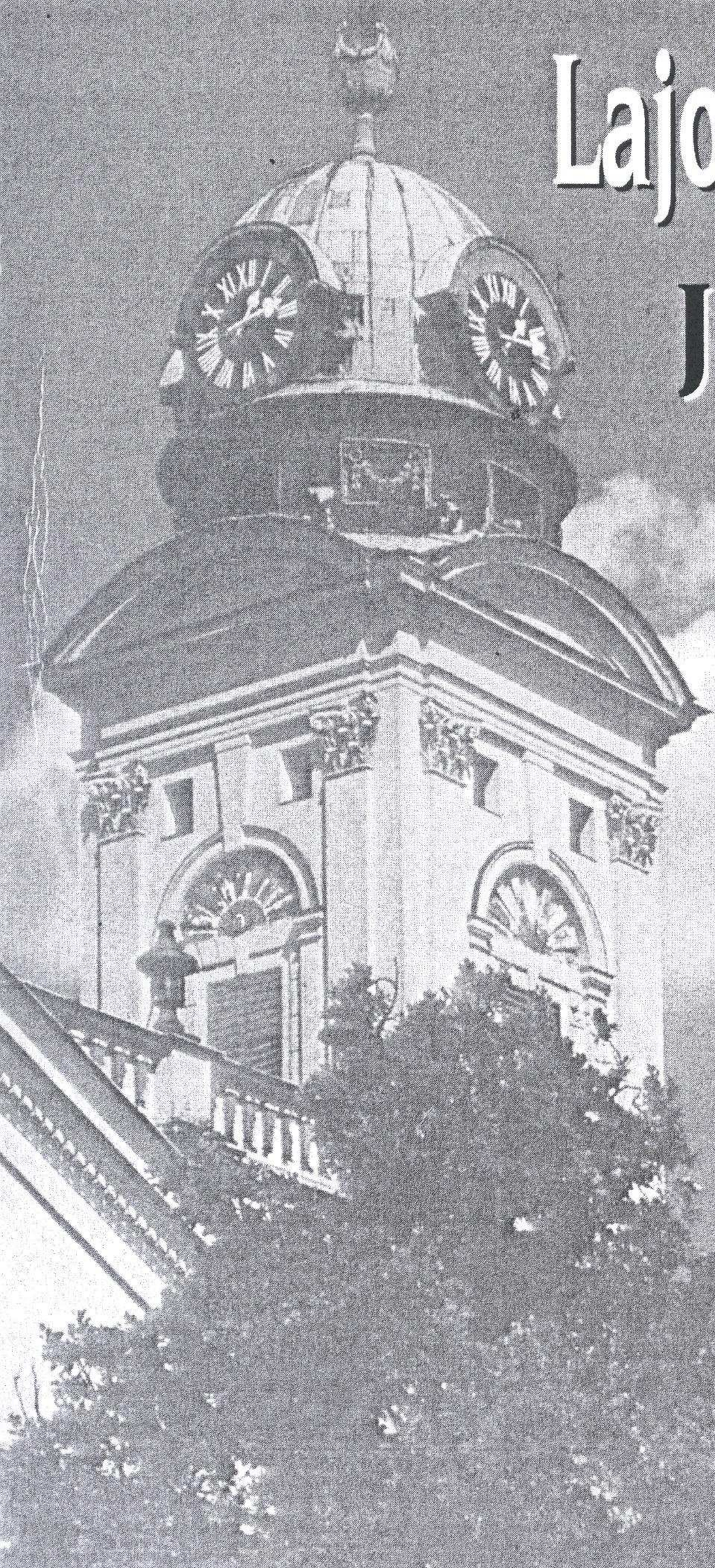
ALLC/ACH '98

CONFERENCE ABSTRACTS

* The Association for Literary and Linguistic Computing
The Association for Computers and the Humanities

Lajos Kossuth University

July, 5-10 1998



Paper

The European LE-PAROLE Project and the Italian Lexical Instantiation

RUIMY N.[1], CORAZZARI O.[2], GOLA E.[2], SPANU A.[2], CALZOLARI N.[1], ZAMPOLLI A.[1]

[1] *Istituto di Linguistica Computazionale del CNR*

[2] *Consorzio Pisa Ricerche*

KEYWORDS: Reusable Resources, Computational Lexicology, Syntax.

CONTACT ADDRESS:
Istituto di Linguistica Computazionale del CNR, v. della Faggiola, 32 - 56126 - Pisa - Italy
E-MAIL: nilda@ilc.pi.cnr.it
FAX: +36 50 556285
PHONE: +36 50 560481

1. INTRODUCTION

The creation of re-usable linguistic data aroused an increasing interest in the NLP community over the last decade. In fact, the lack of large computational lexica and the non-homogeneity of existing resources is a bottleneck for the development of NLP applications. PAROLE, a LE project funded by the CEC DGXIII and executed by the PAROLE Consortium , met this need by building generic and reusable textual and lexical resources in all EU languages.

In this paper we give an overview of the syntactic layer of the Italian Computational Lexicon built in the framework of the PAROLE project, according to the general and language-specific guidelines for lexicon encoding. In addition to — and in compliance with — those specifications, we felt the need, as the encoding process went on, to work out finer-grained indications. We focus here on some of these linguistic and lexicographic criteria which were elaborated using corpus evidence. Then, some information on the syntactic patterns encoded for the main categories, i.e. verbs, nouns and adjectives will allow the lexicon syntactic coverage to be estimated

2. LE-PAROLE PROJECT

PAROLE is the first project producing corpora and lexica in so many languages and built according to the same design principles, same linguistic specifications and representation format. This represents an invaluable achievement, all the more because these resources should constitute a core to be enlarged — following the same principles — at national level.The PAROLE monolingual lexica built for 12 languages[2] consist of 20,000 entries providing morphological and syntactic information. These lexical resources are declarative, theory and application independent, multifunctional and are able to incorporate easily other levels of information or — in virtue of their uniformity — to become multilingual. This approach, which answers the requisite of genericity, explicitness and variability of granularity, ensures a large scale reusability of the produced resources for different application purposes.

Monolingual corpora of at least 20 million words have been developed for 14 languages[3]. Information is encoded following essentially the CES (Corpus Encoding Standard) designed by EAGLES, on the basis of the TEI guidelines. 250,000 running words are tagged and checked at morphosyntactic level, according to language specific instantiations of the EAGLES guidelines. The compatibility of the various corpora is ensured by the adoption of commonly defined criteria for composition, encoding and linguistic annotation.

The PAROLE resources may be used profitably not only by linguists and NLP systems but also as a reference point for different types of research and analyses in the field of Literary Computing and of the Humanities in general.

