M. Mennucci - E. Morreale

AN INTERACTIVE SYSTEM FOR STEM-SUFFIX DISCRIMINATION IN ITALIAN WORDS

1. *Introduction.*

Today, owing to the growing diffusion of on-line processing facilities, the automatic processing of textual data, i.e. of information expressed in natural language, becomes more and more important both for:

1) applications concerning specifically the processing of textual fragments, (studies in linguistic analysis, in information retrieval and question-answering systems, etc.);

2) the implementation of a viable communication process within interactive systems also for problems not specifically linguistic, (such as computer aided instruction, computer aided design, etc.), so that non specialized people can access and use the system without the burden of some more or less rigidly coded command language.

A typical application which combines both of the above aspects can be found in a hospital where the collection and the analysis of clinical data is to be performed automatically. This will be accomplished by a combined hardware-software system capable to process and analyze clinical data expressed in narrative form, e.g. the patients' anamneses; it should furthermore enable the sanitary staff of the hospital to use some form of on-line communication language as explicit as possible.

Any significant processing of information expressed in natural language requires one be able to perform some kind of content analysis on the input data, and to infer some significant associations between these data and particular actions which are considered pertinent in a given context. For instance, in the case of the hospital system mentioned before, the content analysis of patients' anamneses should make it possible to relate the textual data composing an anamnesis with the clinical aspects considered useful for the possible diagnosis and therapy.

The content analysis to be performed on a text requires that both syntactic and semantic features of the text be simultaneously analyzed:

- the syntactic features will lead to the recognition of the simple linguistic elements forming the given, more complex, one;
- the semantic features will associate with the different linguistic elements, composing the text, concepts and actions which are relevant for the particular context in which the text is analyzed; for instance the clinical context in the case of an anamnesis.

In the case of artificial languages, it has become customary to analyze separately these two kinds of features. The much greater complexity of natural languages consists essentially in the fact that such a sharp distinction between these two kinds of features can no longer be made, and therefore, the analysis will require a more unified procedure.

The necessity of a combined syntactic and" semantic " analysis arises even at the very preliminary level we are considering in this work, i.e. at the level of the morphological analysis of the words composing a text. More precisely, we shall consider the problem of an automatic morphological analysis of Italian words, and hence the computer-aided construction of an Italian morphological dictionary.

The necessity of a combined analysis, as mentioned before, arises already at this level as a consequence of the fact that the stem of a word carries the bulk of its semantic value, while the various suffixes that can be appended to it must follow specific morphological rules; they are, furthermore, related to the possible syntactic uses of the word in a sentence and determine, at the same time, its precise meaning in the text.

Actually, from such a dictionary, one would obtain for any Italian word:

- some grammatical qualifications such as substantive, adjective, verb, singular, plural, masculine, feminine, tense, person etc.;
- some semantic qualifications specifying which kinds of relations tie together this word with some other words in the dictionary, according to the particular context in which the text is analyzed.

In such a way, after stem-suffix analysis has been made, grammatical qualifications will mainly result from suffix structures which are common to all the words of the language, while semantic qualifications in any given context will depend only on stems. Therefore the semantic relations constituting these qualifications will be built around the set of stems contained in the dictionary.

According to the above remarks, our work has been oriented toward the study of the flexive structure of Italian words, i.e. the structure according to which, from a single stem, through affixing it with a set of different suffixes, a set of different forms can be derived, each one being qualified by convenient grammatical categories.

The problem of building a morphological dictionary is considered here more from the Information Science point of view than from the linguistic point of view. More precisely, our attention has been mainly focused on the problem of giving some automatic assistance to human operators engaged in the construction of a morphological dictionary, and an interactive system has actually been studied for this purpose. This system has been planned for languages which, like Italian, have a richly flexive structure.

In the present paper, after some short comments about the utility of automatic morphological analysis and the different approaches which can be followed in constructing an Italian morphological dictionary (sec. 2), we give the description of the proposed system (sec. 3) and some concluding remarks (sec. 4).

