A flexible treatment of complex lexical information

Nicoletta Calzolari, M. Luigia Ceccotti, Eugenio Picchi, Adriana Roventini

1. The Italian Machine Dictionary

The evolution of the Italian Machine Dictionary (DMI) is closely connected both to the changes in purposes and interests within the field of Computational Linguistics, and to the developments in data processing technologies. Its current organization is in the form of a very large Lexical Data Base, built on the relational model, to be used both for theoretical investigations in the lexical structure and for many applicative purposes in natural language processing.

We now envisage a new organization of the dictionary data which is aware of the further possibilities offered by computational techniques and devices, and gains profit from the linguistic analysis of the lexical data carried out so far: the two aspects are to be considered in strict connection. We are therefore considering the design and implementation of a new structure which on the one hand maintains the benefits provided by the present database structure (e.g. direct access, interaction, multiple views on the data, etc.), and on the other hand makes the structure of the data independent on the storage devices and open to the addition of new types of information, e.g. predicate-argument structures, surface syntactic structures, case-frames, selection restrictions, semantic features, etc.

The solution we envisage is to design a machine dictionary which is no longer a static and very large set of data, but the combination of a much smaller set of data along with a set of rules which are controlled by procedures operating on them (with obviously no loss of relevant information). Another important requirement is to achieve greater flexibility in the design of the logical organization of the data, in view of a diversified use of the lexical data base, both by different types of procedures (e.g. for lemmatization, parsing, machine (aided) translation, computer assisted instruction, etc.) which need diversified selections of lexical information, and by different kinds of human users.

From the linguistic point of view we can mention the following as typical examples of this kind of solution: a) the new representation of the inflectional structure, and b) the researches we are carrying out on the phenomenon of lexical word-formation (derived words forming a considerable part of the entire lexicon).

2. The representation of inflectional morphology

With regard to the first example, the old dictionary organization consisted in a list of word-forms, each linked to its reference-lemma by means of a pointer.

However some problems arise also in connection with the present database structure on mass storage system, among which the most evident are the following:

(i) the very large amount of storage occupied by the overall dictionary;
(ii) the reliance on the chosen systems of hardware and software: e.g. the dependence on the present physical devices (disks or mass storage systems) and on the virtual storage access method (VSAM) used for the access and indexing of the data;
(iii) as a consequence of (i) and (ii) many difficulties in the portability of the lexical database;
(iv) problems of maintenance of the data.

The new structure is instead made up by a list of lemmas (in a very compressed form), each associated with paradigm codes. Appropriate procedures enable the system to use these codes to generate and to recognize all the existing word-forms.

The codified inflectional morphological system was implemented by using the archive of already existing word-forms. The aim was to automatically obtain the entire series of morphological models for the automatic generation of the word-forms relevant to the whole list of lemmas. The set of lemmas was divided into two groups:

- the first group was formed by all the verbal lemmas;
- the second group was formed by all the non-verbal lemmas.

The verbal lemmas are subdivided into the three following groups:

a) lemmas ending in *-are*;
b) lemmas ending in *-ire*;
c) lemmas ending in *-ere* and in *-rre*.

The non-verbal terms are subdivided into three subsets:

a) lemmas ending in *-e*;
b) lemmas ending in *-o*;
c) lemmas ending in *-a*, plus all the others.

Each subset is defined by tables of endings and also by models of behaviour which enable the generation system to obtain a complete paradigm starting from each lemma (see Fig. 1). A procedure was realized which – for each subset of the lemmas, and starting from a defined set of endings and "flexive" models – analyzes the entire list of word-forms and then produces a number of codes associated to each lemma; each code biunivocally identifies a model of inflectional behaviour. This procedure, recursively employed by introducing new endings and models, makes it possible to completely solve the entire inflectional morpholo-

```
TABELLA IN -ARE

001 03ARE                   F
04      TAB. 2
001 01A                     S3IP
002 01A                     S2MP
003 01O                     S1IP
004 03ANO                   P3IP
03     TAB. 3
001 04IAMO                  P1IP
002 04IAMO                  P1CP
003 04IATE                  P2CP
05     TAB. 4
001 01I                     S2IP
002 01I                     S1CP
003 01I                     S2CP
004 01I                     S3CP
005 03INO                   P3CP
12     TAB. 5
001 04RO'                   S1IF
002 04RAI                   S2IF
003 04RA'                   S3IF
004 05REMO                  P1IF
005 05RETE                  P2IF
006 06RANNO                 P3IF
007 04REI                   S1DP
008 06RESTI                 S2DP
009 06REBBE                 S3DP
010 06REMMO                 P1DP
011 06RESTE                 P2DP

    --------------------

006 12.16cAd-E5--34#
007 12.16cAd-H34E5#
008 1234.16cAd-E5#
009 1.112.412.514.41c.61cA.31d.51d.c1d-E.315.615.b15#
010 12cAd-H34E5$--IEC12#
011 +2U64-$1234cAd-E5#
012 +2I67$12cAd-H34E5#
013 12cAd-E5-4-3#
014 +2U67$12cAd-H34E5#
015 135cAd-E5----V8fAD.312#
016 10A5T9-CC3b-Ed---ECa#
```

Fig. 1: Table of endings and rules for generating the word-forms of a subset of verbs.