2.1. LINGUISTIC SPECIFICATIONS AND REPRESENTATIONAL MODEL FOR THE LEXICON

The PAROLE project linguistic specifications [3],[4], are based on EAGLES recommendations for morphosyntactic information and verb syntax [9] and on the extended GENELEX (GENEric LEXicon) model for morphology and for the handling of non-verb categories. They are implemented in the LE-PAROLE model which provides the overall lexicon architecture and the descriptive language [1].

PAROLE lexica consist of three independent though related levels where morphological, syntactic and semantic information is described. A complete lexical entry is thus a progression through the levels of information encoded. Different sets of descriptive objects are available according to the linguistic level to be handled. At syntactic level (figure 1), the basic formal object is the *Syntactic Unit* (*SynU*) defined by a *Base Description* which describes one syntactic behaviour of a morphological unit and, optionally, *Transformed Description(s)* encoding syntactic transformations of the base structure, e.g.: causative alternation. *Base* and *transformed* descriptions of a *SynU* may be linked to each other through the *Frameset* mechanism. Figure 2 shows, in a macro format elaborated in Pisa, that a Description consists of both a *Construction* encoding information about the syntactic context of the word-entry, i.e. a list of canonically ordered *Positions* (P1-P3 in figure 2), and a *Self* describing the properties/restrictions of the entry in the specific subcategorization frame encoded. Each position filler is a syntactic constituent strongly-bound to the entry and is modelled as a bundle of linguistic information

ranging from syntactic function and syntactic realisation to morphosyntactic or lexical inherent properties as well as link, whenever relevant, to other position fillers. These descriptive objects enable the encoding of all PoSs. Furthermore, entries belonging to different PoSs and which have some kind of relationship, such as verbs and deverbal nouns, may be linked through the *TransfUsyn* device.

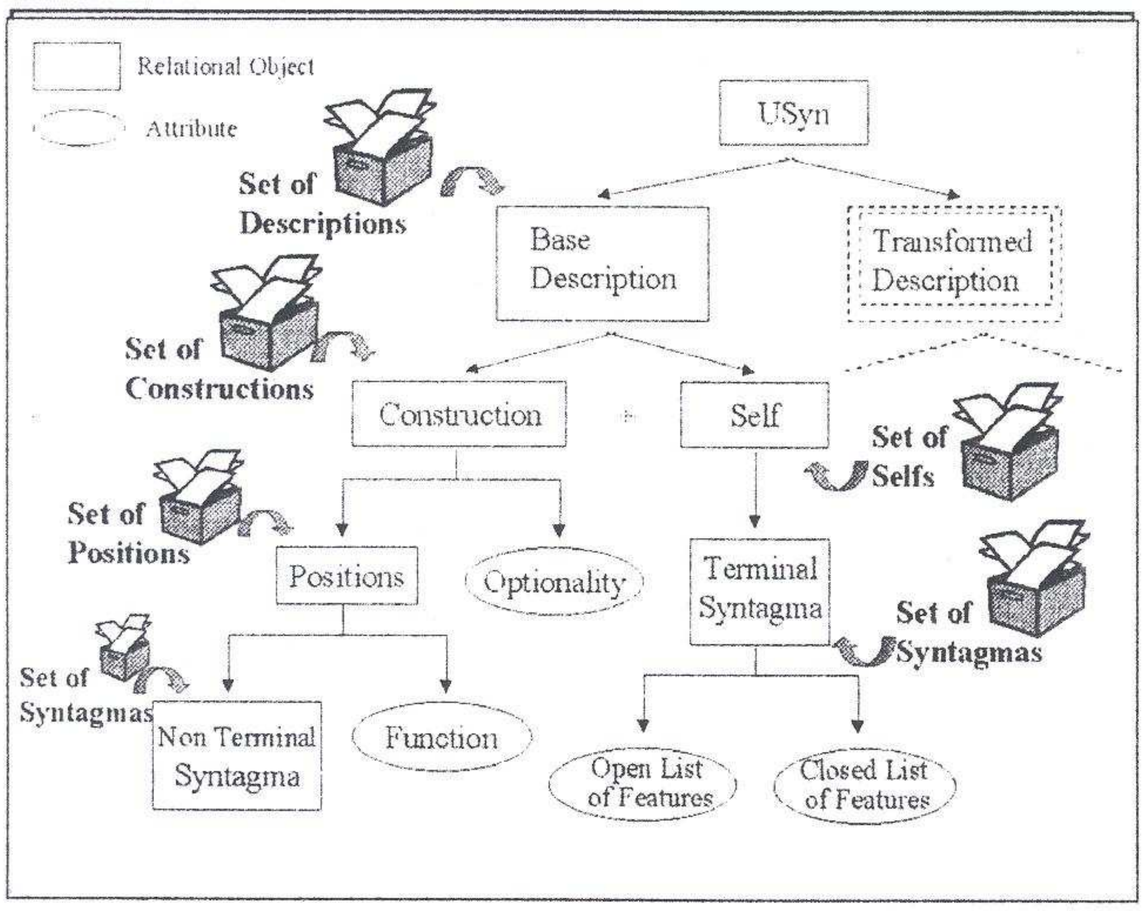


Figure 1: Main objects and attributes at syntactic level.

[SynU: ***chiarire*** [Description: [Construction:
Syntlabel:Clause
P1 :[function:subject]
[cat:np]
[cat:cl] [synsubcat:infcl] [introd:0]
[cat:cl] [synsubcat:thatcl] [mood:sub]
P2[opt:**no**]:[function:object]
[cat:np]
[cat:cl] [synsubcat:thatcl] [mood:**ind**]
[cat:cl] [synsubcat:infcl][introd:di]
[coreference:**I**]
P3[opt:**yes**]:[function:indirectobject]
[cat:pp] [introd:a] [coreference:**I**]]
[SELF: Intervconst: V[func:head][morphsubcat:main]
[aux:**avere**]]].

Figure 2 : Partial representation of an entry information in a macro format (verb *chiarire* 'to clarify').

2.2. THE PAROLE ITALIAN SYNTACTIC LEXICON

The 20,000 one-word lemmas to be encoded in the lexicon were selected among the most frequent words of the ILC Italian Reference Corpus[4] (IRC) [2]. They consist of 3,000 verbs, 3,000 adjectives, 13,000 nouns, 500 adverbs and 500 empty/grammatical words, all belonging to general, modern Italian language.