2. *Computer aided construction of a morphological dictionary.*

Morphology is that part of linguistics which studies word formation and provides us with rules and explanation about the internal structure of a word. Such information about the internal structure of words can be very useful in the analysis of a text at least in four major areas:

1) It helps us control the widening, enrichening and modification of the concepts used within the universe of discourse which is specific to any application context in which textual analysis is performed. In a system in which no morphological analysis is provided any word is considered merely as a " sequence of characters ", each one independent from the others; therefore the addition of any new word will be considered merely as the addition of a new entry to a list of words, even if the new word is simply a new derivation of a stem already present in the dictionary through one or more of its forms.

Furthermore the use of morphological analysis allows one to give semantic qualifications to the stems instead of to the forms, facilitating thus both a more compact and uniform treatment of semantics and the recognition of similarity among aggregates as:

matematica applicata
applicazione della matematica
applicare la matematica
applicazioni matematiche.

2) In languages possessing a rich flexional structure – like Italian – it allows a significant compression in the dimension of the whole dictionary; in languages like Russian or Finnish a ratio between 10-20 can be estimated between a full form and a stem dictionary.

3) It allows the automatic expansion of words in their full flexional schemes.

4) It allows an easier and more uniform association of grammatical categories to the forms composing a text, so that subsequent steps for a more comprehensive content analysis can reach more significant results, even when hampered by some not yet encountered form.

In order to perform an automated morphological analysis of a language, it is mandatory to have a morphological dictionary of that language, i.e. a tabulated and/or algorithmic means, allowing the distinction of stem and affixes within a word.

Without entering here into a complete analysis of such dictionary-creating processes, we can agree on the fact that these processes can be complex and long, in accordance with the large number of forms to be considered and the large number of affixional structures which must be taken into account.

It can therefore be obvious to ask whether one could receive some operational help from some automatic means in constructing this dictionary, and it is in this direction that our approach has been mainly oriented.

In the case of the construction of a general morphological dictionary, perhaps the most immediate approach could be that of starting from a conventional dictionary and expanding all its entries in all their possible forms, according to a set of flexional structures taken from a conventional grammar. The situation appears somehow different when the dictionary is to be used in some specific field of application, for instance that of anamneses analysis in a particular medical field. In fact, in this case it is very difficult to reach any a priori decision about:

1) the selection of the specific words to be included in the dictionary;

2) the selectio n of the non-specific words to be included in the dictionary:

3) the selection of the flexional structures to be considered for the expansion of the above words.

On the other hand, in almost all applications of such kinds, one has a quite large and significant sample of text fragments to analyze, which in some way explicitly defines the area in which linguistic analysis should be made.

According to such considerations, the guidelines which seemed to us worth following for the computer-assisted construction of a dictionary, can be condensed as follows:

1) to start from the collection of text fragments which are already available, i.e. from a set of forms, assuming, as we have verified, that in such texts many forms derived from the same stem are present;

2) to apply some automatic mechanism which can put in evidence, as much as possible, the strong regularities which are characteristic of the Italian morphology, so that human decisions can be applied to classes of words, and to exceptions;

3) to try to take into account as many as possible of the regularities and of the exceptions explicitly outlined by a conventional grammar, in order to enhance the efficiency of the above process.

Therefore, according to the explicit indications (examples) given by a conventional grammar, two sets, E and R, of words will be formed:

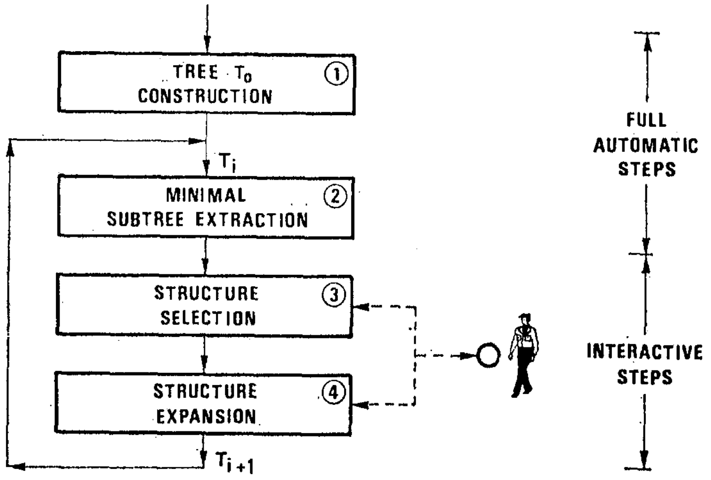
- the exlusion list E, which will contain all those words for which no, or very individual behaviour, is indicated;
- the regularity list R, which will contain some flexionally complete samples (hereafter to be called " templates ") for any flexional structure which is representative of a large class of words.

