance, that the definitional structure based on the relation *ATTITUDINE* (aptitude) has a quantitatively high homogeneity of application with respect to the lemmas in whose definitions the relation is used. In fact, in 50% of the definitions in which this relation appears, it is applied to lemmas whose generic part, i.e. whose main semantic marker, is included in the list of homogeneous markers we have called *STRUMENTO* (instrument) (see Fig. 3). Examples of the recurring generic parts with a high frequency are: *Meccanismo* (mechanism); *Organo* (organ); *Congegno* (contrivance); *Apparato* (apparatus); *Attrezzo* (implement); *Strumento* (instrument); *Arnese* (tool); *Dispositivo* (device); *Apparecchiatura* (apparatus); *Macchina* (machine); *Attrezzatura* (equipment); *Apparecchio* (apparatus).

ACCIARINO	= *Dispositivo atto a determinare l'accensione*
Flint-lock	Tool apt to cause accension
ARCHIPENDOLO	= *Strumento atto a rendere orizzontale una retta*
Plumb-line	Instrument apt to make a straight line horizontal
CARICATORE	= *Attrezzatura atta al carico e allo scarico di materiali*
Loader	Machinery apt to load and unload materials
SPEZZATRICE	= *Macchina del panificio atta a tagliare la pasta in pezzi*
Cutter	Machine of the bakery apt to cut the dough into pieces

Fig. 3. *Examples in which the function ATTITUDINE (aptitude) selects the particular marker STRUMENTO (instrument).*

It is interesting to point out the way in which a certain definition structure can be frequently associated to a certain kind of marker. Other definition structures linked to other functions can make it possible to delimit, within the lexicon, sufficiently homogeneous semantic fields. Since these associations between markers and functions occur in several groups of definitions, we think that this correspondence ' marker-relation ' is not random, but is established for semantic reasons of affinity at a syntagmatic level. It seems possible for us to assert, at this point, that some markers effect a preferential selection toward certain types of defining relations rather than others, and vice versa. If this hypothesis is tested extensively on the lexicon, it can help in reaching a formalization of the semantic information which is in the DMI.

We think that a more complete formalization, in comparison to that obtained by the simple hierarchical organization of the markers, can be achieved by also identifying the other kinds of relations which are different from the hierarchical one. Functions such as those described above will allow:

a) the linking of markers: for example, the pertinence relation *PARTE* (part) makes it possible to link the markers *PERSONA* (person), *UOMO* (man), *DONNA* (woman), with a set of markers such as *MANO* (hand), *CAPELLI* (hair), *BOCCA* (mouth), *TESTA* (head), *CAPO* (head), etc.; and/or

b) the joining of the generic to the specific part of the definitions, for example in the definition of

ACCHIAPPAMOSCHE = *Strumento atto a catturare mosche*
(Fly-swatter = instrument apt to catch flies)

the function *SCOPO* (aim), in its lexicalization *ATTO A* (apt to), links the marker *STRUMENTO* (instrument) to its specification.

For the final structure of the definitions, we think that the markers can either be considered as *n*-place predicates joined to their arguments by these various types of functions, or as nodes of a semantic network linked to the specific part of the definitions, i.e. the other nodes, by arcs which express these various types of functions.

Such relations can be used as the starting point in the study of the use of prepositions and prepositional syntagms in the Italian language and, particularly, in the language of vocabulary definitions.

Unifying these functions is also of great help in structuralizing the definitions, at a higher level of formalization, assisting greatly in the extraction of all the data linked by the same function.

We have also noticed that some types of sentence-structure occur more frequently in the definitions. Besides considering the functions in isolation, we have been working on a quantitative examination of the various possible matchings of these functions among themselves; this has been done with the aim of also identifying the kinds of sentence-structures more frequently used by lexicographers in the compilation of dictionaries. A practical goal for us is to work further towards the unifying of the definitions, by leading them back, as far as possible, to the more frequent and common structures.

2.4. *Perspectives.*

Our research had a number of different aims but was principally directed towards the lexicographic aspect. This aspect consists in an attempt to analyze the defining method adopted by Italian lexicographic tradition as shown by the *Zingarelli*. This analysis has been developed in two different stages:

1) An analysis of the terminology used in the definitions, through the enucleation of markers. We have seen that, among the most frequent lemmas in the definitions (i.e. among those words whose extension is greater or, in other words, whose intension is smaller), those words considered as markers by literature on this subject appear.