3.1. ENCODING CRITERIA AND CORPUS EVIDENCE

As it was demonstrated by experiments performed in the field of semantic disambiguation [11], likewise in a lexicon encoding process there is always a part of lexicographer's subjective assessment. Although this phenomenon is less relevant to syntax than to semantics, it is important to endeavour to reduce as much as possible this subjectivity margin so as to maintain a consistent lexical encoding within a language. On the basis of the PAROLE language-specific guidelines for Italian language [5],[6] which provided the general orientation to be followed, and as the encoding process went on, we therefore worked out finer-grained criteria in order to lead as much as possible the encoding task [8]. The elaboration of some of these criteria was guided by a corpus-based study of phenomena relevant to lexical information. Corpus evidence turned out to be sometimes quite different from what we would expect according to grammar and dictionaries indications. In these cases, we tried to keep a balance between encoding attested patterns only and providing an exhaustive description of all theoretically possible structures, even those not likely to be realized. Corpus data were also used a posteriori to check and tuned intuition-based descriptions. In the following, we provide a few cases of corpus evidence usefulness to establishing verb encoding criteria.

3.1.1. SPLITTING CRITERIA

The extent to which lexical entries are to be split into readings (either into different descriptions of the same entry or into different *SynUs*) is a crucial preliminary step in a lexicon building process. At syntactic level, the reading distinction is clearly syntactic-driven. Besides arity and function assignment differences which were patently criterial for distinguishing different readings, every other syntactic structure variation gave rise to a split. Following are examples of those criteria which either emerged from the analysis of the IRC most frequent words contexts or were adjusted upon checking attested data.

- optionality of a complement in one reading only:

It sometimes happens that complements behave differently as to their optionality in two different readings of a lemma. In this case, two structures were encoded to account for such difference. An example of this phenomenon is embodied by the verb *attraversare* 'to cross / to go through'. Its figurative reading, e.g.: *attraversare*

un momento difficile 'to go through a difficult period' was always expressed in the corpus with an object complement whereas the only occurrences of this verb used without object complement turned out to refer exclusively to the literal reading, i.e.: *i bambini attraversano senza guardare* 'children cross (the road) without looking'

- alternative realisations for a complement in only one reading:

Two intuition-based *SynUs* were written to describe the verb *comprendere* 'include/understand': a Np V Np structure corresponding to the 'include' meaning and another for the 'understand' meaning with Np, completive or infinitive clause object. Corpus evidence revealed the very low frequency (0,1%) of use of *comprendere* 'understand' with infinitive clause object. By contrast, wh-clause object (8,5%) and absolute use (4%) structures, not foreseen initially, were added.

- difference in complement introducers:

The verb *esportare* 'to export' was initially encoded as a tetravalent verb with both origin and goal complements, the latter being introduced by preposition *a* 'to'. Corpus data showed the existence of another frequent structure with unexpressed origin and *in*_Pp goal complement, i.e.: *esportare in Francia* 'to export to France'.

- nominalisation of only one reading:

For some verbs, two different polysemies sharing the same syntactic structure were nonetheless split into two *SynUs* since the verb could be nominalized in only one meaning, as the corpus attested for *doppiaggio* 'dubbing', *rialzo* 'increase', e.g.: *rialzare i prezzi* 'to raise prices'; *rialzare la testa* 'to lift up one's head' / *il rialzo dei prezzi* 'the rise in prices'; **il rialzo della testa.*

3.1.2. LEXICALLY-GOVERNED SYNTACTIC CONTEXTS

As to the notion of frame, the PAROLE guidelines propose a rather liberal definition. A distinction is in fact drawn between lexically-governed and non lexically-governed syntactic contexts rather than between arguments and adjuncts. The determination of which constituents are lexically-selected and which are not is therefore a crucial task to the assignment of the adequate arity. Cases of questionable complements for which no consensual solution was found on linguistic intuition's basis were solved by checking the candidate syntactic patterns against corpus evidence. An element occuring quite often in the context of a given lexical unit is likely to be syntactically strongly-bound to the head and hence to be part of its subcategorization frame. Verbs of feelings, for example, were encoded with a cause complement since 26% of the occurrences of 3 among the most frequent verbs belonging to this class: *lamentarsi* 'to complain', *entusiasmarsi* 'to be excited' and *meravigliarsi* 'to marvel' were followed by a *per* or *di*_ Pp 'for/about'.

3.1.3. COMPLEMENT OPTIONALITY

To assess the optionality of verb complements, we considered only 'nuclear', unmarked sentences, since marked ones allow even the omission of complements usually considered as obligatory. For dubious cases, we referred to corpus data. For example, the verb *autorizzare* 'to authorize' was assigned both a divalent and a trivalent pattern (with infinitive oblique complement) with compulsory complements. On the basis of corpus data, the object of the trivalent pattern was marked as optional since 7% of the verb occurrences were used with an unexpressed object, i.e.: *autorizzare a fare qualcosa* 'to authorize to do something'.

On the other hand, a prototypical oriented movement verb such as *andare* 'to go', encoded as a three argument frame (subject included) with origin and goal complements, turned out to occur in only 1% of the cases with an expressed origin complement. As to the goal, it was realized with an infinitive clause *andare a fare qualcosa* 'to go and do something' much more frequently than we thought (25%). In this case, obviously, all the arguments — even the rarely expressed origin one — were encoded, so that partial patterns be realized through the optionality of complements and the origin complement be recognized whenever occurring.

3.1.4. SYNTACTIC REALIZATION OF ARGUMENTS

A frame position may be instantiated by either one or more alternating fillers, each member of the distribution paradigm being a potential syntagmatic realization of the function associated to that position. Splitting of syntactic descriptions in order to encode separately each alternative realization of an argument might be regarded as an advantageous solution for maintaining the syntactic patterns as simple as possible. However this would increase dramatically the lexicon size and, above all, prevent from keeping trace of linguistically-relevant distributional equivalences occurring in real language use, as attested from corpus data. The clustering of the different realizations of each position in a single description, insofar as all their combinations produce grammatical sentences, was therefore adopted as a linguistically sounder solution. The exhaustivity of our descriptions as to the possible realizations of each argument was checked against corpus data for a core set of highly frequent verbs. For verbs such as *chiarire* 'to clarify', *evitare* 'to avoid' or *confermare* 'to confirm' for example, the corpus analysis confirmed the occurrence of structures with both phrasal and clausal subject and object besides an indirect object complement. It appeared however

that statistically only some of these combinations are significantly used. While clausal complements are relatively frequent, clausal subjects are not and the co-occurrence of clauses filling both subject and object slots is quite rare. Anyway, since in our lexicon no weight is assigned to the occurrence of complements, the usefulness of corpus data was in this case a mere examplification of all possible combinations.