For a specific application to a certain field, for which we have a collection of text-fragments, we shall, first of all, build the set S of all forms contained in it, subtract from S the elements belonging to E, and add to it the elements of R; the resulting set will be the input for the system, to be processed as described in the next section.

3. *Description of the system.*

We shall assume that the text sample to be processed includes, for a significant number of stems, a sufficient number of derived forms, so that the extraction of a significant number of flexional structures is at least in principle feasible.

The basic structure of the system can be schematized as in Fig. 1,



STRUCTURE OF THE SYSTEM

Fig. 1.

in which, after the initial step devoted to the " tree construction ", the remaining ones, i.e.

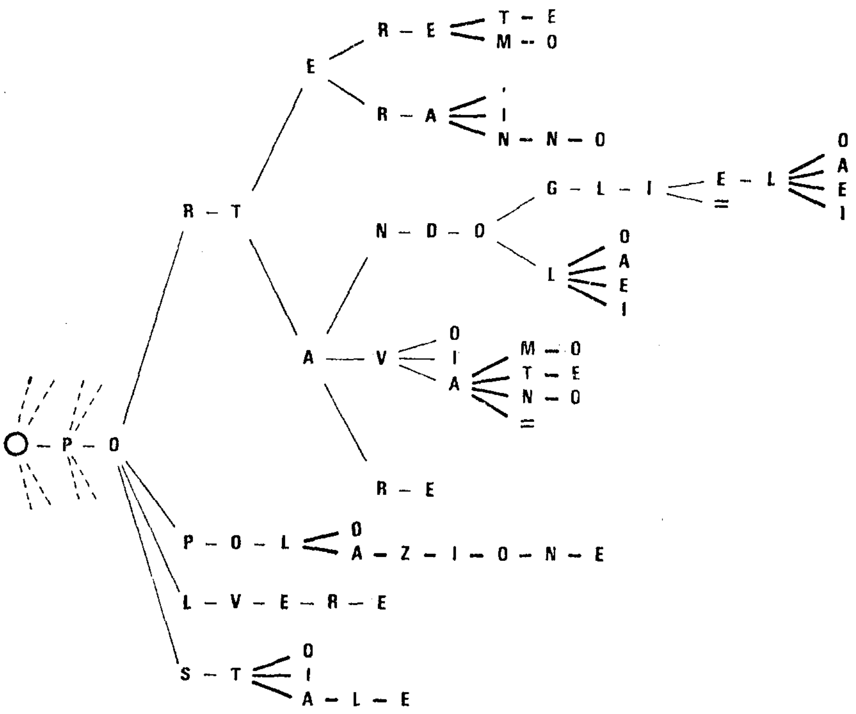
step	function
2	minimal subtrees extraction
3	structure selection
4	structure expansion

can be repeatedly iterates Let us remark that:

- steps 1,2 are fully automatic, while the remaining two involve some amount of interaction with the operator;
- the possible iteration of steps 2,3 and 4 is related with the " articulation depth " of the flexional structures to be recognized.

Along the whole process, the given set of forms is stored and processed as a tree-like data structure, like the one exemplified at the right of Fig. 2 which refers to the excerpt of forms listed at the left. In this structure, each node represents a single character of a form, and an oriented link betwee two nodes represents how these characters follow each other in that form In the figures the character =, denoting the

.....
PORTARE
PORTANDOGLIELE
PORTANDOGLIELO
PORTANDOGLIELI
PORTANDOGLIELA
PORTANDOLI
PORTANDOLE
PORTANDOLO
PORTANDOLA
PORTERAI
PORTERA'
PORTERANNO
PORTAVANO
PORTERETE
PORTEREMO
PORTAVATE
PORTAVI
PORTAVO
POPOLO
POPOLAZIONE
POLVERE
POSTO
POSTI
POSTALE
...



