Nicoletta Calzolari; Maria L. Ceccotti; Adriana Roventini

COMPUTATIONAL TOOLS FOR AN ANALYSIS OF TERMINOLOGICAL DATA IN A GENERAL DICTIONARY

Introduction

This paper describes a study which aims at querying, updating and extending the terminological data set of the ITALIAN MACHINE DICTIONARY DATA-BASE (DMI-DB, cf. Calzolari and Ceccotti 1981, Calzolari et al. 1983), an integrated system based on a relational model. We present some results obtained by terminological information retrieval operations on the DMI-DB.

At present, the DMI-DB, which has been constructed on the basis of a printed dictionary, the VOCABOLARIO DELLA LINGUA ITALIANA (Zingarelli), is constituted by three main relations:

(a) Lemmario, a set of 106,091 lexical entries with grammatical and other types of information associated as attributes;

(b) Formario, the set of 1,016,320 inflectional forms of the lexical entries, and associated attributes;

(c) Definizionario, the set of 185,899 definitions of the lexical entries with other semantic information.

We have concentrated our attention on two domains of the Definizionario relation: the domain of the Definitions attribute, the set of statements representing the meanings of nominal, verbal and adjectival lexical entries, and the domain of the Specialized Language Code attribute containing 106 values which correspond to 106 labels in the VOCABOLARIO DELLA LINGUA ITALIANA. These labels indicate particular fields and sub-fields of knowledge such as linguistics, mathematics and chemistry.

Unfortunately, the information retrieved using these codes as 'direct access keys' represents only a small percentage of the terminological information actually contained in the DMI-DB. The major part of such information cannot be accessed specifically since those records which, for instance, refer to a technical but commonly used term have not been assigned a label by the VOCABOLARIO. In all these cases, in the printed dictionary as well as the DMI-DB as it is now, the technical specification is only evidenced when reading the entire definition.

In order to ensure that the terminological data sets which can be accessed in the DMI-DB are as exhaustive as possible, we are now attempting to retrieve and exploit as much as possible of the terminological information not explicitly marked in the DMI-DB. Our aim is to extract such data from the definitions themselves as automatically as possible.

The retrieval of all records with a particular value in the

Specialized Language Code field is thus a starting point in describing a terminological data sub-model. The definitions registered in this sub-model contain verbal, adjectival and nominal word-forms which can be used as keys to gradually enrich this sub-model. Our analysis, empirically based on views of the terminological data model, aims at (i) identifying new terminological data with links to relevant elements of the definitions and (ii) contributing to a new organization of the Definizionario (cf. Calzolari 1983). This is of great importance with regard to the main objective of our project (cf. Calzolari and Ceccotti 1981): to model the lexical data-base so that it becomes both a valid point of reference for the professional linguist and a flexible tool for use by natural language processing programs. In both approaches it may be interesting to use the DMI-DB terminological views as windows, allowing a rapid and continuous interface between the linguistic reality and its model.

An example of a terminological sub-set in the DMI

As an example of a terminological sub-set taken from the Definition set of the DMI-DB, retrieval was performed of those 1,078 terms in the VOCABOLARIO which have one or more definitions with the label 'law'. Some quantitative data and observations on this 'legal' terminological sub-set are reported below.

(1) 15% of the terms are variants, mostly coded 'archaism', of other terms, e.g. volontà, volontade, volontate;

(2) 65% of the definitions are of nouns, 15% verbs, 20% adjectives;

(3) 8% of the definitions are locutions (asse 'asse ereditario', capacità 'capacità giuridica'), 6% are synonymic expressions (conduttore 'affituario/locatario', pontaggio 'pontatico'), 85% are 'Aristotelian' definitions consisting of genus and differentia specifica (assenso 'manifestazione di volontà che condiziona un atto', bannalità 'diritto del feudatario di imporre attività agricole');

(4) legal locutions are very numerous, constituting 10% of the expressions registered in all the 106 terminological sub-sets of the DMI-DB;

(5) it is possible to recognize several complete groups of terms which share the base form and differ only in their morphosyntactic function, e.g. imputabile, imputabilità, imputare, imputato, imputazione, but often a term from the same derivational paradigm and semantic field is absent, e.g. the nominal term garanzia is missing from the group garante, garantire, garantito, garentire;

(6) in this terminological sub-set we find many monosemic technical terms (declaratoria 'provvedimento giurisprudenziale avente carattere dichiarativo'), but also commonly used general words (cugino 'parente in linea collaterale') or terms which belong more to the history of law (curiato 'della Curia della Roma antica').