4. AN OVERVIEW OF SYNTACTIC PATTERNS

While allowing a very fine-grained description, the PAROLE model enables for a variable granularity beyond a core of mandatory information to be encoded in all 12 lexica. For Italian, all of the general properties shared by whole word classes (e.g. passivization, pro-drop, subject and object pronominalization and postposed subject, for verbs) and derivable by virtue of the membership of a lemma to a class, are assumed to be within the competence of the grammar rather than of the lexicon. Only the idiosyncratic behaviour w.r.t. to grammatical rule's application is therefore stipulated in the lexicon. A syntactic entry encodes the specific properties / restrictions of a lemma and of its subcategorizing elements in a given syntactic structure: it describes the lexically-governed syntactic context. For frame-bearing elements, in particular, each argument is provided with information concerning its optionality, its syntagmatic realization(s) and syntactic function, any relevant constraint at morphosyntactic or lexical level, such as clause type, mood, number and lexical specification of clausal or phrasal complements introducers, as well as any link, whenever relevant, to other arguments, e.g.: agreement and coreference information. Besides, any constraint enforced on the headword, in the specific structure being described, i.e.: auxiliary selection for verbs, mass/count distinction for nouns, pre or postnominal position for adjectives, etc. is encoded.

In the Italian lexicon, besides adverbs and empty words, zero to tetravalent structures of 3,000 intransitive, transitive, pronominal, reflexive and reciprocal verbs were described. Modal verbs as well as subject and object predicate, control, raising, and impersonal constructions were handled. 13,000 concrete and abstract simple nouns as well as deverbal nouns with up to 4 clausal or phrasal arguments were encoded. 3,000 adjectives in predicative and/or attributive use, non predicative uses, non valent and valent adjectives with phrasal and clausal complements and impersonal structures were accounted for.

If we consider as a 'syntactic pattern' the *whole* set of information encoded in an entry, quite a high number of patterns were distinguished, given the amount of entries encoded and the descriptive granularity.

	Verbs	Nouns	Adjectives
syntactic patterns	794	220	95

If we abstract from these highly specified patterns any information on complement optionality as well as on lexical / morphosyntactic constraints on complements realization or on headword and if we consider only the number of arguments, their function and syntagmatic realization, the number of more general structures identified are reduced by around 80% wrt. the specified ones.

tb	Verbs	Nouns	Adjectives
syntactic patterns	174	44	22

5. FINAL REMARKS

The complexity and elevated cost of creation of language resources has induced the scientific community to pay more and more attention to the issue of reusability of existing data. Unfortunately, language resources are too often created from specialized approaches which render the resulting data inadequate for further uses. Resources must in fact meet a certain number of requirements in order to be reusable: the databases produced must be generic, the data uniformly structured and the descriptions precise and explicit.

For the first time, with the LE-PAROLE project, lexica in 12 languages of the European Union have been built according to the same principles. The PAROLE lexica share in fact the same theory and application-independent linguistic specifications, a global architecture, a core set of information content, a descriptive language, management tool and SGML exchange format. PAROLE lexical resources, conceived as generic lexica easily usable by both humans and language processing systems, encode the basic information required by most NLP applications. These characteristics which answer the requisite of genericity, explicitness, and variability of granularity confer a considerable value to the produced resources. They ensure their intra and inter consistency, an easy maintenance of data and a large scale reusability in different theoretical and application frameworks, among which NLP systems development, information retrieval, language learning and machine translation applications. The PAROLE resources, which will be broadly available trough ELRA, are also most relevant to the literary community.

The Italian instantiation of the PAROLE syntactic lexicon presents many interesting aspects. First of all the fact of encoding wide coverage, general and modern language, thanks to a corpus frequency-based lemmas acquisition. Moreover, its

computational nature enables the handling of a large amount of entries as well as a coherent and standardised encoding of information. Partial knowledge, relevant for specific NLP application-dependent models of data and applicative dictionaries can be derived from this repository of information, by mapping the application model from the generic one. Besides, while maintaining its own specificity regarding some encoding decisions as well as a large number of language-specific phenomena whose treatment was partly guided by corpus evidence, it shares with all PAROLE lexica the approach to the conceptual and representational model, the core set of information encoded and the representation type. This membership in a network of European monolingual lexica, which thus implies the possibility of comparison, of creation of multilingual links, and of use in multilingual NLP applications contributes undoubtedly to increase its value.

NOTES

[1] The current Consortium is formed by the following partners: Consorzio Pisa Ricerche (coordinator); GSI-Erli; Institute for Language and Speech Processing (ILSP); Institut d'Estudis Catalans (IEC);University of Birmingham; Institute for Language, Speech and Hearing - Univ. of Sheffield (ILASH); Det Danske Sprog- og Litteraturselskab (DSL); Center for Sprogteknologi (CST); Institiúid Teangeolaíochta Éireann (ITÉ); Dept. of Swedish, Språkdata - Göteborgs Universitet; Department of General Linguistics - University of Helsinki; Instituut voor Nederlandse Lexicologie (INL); Université de Liège BELTEXT; Centro de Linguística da Universidade de Lisboa (CLUL); Instituto de Engenharia de Sistemas e Computadores (INESC); Fundacion Bosch Gimpera Universitat de Barcelona; Institut für Deutsche Sprache (IDS); Institut National de la Langue Française, CNRS (INaLF).

[2]Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

[3]Catalan, Belgian-French, Danish, Dutch, English, Finnish, French, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish.

[4] A textual corpus available at the Pisa Institute of Computational linguistics. This corpus consists of 12,750,000 word tokens from newspapers, magazines, novels, short stories, technical reports, handbooks and scientific texts.

6. REFERENCES

1. GENELEX Consortium (1993) EUREKA PROJECT GENELEX Report on Syntactic Layer, 4.0.

2. Bindi, R., Monachini, M., Orsolini, P. (1991) *Italian Reference Corpus. General Information and Key for Consultation,* ILC-TLN-1991-1, ILC-CNR, Pisa.

3. Calzolari, N., Montemagni, S., Pirrelli, V. (1996) Verb Subcategorization, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa.

4. Flores, S. (1996) Nouns, Adjectives, Adverbs and Prepositions, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Paris, GSI-ERLI.

5. Montemagni, S. & Pirrelli, V. (1996a) Verb Subcategorization in Italian, Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon, Pisa.

6. Montemagni, S. & Pirrelli, V. (1996b) Noun, Adjective, Adverb and Preposition Subcategorization in Italian, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa.

7. Renzi, L. & Salvi G. (Eds.) (1988) *Grande Grammatica italiana di consultazione*, voll. I-III, 1988/91/95, Il Mulino, Bologna.

8. Ruimy, N., Battista, M., Corazzari, O., Gola, E., Spanu A. (1998) Italian Lexicon Documentation, LE-PAROLE, WP3.11, Pisa.

9. Sanfilippo, A. et al. (1996) Subcategorization Standards, Report of the Eagles/Lexicon/ Syntax Group.

10. Schwarze, C. (1995) Grammatik der italienischen Sprache, vol. 2., verbesserte Auflage, Niemeyer Verlag, Tuebingen.

11. Véronis, J., Houitte, V., Jean, C., 1998, Methodology for the construction of test material for the evaluation of word sense disambiguation systems, WLSS98, Pisa.