AN EXCERPT OF THE INITIAL FORM TREE T_0

Fig. 2.

end of each form, has been omitted wherever unnecessary. In our system, the realization of this data structure is performed by step 1, which reads the different forms and builds up the initial tree T. At the end of the Ith iteration the input tree T_{i-s} will be processed and transformed into the new one T_i, which possibly will be processed similarly by the next iteration. Obviously, forms derived from the same stem constitute a subtree of the whole tree (see Fig. 3). In such a subtree, unless it is an improper one, we distinguish:

– a " cut node " (c.n.), i.e. the leftmost branching node in the tree;

– a stem, i.e. the sequence of characters preceding and including the c.n.;

– a " flexional structure ", i.e. the set of branches starting from the c.n. (but not including it).

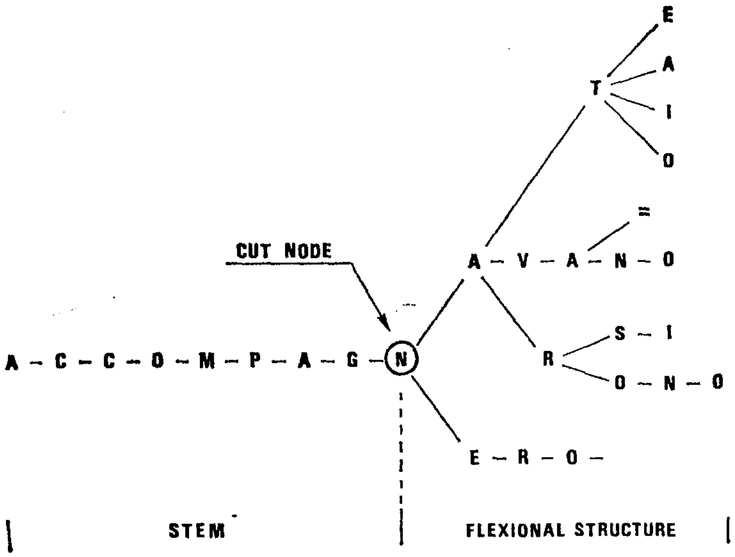
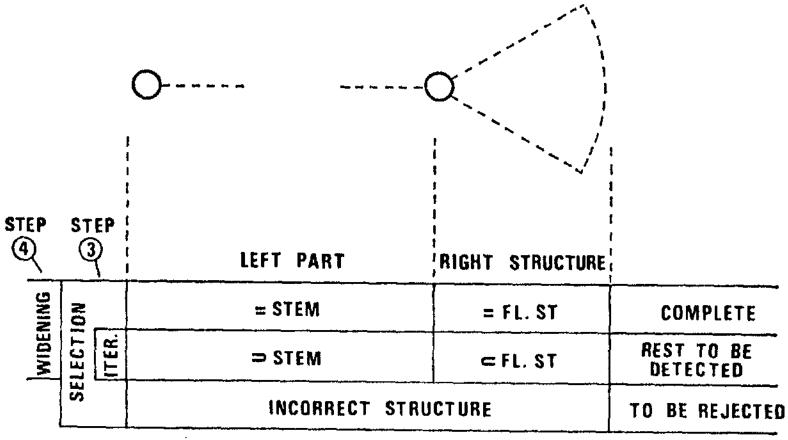


Fig. 3

It must be noticed that, if we extract a generic subtree from the form tree, the left and the right part does not generally coincide with the stem and the flexional structure, so we call them " left part " and " right structure ". Three different stituations (as Fig. 4 shows) may arise:



	LEFT PART	RIGHT STRUCTURE	
	= STEM	= FL. ST	COMPLETE
	⊃ STEM	⊂ FL. ST	REST TO BE DETECTED
	INCORRECT STRUCTURE		TO BE REJECTED

Fig. 4

1) the " left part " coincides with the stem; the " right structure " coincides with the flexional structure; no further operations are needed;

2) the " left part " contains the stem plus some characters of the flexional structure; the " right structure " is properly contained in the flexional structure; further iterations are needed;

3) the forms in the subtree are not derived from the same stem and the structure must be rejected.