DMI on Pc by E.Picchi

```
Num.Lemma :  5014000

    Lemma : ANDARE

 Cod.Flex : 015
Zing/Garz : *
  Cod.Uso :
 Cod.Gram : 075 VI
```

1	ANDARE	F
2	VADO	S1IP
3	VO	S1IP
4	VO'	S1IP
5	VAI	S2IP
6	VA'	S3IP
7	ANDIAMO	P1IP
8	ANDATE	P2IP
9	VANNO	P3IP
10	ANDAVO	S1II
11	ANDAVI	S2II
12	ANDAVA	S3II
13	ANDAVAMO	P1II
14	ANDAVATE	P2II
15	ANDAVANO	P3II
16	ANDAI	S1IR
17	ANDASTI	S2IR
18	ANDO'	S3IR
19	ANDAMMO	P1IR
20	ANDASTE	P2IR
21	ANDARONO	P3IR

Fig. 2: Automatic generation of the word-forms of the verb ANDARE.

gical system including the irregularities which are thus inserted within a general model.

The procedure described exploits this system for the automatic generation of all the word-forms starting from the list of lemmas (see Fig. 2), whereas another procedure makes it possible to recognize the word-forms, associating to each word-form to be analyzed the grammatical codes and the appropriate lemma (or the lemmas in the case of homographs). The set of these procedures and of the codes associated to each lemma replaces the entire existing list of forms with the advantage of a significant saving in the space necessary for the memorization and functioning of the system.

3. Derivational morphology

As far as morphology in general is concerned, we are therefore about to create something like a two-level structure, where the lower level formalizes the inflectional morphology (described above), and the upper level represents the derivational morphology, with considerable advantages, not only from the point of view of minimizing the volume of space occupied, but also, from the linguistic point of view, for a homogeneous representation of semantically connected families of words.

For word-formation, we are following two approaches: a) semi-automatic treatment of selected subsets of derived words, introducing a number of homogeneous codes for their meanings, b) semi-automatic identification of families of derivatives.

With regard to the first approach, the starting point was the analysis of the metalanguage of definitions. It is possible to interactively extract in the data base subsets of words ending with selected suffixes, and their natural language definitions. Semantic regularities expressed by typical definitional patterns were identified, and a normalization or standardization for similar patterns was studied. Each entry was supplied, where possible, with a label for a semantic rule, and with the numerical key acting as a pointer to the base-lemma in our data base. The rules operate as a kind of redundancy rules, and automatically give the meaning or part of the meaning for each derived word (for more details, see Calzolari et al., 1985).

With regard to the second approach, we first examined the subset of all the verbs of the first conjugation (8500 verbs ending in -ARE). Each stem was linked to all the words in the dictionary beginning with the stem and ending with one of the possible Italian suffixes taken from Italian grammars, and the frequency of each suffix was recorded. It was thus possible to automatically identify the most "productive" suffixes and to produce a separate list where each suffix was coded as a number. A matrix was then obtained where each stem was associated with the numbers representing its set of accepted suffixes. Appropriate procedures and sortings on the numerical codes of this large matrix allowed those stems with one or more suffixes in common to be grouped together, and those suffixes which more frequently occur together also to be individuated.

The following are the most frequent deverbal suffixes in our corpus (the total occurrences and the number of times in which they appear alone, i.e. they are the only suffix for a given stem, are shown):

suffix	**total**	**alone**
– AMENTO	1909	289
– ATO	1838	163
– ATORE	1630	79
– AZIONE	1407	291
– ATURA	1124	162
– ABILE	665	16
– ANTE	638	52
– ATA	612	76

The most frequent combinations of two suffixes are the following:

– AMENTO	– ATO	765
– ATO	– ATORE	752
– AMENTO	– ATORE	744
– ATORE	– AZIONE	553
– ATO	– AZIONE	488
– ATO	– ATURA	465
– AMENTO	– ATURA	434
– AMENTO	– AZIONE	429
– ABILE	– ATORE	421
– ATORE	– ATURA	410