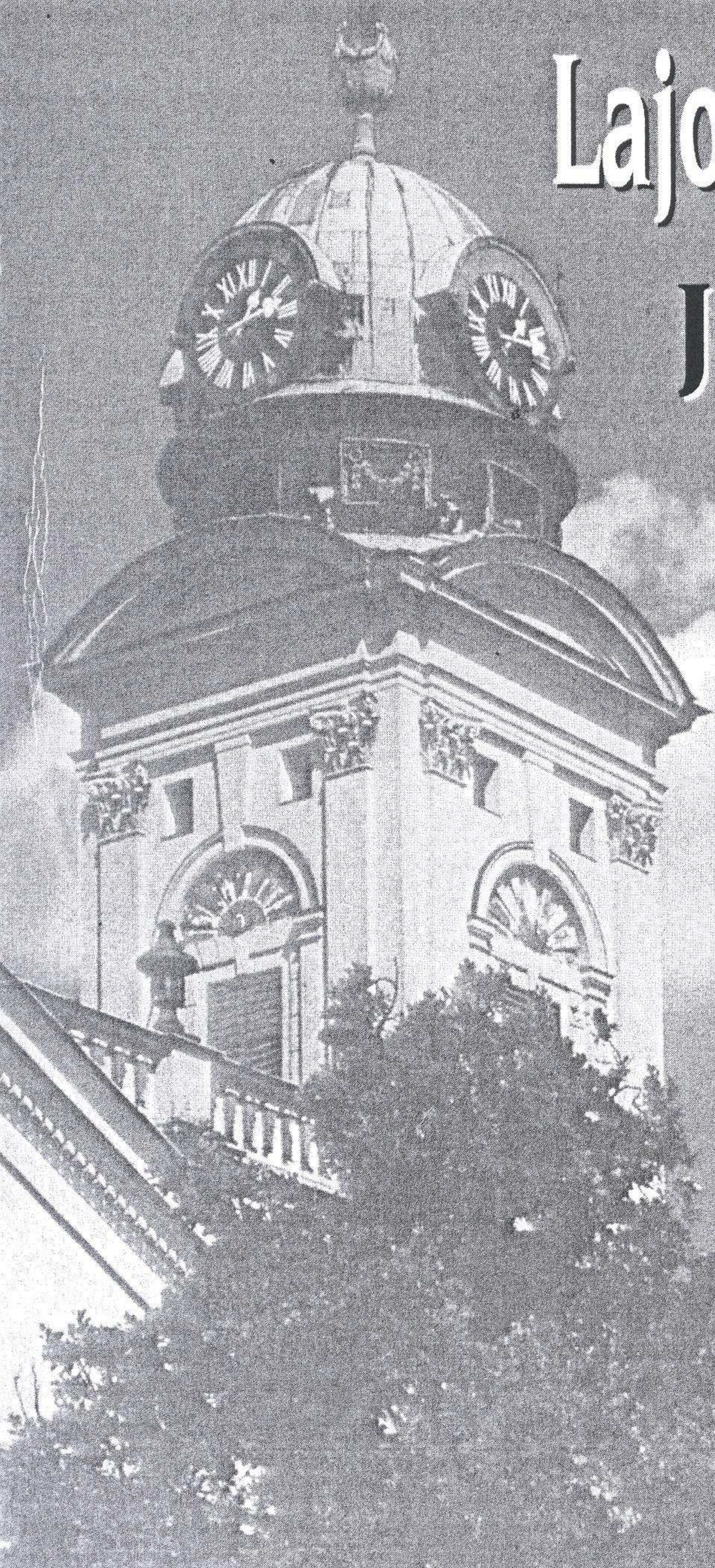
ALLC/ACH '98

CONFERENCE ABSTRACTS

* The Association for Literary and Linguistic Computing
The Association for Computers and the Humanities

Lajos Kossuth University

July, 5-10 1998



Paper

The European LE-PAROLE Project and the Italian Lexical Instantiation

RUIMY N.[1], CORAZZARI O.[2], GOLA E.[2], SPANU A.[2], CALZOLARI N.[1], ZAMPOLLI A.[1]

[1] *Istituto di Linguistica Computazionale del CNR*

[2] *Consorzio Pisa Ricerche*

KEYWORDS: Reusable Resources, Computational Lexicology, Syntax.

CONTACT ADDRESS:
Istituto di Linguistica Computazionale del CNR, v. della Faggiola, 32 - 56126 - Pisa - Italy
E-MAIL: nilda@ilc.pi.cnr.it
FAX: +36 50 556285
PHONE: +36 50 560481

1. INTRODUCTION

The creation of re-usable linguistic data aroused an increasing interest in the NLP community over the last decade. In fact, the lack of large computational lexica and the non-homogeneity of existing resources is a bottleneck for the development of NLP applications. PAROLE, a LE project funded by the CEC DGXIII and executed by the PAROLE Consortium , met this need by building generic and reusable textual and lexical resources in all EU languages.

In this paper we give an overview of the syntactic layer of the Italian Computational Lexicon built in the framework of the PAROLE project, according to the general and language-specific guidelines for lexicon encoding. In addition to — and in compliance with — those specifications, we felt the need, as the encoding process went on, to work out finer-grained indications. We focus here on some of these linguistic and lexicographic criteria which were elaborated using corpus evidence. Then, some information on the syntactic patterns encoded for the main categories, i.e. verbs, nouns and adjectives will allow the lexicon syntactic coverage to be estimated

2. LE-PAROLE PROJECT

PAROLE is the first project producing corpora and lexica in so many languages and built according to the same design principles, same linguistic specifications and representation format. This represents an invaluable achievement, all the more because these resources should constitute a core to be enlarged — following the same principles — at national level.The PAROLE monolingual lexica built for 12 languages[2] consist of 20,000 entries providing morphological and syntactic information. These lexical resources are declarative, theory and application independent, multifunctional and are able to incorporate easily other levels of information or — in virtue of their uniformity — to become multilingual. This approach, which answers the requisite of genericity, explicitness and variability of granularity, ensures a large scale reusability of the produced resources for different application purposes.

Monolingual corpora of at least 20 million words have been developed for 14 languages[3]. Information is encoded following essentially the CES (Corpus Encoding Standard) designed by EAGLES, on the basis of the TEI guidelines. 250,000 running words are tagged and checked at morphosyntactic level, according to language specific instantiations of the EAGLES guidelines. The compatibility of the various corpora is ensured by the adoption of commonly defined criteria for composition, encoding and linguistic annotation.

The PAROLE resources may be used profitably not only by linguists and NLP systems but also as a reference point for different types of research and analyses in the field of Literary Computing and of the Humanities in general.