We are actually interested in extracting, from the given set of forms, all the subtrees related to the same stem, possibly after having widened these subtrees to their largest linguistically consistent expansion, so that both the left part and the associated flexional structure reach their most stable form. For this purpose, two basic operations must be performed:

1) extraction of – even incomplete – subtrees related to the same stem; this process will result essentially from the, possibly iterated, extraction of " minimal subtrees " to be defined below;

2) widening, as much as possible, of the flexional structures extracted before; this process will benefit from the introduced templates, and will require some amount of interaction with the operator.

In order to describe the functions performed in step 2, let us define a " minimal subtree " of a given tree as the tree of all (at least two) branches which emerge from the same node, and which do not give rise to any other proper branching. Let us call " right part " any set of characters constituting a branch in a minimal subtree, and " right structure " the set of all right parts in a minimal subtree.

The result of applying step 2 to our example is shown in Fig. 2, where all minimal subtreees have been marked by bold lines. As this example shows, the majority of minimal subtrees so extracted refer to the same stem, and this result has beeen strongly confirmed by our experience on Italian words. A more detailed analysis of the different situations which can arise during the extraction of minimal subtrees is made through Fig. 5:

– In Fig. 5 *a*, the extracted minimal subtree represents exactly a flexional structure;

– in Fig. 5 *b*, the detected minimal subtree gives us a right structure constituting only a part of a broader flexional structure which will be detected by further iterations;

– in Fig. 5 *c*, the detected minimal subtree gives us a right structure which does not represent a flexional structure and should therefore be discarded.

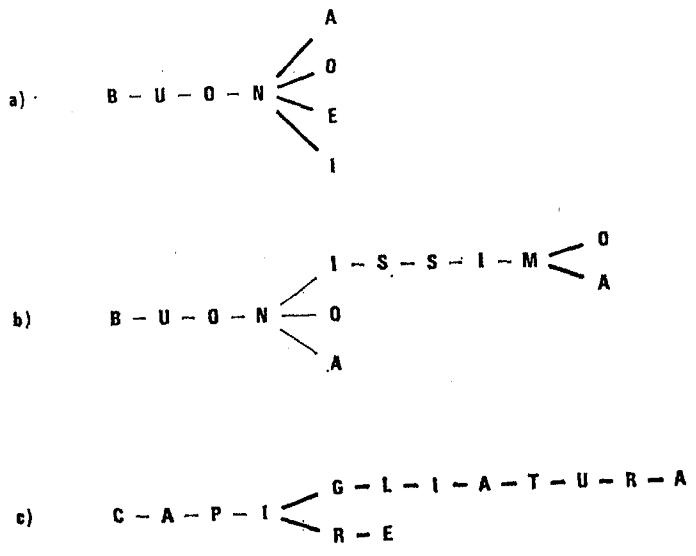


Fig. 5

According to this discussion, the required filtering of minimal subtrees resulting from step 2 is performed by step 3, mainly on the basis of the number of occurrences of structures and of their terminations, and under the control of the operator.

Let $\sigma = \{\tau_1, \tau_2 ...., \tau_k\}$ denote any structure composed of k terminations $\tau_1 \ \tau_2 ......, \tau_k$, and let $\omega(\sigma)$ and $\omega(\tau_i)$ denote the number of occurrences of a given structure σ and of a given termination τ_i respectively. Assuming that two threshold values s and $t \leqslant s$ have been selected for the number of occurrences of structures and terminations respectively, the filtering of structures is performed as follows: for any structure

1) if $\omega(\sigma) > s$ then the structure σ is retained;

in the case $\omega(\sigma) \leqslant s$ the occurrences $\omega(\tau_i)$ of terminations τ_i are matched against the threshold t, and decisions are taken according to one of the following three possible outcomes:

2*a*) if $\omega(\tau_i) > t$ for $i = 1, k$, then the structure σ is retained;

2*b*) If $\omega(\tau_i) \leqslant t$ for i = 1, k, then the structure σ is rejected;

2*c*) in the remaining case, the set $\sigma = \{\tau_1, \tau_2 \ldots, \tau_k\}$ is properly partitioned by the threshold into two non empty subsets $\sigma' = \{\tau'_1, \tau'_2 \ldots,$

$\tau'_{k'}\}$ with $\omega(\tau'_i) > t$, and $\sigma'' = \{\tau''_1, \tau''_2 ...., \tau''_{k''}\}$ with $\omega(\tau''_i) \leqslant t$, and one of the three following actions can be optionally applied:

- the structure σ is rejected;
- only the substructure σ' is retained;
- the structure σ is displayed and further decisions about it are left to the operator.