Diderot pointed out (cf. Guilbert 1973) that a dictionary is the mirror of a people's knowledge, and that a comparison of dictionaries from different periods would allow us to verify the development of a society. We considered how these terms labelled 'legal'

in the tenth edition of the VOCABOLARIO DELLA LINGUA ITALIANA (1971) had been treated in a previous edition (the second, 1922) of the same dictionary. The quantitative results of the comparison were as follows:

(1) 40% of the terms are recorded in both editions;

(2) 50% of the terms from the 1971 edition are present in the 1922 edition, but without any label;

(3) 10% of the terms from the 1971 edition are missing in the 1922 edition.

Furthermore, the terms which are missing in the 1922 edition are mostly derivatives and participles. In order to verify the inverse phenomenon, i.e. the disappearance in the 1971 edition of codes or lexical entries recorded in 1922, we compared by hand the sub-set of entries beginning with the letter 'A' in both editions. The surprising result of this comparison was that the two dictionaries contain altogether 187 labelled terms, but only 34 terms are recorded the same way in both.

Extension of a sub-set with information retrieval techniques

Given the inadequacy of the terminological sub-set extracted via the Specialized Language labels, the next step in the analysis was the exploitation of the capability offered by the design of the Definitional part of the DMI-DB (cf. Calzolari 1983) to directly access each word form in each definition, and to connect each definiens word with its definiendum.

This is a really dynamic way of consulting the dictionary, the search for desired terminological data being conducted, with information retrieval techniques, on relations which virtually restructure the dictionary's definitional content similar to a thesaurus. Ad-hoc procedures are in fact creating within the lexicon chains of synonyms or quasi-synonyms, of hyponyms and hyperonyms (superordinates). A number of different subdivisions are thus obtained, within the same initial lexical set, according to the specific and momentary interest of the user consulting the DB.

We have used some of these search procedures to extract new data in order to integrate the first sub-set. Retrieval was performed, throughout the entire lexicon and by means of key-words, of all the other entries connected to the key-words by hyponymic relation or by a more general relation of 'belonging to the same semantic field'.

With this purpose in mind we have interactively retrieved from the DB all lexical entries in whose definitions key-words important to a particular field are present, e.g. diritto, legge, giuridico, giurisdizionale in the 'legal' field.

Table 1

	total	with code	code to be added
AUTORITA'	199	21	48
AVVOCATI	5		5
AVVOCATO	22	3	
DELINQUENZA	2		2
DENUNCIA	8	1	3
DIRITTO	332	49	208
GIUDICE	27	9	12
GIUDICI	17	2	11
GIUDIZIALE	5	2	3
GIUDIZIARI	9	3	6
GIUDIZIARIA	23	10	13
GIUDIZIARIE	3		3
GIUDIZIARIO	29	7	21
GIUDIZIO	124	27	28
GIURIDICA	56	15	41
GIURIDICHE	6		6
GIURIDICI	19	11	8
GIURIDICO	66	25	40
GIURISDIZIONALE	20	12	8
GIURISDIZIONALI	5	2	3
GIURISDIZIONE	36		35
GIURISPRUDENZIALE	2	2	
GIURISTA	6		6
GIURISTI	2		2
GIUSTIZIA	44	5	15
IMPUTATI	4		3
IMPUTATO	22	8	13
IMPUTAZIONE	2	1	1
IMPUTAZIONI	3	3	
LEGALE	20	1	9
LEGALI	4	2	2
LEGGE	145	31	67
LEGGI	64	3	29
LEGISLATORE	2	1	1
LEGISLAZIONE	4		1
PENALE	33	8	25
PENALI	6	2	4
PROCESSI	36	1	4
PROCESSO	250	23	27
PROCESSUALE	21	13	9
PROCESSUALI	6	4	2
REATO	66	30	36
SENTENZA	28	7	16
SENTENZE	8	1	3

Table 1 presents some quantitative data concerning

(a) terms used as key-words in the DMI-DB searches (first column);

(b) how many lexical entries carry a given key-word in one of their definitions (second column);

(c) how many of these entries carry the Specialized Language code 'legal' (third column);

(d) the number of definitions, at present without any code, which according to our model ought to be marked with the code label 'legal' (fourth column).