2.1. LINGUISTIC SPECIFICATIONS AND REPRESENTATIONAL MODEL FOR THE LEXICON

The PAROLE project linguistic specifications [3],[4], are based on EAGLES recommendations for morphosyntactic information and verb syntax [9] and on the extended GENELEX (GENEric LEXicon) model for morphology and for the handling of non-verb categories. They are implemented in the LE-PAROLE model which provides the overall lexicon architecture and the descriptive language [1].

PAROLE lexica consist of three independent though related levels where morphological, syntactic and semantic information is described. A complete lexical entry is thus a progression through the levels of information encoded. Different sets of descriptive objects are available according to the linguistic level to be handled. At syntactic level (figure 1), the basic formal object is the *Syntactic Unit* (*SynU*) defined by a *Base Description* which describes one syntactic behaviour of a morphological unit and, optionally, *Transformed Description(s)* encoding syntactic transformations of the base structure, e.g.: causative alternation. *Base* and *transformed* descriptions of a *SynU* may be linked to each other through the *Frameset* mechanism. Figure 2 shows, in a macro format elaborated in Pisa, that a Description consists of both a *Construction* encoding information about the syntactic context of the word-entry, i.e. a list of canonically ordered *Positions* (P1-P3 in figure 2), and a *Self* describing the properties/restrictions of the entry in the specific subcategorization frame encoded. Each position filler is a syntactic constituent strongly-bound to the entry and is modelled as a bundle of linguistic information

ranging from syntactic function and syntactic realisation to morphosyntactic or lexical inherent properties as well as link, whenever relevant, to other position fillers. These descriptive objects enable the encoding of all PoSs. Furthermore, entries belonging to different PoSs and which have some kind of relationship, such as verbs and deverbal nouns, may be linked through the *TransfUsyn* device.

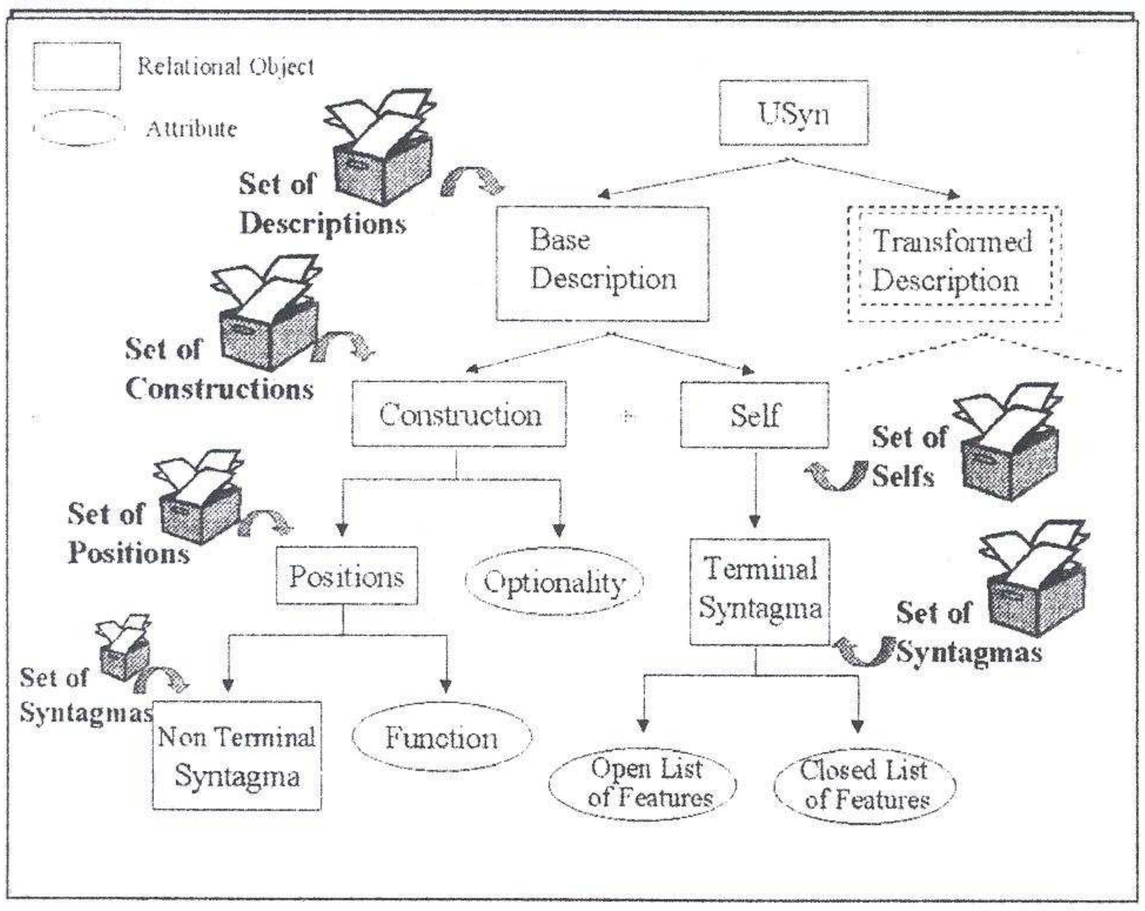


Figure 1: Main objects and attributes at syntactic level.

[SynU: ***chiarire*** [Description: [Construction:
Syntlabel:Clause
P1 :[function:subject]
[cat:np]
[cat:cl] [synsubcat:infcl] [introd:0]
[cat:cl] [synsubcat:thatcl] [mood:sub]
P2[opt:**no**]:[function:object]
[cat:np]
[cat:cl] [synsubcat:thatcl] [mood:**ind**]
[cat:cl] [synsubcat:infcl][introd:di]
[coreference:**I**]
P3[opt:**yes**]:[function:indirectobject]
[cat:pp] [introd:a] [coreference:**I**]]
[SELF: Intervconst: V[func:head][morphsubcat:main]
[aux:**avere**]]].

Figure 2 : Partial representation of an entry information in a macro format (verb *chiarire* 'to clarify').

2.2. THE PAROLE ITALIAN SYNTACTIC LEXICON

The 20,000 one-word lemmas to be encoded in the lexicon were selected among the most frequent words of the ILC Italian Reference Corpus[4] (IRC) [2]. They consist of 3,000 verbs, 3,000 adjectives, 13,000 nouns, 500 adverbs and 500 empty/grammatical words, all belonging to general, modern Italian language.