The above filtering scheme could be further refined by taking into account even some other parameters, as for instance: the length (number of characters), or the mean length of the right parts within the structure σ.

After this selection step, all the words related to rejected structures will be removed from the tree, and will be collected as a part of the output of the system.

In any iteration, after partial flexional structures have been detected and selected by steps 2 and 3, step 4 will perform the possible widening of these structures as follows, on the basis of similarity matching between structures and under operator control. For every pair of structures σ_i, σ_j i $\neq$ j a similarity value $X(\sigma_i, \sigma_j)$ is computed through a properly selected function X, which evaluates the similarities between two structures. For instance one could select:

$$X = max \left(\frac{C}{N_i + C}, \frac{C}{N_j + C} \right),$$

or else:

$$X = \frac{1}{2} \left(\frac{C}{(N_i + C} + \frac{C}{N_i + C} \right),$$

where:

$C =$ number of characters in $\gamma = \sigma_i \cap \sigma_j$ (see Fig. 6)
$N =$ » » » » $\sigma_i - \gamma$
$N =$ » » » » $\sigma_j - \gamma$

According to the decreasing order of their similarity values, pairs of structures σ_i, σ are displayed to the operator, for decision about the possibility of widening σ_i and/or σ_j in the structure $\sigma_i \cup \sigma_j$. In this phase, the operator can communicate his decisions through a set of commands which allow him to manipulate structures, transfer forms from a structure to another, etc. Commands will be also provided for displaying entities (structures, forms related to a given structure, etc.) useful to

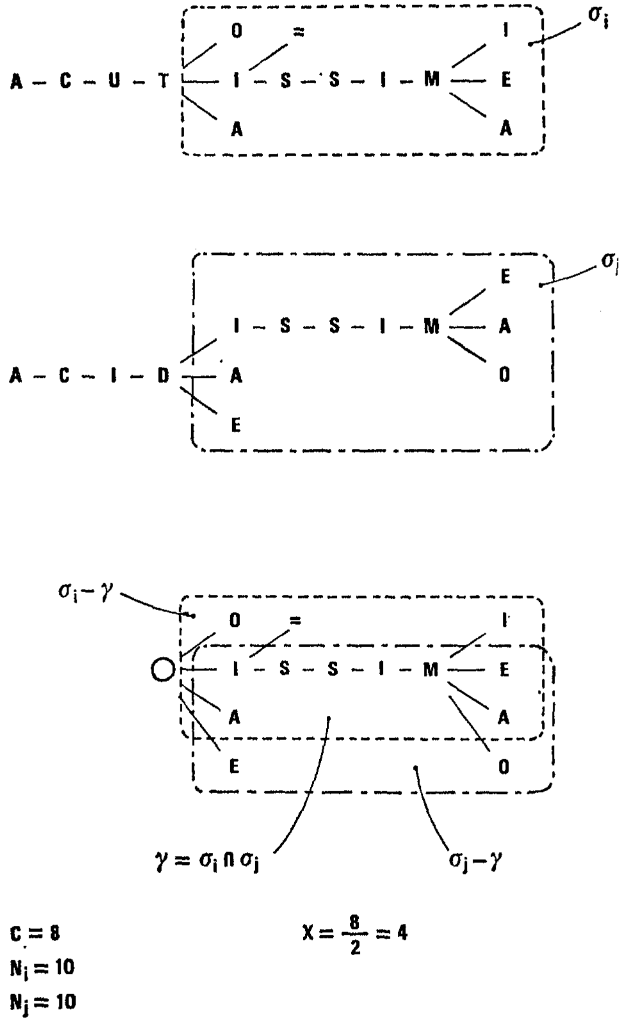


Fig. 6

evaluate and possibly adjust parameters, as thresholds *t* and *s* in step 3, which control the process.