2) A check of the definitions considered from the point of view of their structure. This emphasized the very high frequency of certain types of functional syntagms as being more suitable in compiling definitions. It will be interesting to have a comparative examination with dictionaries of other languages.

The semantic aspect is very closely related to the lexicographic aspect of this study. Our aim was to give a hierarchical type of organization, even if provisional, to the large set of Italian nouns at our disposal. In doing so, we have taken what in our opinion is the first step towards a decomposition of a meaning into distinctive markers, i.e. the attribution as main semantic marker of the lemma which is at an immediately higher level in a hierarchical scale. Many hierarchical scales can be individualized in the lexicon, or more precisely among the meanings of the lexical items.

We have also begun, through the study of prepositional functions, the second step in the decomposition of a meaning into markers: the linking of markers with other markers, the individualization of the different kinds of relations which exist among markers, and of those relations which exist between primary and secondary markers expressed respectively by the generic and the specific part of the definitions.

There is also an important practical aspect of this work: that of making the definitions of the DMI more uniform from a semantic point of view. This is achieved by indicating the semantic uniformities which are latent under the different lexicalizations of the same markers or of identical relations, and by reducing these diversities of lexical forms to one single symbol reflecting their uniformity. This will make the looking up of the DMI easier.

This work should also be of relevance, at a future date, in connection with an analysis of the verb which takes into consideration the above mentioned analyses of the noun at a level of selectional restrictions at first, and, later, extends these analyses to the level of " knowledge of the world ". Thus, we feel that our work can provide a first step for a future utilization of the DMI in syntactic and semantic analyses of the Italian language.

REFERENCES

M. ALINEI, *La struttura del lessico*, Il Mulino, Bologna, 1974.

M. BIERWISCH, *On certain problems of semantic representations*, in « Foundations of Language », V (1969), pp. 153-184.

L. BLOOMFIELD, *Language*, New York, 1933.

N. CALZOLARI, *An empirical approach to circularity in dictionary definitions*, in « Cahiers de Lexicologie », XXXI (1977) 2.

N. CALZOLARI, L. MORETTI, *A method for a normalization and a possible algorithmic treatment of definitions in the Italian Dictionary*, presented at ICCL, 6th International Conference on Computational Linguistics (COLING '76), Ottawa, 1976.

P. CASTROGIOVANNI, A. TELARA, *Primi risultati di un'analisi statistica morfologica e lessicale delle risposte al test di Rorschach nella prospettiva di uno studio dei rapporti tra psicologia e linguaggio*, in A. ZAMPOLLI (ed.), 1973a, pp. 307-324.

E. CHARNIAK, Y. WILKS (eds.), *Computational Semantics*, North-Holland, Amsterdam, 1976.

C. CIAMPI, *Les projets de recherche automatique des informations juridiques dans l'Institut pour la documentation juridique du Conseil National des Recherches*, in A. ZAMPOLLI (ed.), 1973a, pp. 249-268.

A. M. CIRESE, *Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens*, in A. ZAMPOLLI (ed.), 1973a, pp. 209-231.

P. COLE, J. SADOCK (eds.), *Syntax and Semantics: Grammatical Relations*, Academic Press, New York, 1977.

F. DIMITRESCU, *Projet d'un dictionnaire de la langue roumaine du XVI siècle*, in A. ZAMPOLLI (ed.), 1973a, pp. 41-48.

A. DURO, *Élaborations électroniques de textes effectuées par l'Accademia della Crusca, pour la préparation du dictionnaire historique de la langue italienne*, in A. ZAMPOLLI (ed.), 1973a, pp. 33-76.

C. FILLMORE, *The Case for Case*, in E. BACH, R. T. HARMS (eds.), *Universals in Linguistic Theory*, Holt, Rinehart & Winston, New York, 1968, pp. 1-88.