3.1. ENCODING CRITERIA AND CORPUS EVIDENCE

As it was demonstrated by experiments performed in the field of semantic disambiguation [11], likewise in a lexicon encoding process there is always a part of lexicographer's subjective assessment. Although this phenomenon is less relevant to syntax than to semantics, it is important to endeavour to reduce as much as possible this subjectivity margin so as to maintain a consistent lexical encoding within a language. On the basis of the PAROLE language-specific guidelines for Italian language [5],[6] which provided the general orientation to be followed, and as the encoding process went on, we therefore worked out finer-grained criteria in order to lead as much as possible the encoding task [8]. The elaboration of some of these criteria was guided by a corpus-based study of phenomena relevant to lexical information. Corpus evidence turned out to be sometimes quite different from what we would expect according to grammar and dictionaries indications. In these cases, we tried to keep a balance between encoding attested patterns only and providing an exhaustive description of all theoretically possible structures, even those not likely to be realized. Corpus data were also used a posteriori to check and tuned intuition-based descriptions. In the following, we provide a few cases of corpus evidence usefulness to establishing verb encoding criteria.

3.1.1. SPLITTING CRITERIA

The extent to which lexical entries are to be split into readings (either into different descriptions of the same entry or into different *SynUs*) is a crucial preliminary step in a lexicon building process. At syntactic level, the reading distinction is clearly syntactic-driven. Besides arity and function assignment differences which were patently criterial for distinguishing different readings, every other syntactic structure variation gave rise to a split. Following are examples of those criteria which either emerged from the analysis of the IRC most frequent words contexts or were adjusted upon checking attested data.

- optionality of a complement in one reading only:

It sometimes happens that complements behave differently as to their optionality in two different readings of a lemma. In this case, two structures were encoded to account for such difference. An example of this phenomenon is embodied by the verb *attraversare* 'to cross / to go through'. Its figurative reading, e.g.: *attraversare*

un momento difficile 'to go through a difficult period' was always expressed in the corpus with an object complement whereas the only occurrences of this verb used without object complement turned out to refer exclusively to the literal reading, i.e.: *i bambini attraversano senza guardare* 'children cross (the road) without looking'

- alternative realisations for a complement in only one reading:

Two intuition-based *SynUs* were written to describe the verb *comprendere* 'include/understand': a Np V Np structure corresponding to the 'include' meaning and another for the 'understand' meaning with Np, completive or infinitive clause object. Corpus evidence revealed the very low frequency (0,1%) of use of *comprendere* 'understand' with infinitive clause object. By contrast, wh-clause object (8,5%) and absolute use (4%) structures, not foreseen initially, were added.

- difference in complement introducers:

The verb *esportare* 'to export' was initially encoded as a tetravalent verb with both origin and goal complements, the latter being introduced by preposition *a* 'to'. Corpus data showed the existence of another frequent structure with unexpressed origin and *in*_Pp goal complement, i.e.: *esportare in Francia* 'to export to France'.

- nominalisation of only one reading:

For some verbs, two different polysemies sharing the same syntactic structure were nonetheless split into two *SynUs* since the verb could be nominalized in only one meaning, as the corpus attested for *doppiaggio* 'dubbing', *rialzo* 'increase', e.g.: *rialzare i prezzi* 'to raise prices'; *rialzare la testa* 'to lift up one's head' / *il rialzo dei prezzi* 'the rise in prices'; **il rialzo della testa.*

3.1.2. LEXICALLY-GOVERNED SYNTACTIC CONTEXTS

As to the notion of frame, the PAROLE guidelines propose a rather liberal definition. A distinction is in fact drawn between lexically-governed and non lexically-governed syntactic contexts rather than between arguments and adjuncts. The determination of which constituents are lexically-selected and which are not is therefore a crucial task to the assignment of the adequate arity. Cases of questionable complements for which no consensual solution was found on linguistic intuition's basis were solved by checking the candidate syntactic patterns against corpus evidence. An element occuring quite often in the context of a given lexical unit is likely to be syntactically strongly-bound to the head and hence to be part of its subcategorization frame. Verbs of feelings, for example, were encoded with a cause complement since 26% of the occurrences of 3 among the most frequent verbs belonging to this class: *lamentarsi* 'to complain', *entusiasmarsi* 'to be excited' and *meravigliarsi* 'to marvel' were followed by a *per* or *di*_ Pp 'for/about'.

3.1.3. COMPLEMENT OPTIONALITY

To assess the optionality of verb complements, we considered only 'nuclear', unmarked sentences, since marked ones allow even the omission of complements usually considered as obligatory. For dubious cases, we referred to corpus data. For example, the verb *autorizzare* 'to authorize' was assigned both a divalent and a trivalent pattern (with infinitive oblique complement) with compulsory complements. On the basis of corpus data, the object of the trivalent pattern was marked as optional since 7% of the verb occurrences were used with an unexpressed object, i.e.: *autorizzare a fare qualcosa* 'to authorize to do something'.

On the other hand, a prototypical oriented movement verb such as *andare* 'to go', encoded as a three argument frame (subject included) with origin and goal complements, turned out to occur in only 1% of the cases with an expressed origin complement. As to the goal, it was realized with an infinitive clause *andare a fare qualcosa* 'to go and do something' much more frequently than we thought (25%). In this case, obviously, all the arguments — even the rarely expressed origin one — were encoded, so that partial patterns be realized through the optionality of complements and the origin complement be recognized whenever occurring.

3.1.4. SYNTACTIC REALIZATION OF ARGUMENTS

A frame position may be instantiated by either one or more alternating fillers, each member of the distribution paradigm being a potential syntagmatic realization of the function associated to that position. Splitting of syntactic descriptions in order to encode separately each alternative realization of an argument might be regarded as an advantageous solution for maintaining the syntactic patterns as simple as possible. However this would increase dramatically the lexicon size and, above all, prevent from keeping trace of linguistically-relevant distributional equivalences occurring in real language use, as attested from corpus data. The clustering of the different realizations of each position in a single description, insofar as all their combinations produce grammatical sentences, was therefore adopted as a linguistically sounder solution. The exhaustivity of our descriptions as to the possible realizations of each argument was checked against corpus data for a core set of highly frequent verbs. For verbs such as *chiarire* 'to clarify', *evitare* 'to avoid' or *confermare* 'to confirm' for example, the corpus analysis confirmed the occurrence of structures with both phrasal and clausal subject and object besides an indirect object complement. It appeared however

that statistically only some of these combinations are significantly used. While clausal complements are relatively frequent, clausal subjects are not and the co-occurrence of clauses filling both subject and object slots is quite rare. Anyway, since in our lexicon no weight is assigned to the occurrence of complements, the usefulness of corpus data was in this case a mere examplification of all possible combinations.