After having identified – in the list of combinations of two suffixes – the most productive suffixes (– ABILE, – ABILITA', – AMENTO, – ANTE, – ATA, – ATIVO, – ATO, – ATOIO, – ATORE, – ATORIO, – ATRICE, – ATURA, – AZIONE, – IO, – OSO), we obtained the data relevant to the groups of the three most frequent suffixes occurring together with the same base:

– AMENTO	– ATO	– ATORE	462
– ATO	– ATORE	– AZIONE	322
– AMENTO	– ATORE	– AZIONE	299
– AMENTO	– ATO	– ATURA	283
– AMENTO	– ATO	– AZIONE	270
– ABILE	– ATO	– ATORE	268
– ABILE	– AMENTO	– ATORE	258
– ABILE	– ATORE	– AZIONE	253
– ATO	– ATORE	– ATURA	252
– AMENTO	– ATORE	– ATURA	246
– ABILE	– AMENTO	– ATO	220
– ABILE	– ATO	– AZIONE	207

This type of data also offers the opportunity to perform different analyses, as for example:

– to check whether the recurrence of certain couples of suffixes corresponds to a recurrence of identical changes of meaning;
– to check whether the groups of verbs associated with certain types of suffixes (and thus changes of meaning) identify groups with other significant regularities;
– to extend this type of procedure to the verbs (fewer in number) ending in -ERE, -IRE, -RRE;
– to check the features of the verbs which have produced no derivatives in the Italian lexical system.

The families of derivatives made evident can be handled, in a very compact way, by means of codes assigned to the stems, where each code bears the information both on the attested suffixes for the stem, and on the central or regular changes of meaning which the suffixes imply when applied to the base.

The above described treatment of morphological data as well as the design of a new logical organization of the entire lexical structure is necessary in order to allow a better portability of the overall system, and also to broaden the possibility of utilization of the dictionary in the new structure on a complete series of computers ranging from personal computers to the largest machines.

In the same perspective it is our intention to define a number of basic or primitive functions for accessing the system. These functions should reflect all the possible ways of accessing elementary pieces of information, in order to meet all the basic needs which can be foreseen. They can then be used as interchangeable software modules and can be called either by external programs in all the possible combinations or by a non-procedural query language for the human users. The objective of gaining in flexibility must always be taken into consideration when designing the logical organization of the data, in view of the possibility of querying or extracting specified subsets of data, as more suitable for any type of individual application.

References

Aronoff, Mark (1975), *Word formation in Generative Grammar*, Cambridge (Mass.): MIT Press.

Beard, Robert E. (1981), "On the question of lexical regularity", in: *Journal of Linguistics,* 17, 1: 31–37.

Byrd, Roy J./Klavans, Judith L./Aronoff, Mark/Anshen, Frank (1986), "Computer Methods for Morphological Analysis", in: *Proceedings of the Association for Computational Linguistics,* 120–127.

Calzolari, Nicoletta (1983 a), "On the Treatment of Derivatives in a Lexical Database", in: *Computers in Literary and Linguistic Research, Proceedings of the VII ALLC Symposium,* Pisa, June 1982, Pisa: Giardini, III, Supplement, 103–113.

Calzolari, Nicoletta (1983 b), "Lexical definitions in a computerized dictionary", in: *Computers and Artificial Intelligence,* II, 3: 225–233.

Calzolari, Nicoletta (1983 c), "Per un'analisi formale della derivazione in italiano: metodologia di lavoro e primi risultati", in: *Studi di Lessicografia Italiana,* Firenze: Accademia della Crusca, Vol. V, 229–242.

Calzolari, Nicoletta/Ceccotti, M. Luigia/Roventini, Adriana (1985), "Some generalizations on classes of derivatives", in: *Proceedings of the XII International ALLC Conference: Methodes quantitatives et informatiques dans l'étude des textes,* Geneve, Slatkine, 161–166.

Clark, Eve V./Clark, Herbert H. (1979), "When nouns surface as verbs", in: *Language,* 55, 4: 767–811.

Jackendoff, Roy (1975), "Morphological and semantic regularities in the lexicon", in: *Language,* 51, 3: 639–671.

Picchi, Eugenio (1983), "Textual Data Base", in: *Proceedings of the International Conference on Data Bases in the Humanities and Social Sciences,* New Brunswick (New Jersey).

Picchi, Eugenio/Calzolari, Nicoletta (1985), "Textual perspectives through an automatized lexicon", in: *Proceedings of the XII International ALLC Conference,* Nice.