It is interesting to note in Table 1 that some key-words practically select only terms belonging to the 'legal world'. However, there are other terms used here as key-words because of their legal connotations which show a wide currency in the 'ordinary' language with a more generic sense. It is in this non-technical broader sense, i.e. as non-marked words, that they are used for defining non-legal words. Moreover, certain key-words are grammatical homographs (such as legge, which is also 3rd person singular of the verb leggere 'to read') or lexical homographs (such as diritto, which is also used as an adjective with the meaning 'straight'); obviously retrieval based merely on the occurrence of graphic form cannot resolve such ambiguities. This could only be achieved by means of a definitional corpus which is both grammatically and semantically tagged.

We also see from Table 1 that with a simple search based on the presence in the definitions of only 44 key-words we are able to enrich the specialized sub-lexicon of some 779 terms almost mechanically for some sub-sets and by manual intervention for the others. Obviously some difficulties can arise as far as the assignment of codes is concerned, and these should be discussed with experts in the respective fields. Usually these doubtful cases depend on the position of the lexical item along the boundary line between the 'specialized' vocabulary of a technical field and the 'general' vocabulary of more common currency.

Some interesting results emerge from an analysis of each of the extracted sub-sets, for example the different ways in which the retrieved terms are grouped around the key-word. Consider the word contratto, which appears in 125 definitions. We observe that the retrieved terms from a particular lexical or semantic field manifest quite well-defined 'roles' in the respective field. We find a large and homogeneous core of 64 terms linked to contratto by hyponymic relation, e.g. assicurazione 'contratto che prevede indennizzo di danno in oggetto'. Formally this kind of relation is shown by the position of the key-word in the so-called genus part of the definition. These terms therefore denote as many 'types' of different contracts. Other, quantitatively more restricted but semantically relevant, sub-sets are the following:

11 verbs denoting actions connected with the term contratto, e.g. assicurare 'concludere un contratto di assicurazione';

15 nouns denoting persons variously involved in the action of a contract, e.g. assicurante 'chi conclude un contratto di assicurazione';

11 abstract nouns (usually deverbal nominalizations) referring to actions connected with a contract, e.g. disdetta 'dichiarazione unilaterale di volersi scioglier da contratto';

7 names of documents, e.g. abbonamento 'documento che attesta l'esistenza del contratto';

6 nouns referring to the money involved in a contract, e.g. abbonamento 'canone da pagarsi in base al contratto'.

This kind of information could be very useful in a subsequent semantic sub-categorization of these legal terms.

Conclusion

For the selection from the DMI-DB of the data related to these lexical sub-sets, the model appears suitable for extracting not only the lexical entries, but also the definitions associated with them. An analysis of the definitional data can in fact have the following purposes:

(a) to examine the definitions themselves, in order to find a way of standardizing their structure;

(b) to examine the different types and levels of relationships and links between lexical entries, in order to arrive at a first formalization of some of the relations;

(c) to discover any circularities or non-defining cross-references, in order to eliminate them;

(d) to evaluate the distinction between properly specialized terms and words which though originally from a technical field have spread into the general vocabulary, or vice versa.

The material thus extracted, analyzed and selected could serve to enrich already existing thesauri and allow the extension of terminological lexica within the DMI-DB.

References

Calzolari, N. (1983) "Lexical definitions in a computerized dictionary" Computers and Artificial Intelligence 3, 3: 225-233

Calzolari, N. and Ceccotti, M.L. (1981) "Organizing a large-scale lexical database" in Actes du Congrès International Informatique et Sciences Humaines ed. by L. Delatte. Liège: L.A.S.L.A.

Calzolari, N., Ceccotti, M.L., Roventini, A. (1983) "La terminologia giuridica in un dizionario di lingua: strumenti informatici per valutarne consistenza e tipologia" in Atti del IIIo Congresso Internazionale 'L'informatica giuridica e le comunità nazionali ed internazionali'. Roma: Corte Suprema di Cassazione/Centro Elettronico di Documentazione

Guilbert, L. (1973) "La spécificité du terme scientifique et technique" Langue française 17: 5-18