C. FILLMORE, *Scenes-and-frame semantics*, in A ZAMPOLLI (ed.), 1977b.

J. GREENBERG, *Some universals of grammar with particular reference to the order of meaningful elements*, in J. GREENBERG (ed.), *Universals of Language*, MIT Press, Cambridge (Mass.), 1966.

A. GRILLI, N. MARINONE, A. ZAMPOLLI, D. A. BROGNA, V. LOMANTO, L. FIOCCHI, *Concordanza dei grammatici latini*, in *Supplemento agli Atti dell'Accademia delle Scienze*, Torino, vol. 112, 1978.

M. GROSS, *Methodes en Syntaxe*, Paris, 1975.

R. S. JACKENDOFF, *On some questionable arguments about quantifiers and negation*, in « Language », XLVII (1971) 2, pp. 282-297.

R. S. JACKENDOFF, *Semantic Interpretation*, MIT Press, Cambridge (Mass.), 1972.

H. H. JOSSELSON, *The Lexicon: a System of Matrices of Lexical Units and their Properties*, in « ICCL », 1969.

J. J. KATZ, *Semantic Theory*, Harper & Row, New York, 1972.

J. J. KATZ, J. FODOR, *The structure of a semantic theory*, in « Language », XXXIX (1963), pp. 170-210.

G. LEECH, *Semantics*, Penguin, London, 1974.

J. D. MC CAWLEY, *The role of semantics in a grammar*, in E. BACH, R. T. HARMS, *Universals in Linguistic Theory*, Holt, Rinehart & Winston, New York, 1968, pp. 125-169.

J. MACNAMARA, *Parsimony and the lexicon*, in « Language », XLVII (1971) 2, pp. 359-374.

J. PETÖFI, *Lexicology, Encyclopaedic Knowledge, Theory of Text*, in « Cahiers de Lexicologie », XXIX (1976) 2, pp. 25-41.

R. RUSTIN (ed.), *Natural Language Processing*, Algorithmic Press, New York, 1973.

C. SEGRE, A. ZAMPOLLI, *La concordanza diacronica del « Furioso »*, in *Atti del Convegno di Studi « Lingua, stile e tradizioni delle opere dell'Ariosto »*, (Reggio Emilia-Ferrara), 1974.

R. SIMMONS, *Semantic networks: their computation and use for understanding English sentences*, in R. SCHANK, K. COLBY (eds.), *Computer Models of Thought and Language*, Freeman, San Francisco (Calif.), 1973.

D. D. STEINBERG, L. A. JAKOBOVITS (eds.), *Semantics*, Cambridge University Press, Cambridge, 1971.

U. WEINREICH, *Explorations in semantic theory*, in T. A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. III, Mouton, The Hague, 1966.

A. ZAMPOLLI, *Projet d'un dictionnaire italien du machine, Intervention*, in R. BUSA (ed.), *De lexico electronico latino*, Pisa, 1968.

A. ZAMPOLLI, *Nota tecnica*, in A. M. BARTOLETTI COLOMBO (ed.), *La Costituzione della Repubblica Italiana, Testi, Indici, Concordanze*, Firenze, 1971.

A. ZAMPOLLI (ed.), *Linguistica matematica e calcolatori*, Firenze, 1973a.

A. ZAMPOLLI, *Humanities Computing in Italy*, in « Computers and the Humanities », 7 (1973b) 6, pp. 343-360.

A. ZAMPOLLI, *La section linguistique du CNUCE*, in A. ZAMPOLLI (ed.), 1973a, pp. 133-199.

A. ZAMPOLLI, *L'elaborazione elettronica dei dati linguistici: stato delle ricerche e prospettive*, in *Atti del Convegno sul tema « Le tecniche di classificazione e loro applicazione linguistica »*, Accademia Nazionale dei Lincei, Roma, 1975.

A. ZAMPOLLI, *Trattamento automatico di dati linguistici e linguistica quantitativa*, in SOCIETÀ DI LINGUISTICA ITALIANA, *Dieci anni di linguistica italiana (1965-1975)*, Roma, 1977a, pp. 349-370.

A. ZAMPOLLI (ed.), *Linguistic Structures Processing*, North-Holland, Amsterdam, 1977b.

N. ZINGARELLI, *Vocabolario della Lingua Italiana*, X ed., Zanichelli, Bologna, 1970.