4. AN OVERVIEW OF SYNTACTIC PATTERNS

While allowing a very fine-grained description, the PAROLE model enables for a variable granularity beyond a core of mandatory information to be encoded in all 12 lexica. For Italian, all of the general properties shared by whole word classes (e.g. passivization, pro-drop, subject and object pronominalization and postposed subject, for verbs) and derivable by virtue of the membership of a lemma to a class, are assumed to be within the competence of the grammar rather than of the lexicon. Only the idiosyncratic behaviour w.r.t. to grammatical rule's application is therefore stipulated in the lexicon. A syntactic entry encodes the specific properties / restrictions of a lemma and of its subcategorizing elements in a given syntactic structure: it describes the lexically-governed syntactic context. For frame-bearing elements, in particular, each argument is provided with information concerning its optionality, its syntagmatic realization(s) and syntactic function, any relevant constraint at morphosyntactic or lexical level, such as clause type, mood, number and lexical specification of clausal or phrasal complements introducers, as well as any link, whenever relevant, to other arguments, e.g.: agreement and coreference information. Besides, any constraint enforced on the headword, in the specific structure being described, i.e.: auxiliary selection for verbs, mass/count distinction for nouns, pre or postnominal position for adjectives, etc. is encoded.

In the Italian lexicon, besides adverbs and empty words, zero to tetravalent structures of 3,000 intransitive, transitive, pronominal, reflexive and reciprocal verbs were described. Modal verbs as well as subject and object predicate, control, raising, and impersonal constructions were handled. 13,000 concrete and abstract simple nouns as well as deverbal nouns with up to 4 clausal or phrasal arguments were encoded. 3,000 adjectives in predicative and/or attributive use, non predicative uses, non valent and valent adjectives with phrasal and clausal complements and impersonal structures were accounted for.

If we consider as a 'syntactic pattern' the *whole* set of information encoded in an entry, quite a high number of patterns were distinguished, given the amount of entries encoded and the descriptive granularity.

	Verbs	Nouns	Adjectives
syntactic patterns	794	220	95

If we abstract from these highly specified patterns any information on complement optionality as well as on lexical / morphosyntactic constraints on complements realization or on headword and if we consider only the number of arguments, their function and syntagmatic realization, the number of more general structures identified are reduced by around 80% wrt. the specified ones.

tb	Verbs	Nouns	Adjectives
syntactic patterns	174	44	22

5. FINAL REMARKS

The complexity and elevated cost of creation of language resources has induced the scientific community to pay more and more attention to the issue of reusability of existing data. Unfortunately, language resources are too often created from specialized approaches which render the resulting data inadequate for further uses. Resources must in fact meet a certain number of requirements in order to be reusable: the databases produced must be generic, the data uniformly structured and the descriptions precise and explicit.

For the first time, with the LE-PAROLE project, lexica in 12 languages of the European Union have been built according to the same principles. The PAROLE lexica share in fact the same theory and application-independent linguistic specifications, a global architecture, a core set of information content, a descriptive language, management tool and SGML exchange format. PAROLE lexical resources, conceived as generic lexica easily usable by both humans and language processing systems, encode the basic information required by most NLP applications. These characteristics which answer the requisite of genericity, explicitness, and variability of granularity confer a considerable value to the produced resources. They ensure their intra and inter consistency, an easy maintenance of data and a large scale reusability in different theoretical and application frameworks, among which NLP systems development, information retrieval, language learning and machine translation applications. The PAROLE resources, which will be broadly available trough ELRA, are also most relevant to the literary community.

The Italian instantiation of the PAROLE syntactic lexicon presents many interesting aspects. First of all the fact of encoding wide coverage, general and modern language, thanks to a corpus frequency-based lemmas acquisition. Moreover, its

computational nature enables the handling of a large amount of entries as well as a coherent and standardised encoding of information. Partial knowledge, relevant for specific NLP application-dependent models of data and applicative dictionaries can be derived from this repository of information, by mapping the application model from the generic one. Besides, while maintaining its own specificity regarding some encoding decisions as well as a large number of language-specific phenomena whose treatment was partly guided by corpus evidence, it shares with all PAROLE lexica the approach to the conceptual and representational model, the core set of information encoded and the representation type. This membership in a network of European monolingual lexica, which thus implies the possibility of comparison, of creation of multilingual links, and of use in multilingual NLP applications contributes undoubtedly to increase its value.

NOTES

[1] The current Consortium is formed by the following partners: Consorzio Pisa Ricerche (coordinator); GSI-Erli; Institute for Language and Speech Processing (ILSP); Institut d'Estudis Catalans (IEC);University of Birmingham; Institute for Language, Speech and Hearing - Univ. of Sheffield (ILASH); Det Danske Sprog- og Litteraturselskab (DSL); Center for Sprogteknologi (CST); Institiúid Teangeolaíochta Éireann (ITÉ); Dept. of Swedish, Språkdata - Göteborgs Universitet; Department of General Linguistics - University of Helsinki; Instituut voor Nederlandse Lexicologie (INL); Université de Liège BELTEXT; Centro de Linguística da Universidade de Lisboa (CLUL); Instituto de Engenharia de Sistemas e Computadores (INESC); Fundacion Bosch Gimpera Universitat de Barcelona; Institut für Deutsche Sprache (IDS); Institut National de la Langue Française, CNRS (INaLF).

[2]Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

[3]Catalan, Belgian-French, Danish, Dutch, English, Finnish, French, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish.

[4] A textual corpus available at the Pisa Institute of Computational linguistics. This corpus consists of 12,750,000 word tokens from newspapers, magazines, novels, short stories, technical reports, handbooks and scientific texts.

6. REFERENCES

1. GENELEX Consortium (1993) EUREKA PROJECT GENELEX Report on Syntactic Layer, 4.0.

2. Bindi, R., Monachini, M., Orsolini, P. (1991) *Italian Reference Corpus. General Information and Key for Consultation,* ILC-TLN-1991-1, ILC-CNR, Pisa.

3. Calzolari, N., Montemagni, S., Pirrelli, V. (1996) Verb Subcategorization, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa.

4. Flores, S. (1996) Nouns, Adjectives, Adverbs and Prepositions, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Paris, GSI-ERLI.

5. Montemagni, S. & Pirrelli, V. (1996a) Verb Subcategorization in Italian, Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon, Pisa.

6. Montemagni, S. & Pirrelli, V. (1996b) Noun, Adjective, Adverb and Preposition Subcategorization in Italian, *Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa.

7. Renzi, L. & Salvi G. (Eds.) (1988) *Grande Grammatica italiana di consultazione*, voll. I-III, 1988/91/95, Il Mulino, Bologna.

8. Ruimy, N., Battista, M., Corazzari, O., Gola, E., Spanu A. (1998) Italian Lexicon Documentation, LE-PAROLE, WP3.11, Pisa.

9. Sanfilippo, A. et al. (1996) Subcategorization Standards, Report of the Eagles/Lexicon/ Syntax Group.

10. Schwarze, C. (1995) Grammatik der italienischen Sprache, vol. 2., verbesserte Auflage, Niemeyer Verlag, Tuebingen.

11. Véronis, J., Houitte, V., Jean, C., 1998, Methodology for the construction of test material for the evaluation of word sense disambiguation systems, WLSS98, Pisa.