After having shown the functions performed by the single steps of the system, let us discuss through the example in Fig. 7 how the system works by iterating steps 2, 3 and 4. Let us suppose that the nine-

SUFFIXES	
1	TE
2	MO
3	'
4	I
5	NNO
6	O
7	A
8	E
9	NO
10	LI
11	RE ①
12	RA ②
13	EL ③
14	A ④
15	GLI ⑥
16	L ③
17	NDO ⑧
18	V ⑦
19	RE
20	E ⑤
21	A ⑨

STRUCTURES		
①	1	TE
	2	MO
②	3	'
	4	I
	5	NNO
③	6	O
	7	A
	8	E
	4	I
④	2	MO
	1	TE
	9	NO
	10	␣
⑤	11	RE ①
	12	RA ②
⑥	13	EL ③
	10	␣
⑦	6	O
	4	I
	14	A ④
⑧	15	GLI ⑥
	16	L ③
⑨	17	NDO ⑧
	18	V ⑦
	19	RE

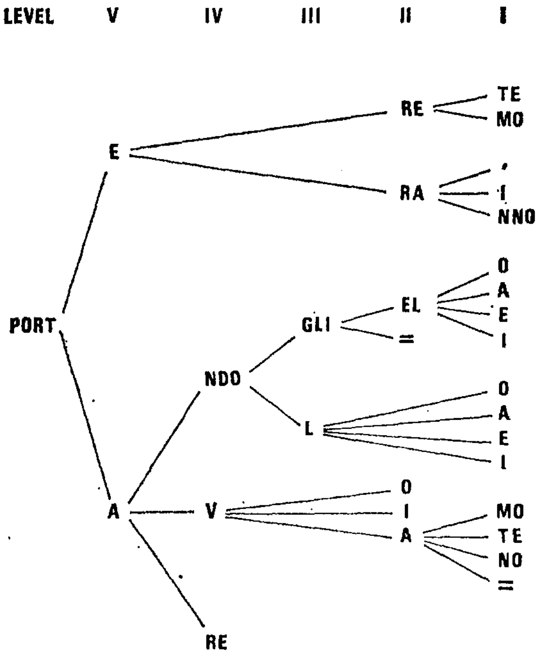


Fig. 7.

teen uppermost forms shown in Fig. 2 have been submitted to the system. The set of structures formed by the first iteration would be the one shown in Fig. 7 under level 1. When these structures have been detected (for the sake of simplicity we have omitted the effects of possible expansions eventually generated by step 4), each one of them will be condensed in a single node representing the identification of that structure and new iterations will then be applied. In this way, all the further levels, shown in Fig. 7, of the flexional structure will successively emerge, converging at the end to a stable left part (stem) and a stable right structure (flexional structure) associated with it.

In any new iteration, as already remarked about step 3, the form tree T_0 will be progressively stripped of some forms, and the process will stop when it vanishes. At this point, all the material resulting from the process, i.e. left parts and flexional structures, will constitute a base suitably structured to receive the grammatical qualifications needed in a morphological dictionary.

4. *Concluding remarks.*

Some parts (steps 1, 2) of the system described above have already been implemented, while implementation is presently under way for the others (steps 3,4). The set of commands already implemented allows one to perform, on a non automatic basis, all the interactive operations by steps 3 and 4.

At the present stage of implementation, a teletype is used for the interaction. However, in order to speed up the interactive communication, the use of an alphanumeric video display equipped with a light pen, is planned.

So far some experiments have been made with the implemented parts of the system, by processing a small sample of texts, i.e. 30 neurological anamneses including about 2000 forms. These tests, even though incomplete, have shown that the process of grouping together forms derived from the same stem is largely satisfactory, as it has given after the first application of steps 1,2 and 3 (in a non-automatical version), a very low fraction of still incorrect groups, i.e. groups of forms non derivable from the same stem.

These preliminary results allow us to think that, after the implementation of the interactive portion of the system is completed, the system itself can be a useful tool for building morphological dictionaries for specific applications involving the analysis of textual data expressed in a richly flexive language as Italian. The main advantages that such a system can offer can be summarized in the following points:

1) The textual data to be analyzed constitute by themselves the input to the system, and no further input is required;

2) human operations are restricted mainly to validating and/or modifying structures already prepared by the system;

3) manual interventions can be applied to whole classes of morphologically homogeneous forms;

4) grammatical qualifications can be applied to whole classes of morphologically homogeneous forms.