THE *PAROLE* MODEL AND THE ITALIAN SYNTACTIC LEXICON

NILDA RUIMY, ORNELLA CORAZZARI,
ELISABETTA GOLA, ANTONIETTA SPANU,
NICOLETTA CALZOLARI, ANTONIO ZAMPOLLI

Abstract - *This paper presents an overview of a large scale Syntactic Computational Lexicon of Italian. This lexicon was elaborated in the framework of the EC funded LE-PAROLE project, which developed core, generic and re-usable written language resources in 12 EU languages. All monolingual lexica were built according to the same design principles, same linguistic specifications and representation format. The PAROLE Italian lexicon is representative of modern Italian language use. The entries were selected on a frequency basis from the ILC Corpus and the syntactic structures encoded were partly inferred from their contexts of occurrence. Both the general structure of a PAROLE lexicon and the specificity of its Italian instantiation are presented. Some language-specific linguistic and lexicographic options concerning crucial issues to a lexicon building process are illustrated. An overview of the syntactic structures encoded for verbs, nouns and adjectives allows lexicon syntactic coverage as well as description fine-grainedness to be estimated.*

Keywords - *lexicon, morphology, syntax, reusability*

1. INTRODUCTION

The development of re-usable linguistic data has aroused an increasing interest in the NLP community over the last decade due to the complexity and huge cost of creation of new language resources. 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 carried out by the PAROLE Consortium[1], met this need by building core generic and reusable textual and lexical resources in all EU languages.

The LE-PAROLE project is the second part of a three-phase program. This program included the MLAP PP-PAROLE project (Zampolli, 1994) for the elaboration of the linguistic specifications to be followed in the LE-PAROLE project which developed core, generic, multifunctional and re-usable harmonised written language resources, i.e. corpora for 14 languages and morphological and syntactic electronic lexica for 12 languages of the European Union. The third phase is the LE-SIMPLE project (Ruimy *et al.*, this volume), which has just been concluded, aimed at the addition of a layer of semantic information to PAROLE lexica.

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 now constitute a core to be enlarged following the same principles at the national level[2].

The aim of favouring the reusability of the resources can be achieved by relying on the most generic lexical architecture and descriptive language. These lexical resources are in fact declarative, theory and application independent, multifunctional and are able to easily incorporate other levels of information or, in virtue of their uniformity, to become multilingual. The lexica follow the PAROLE project linguistic specifications (Calzolari *et al.*,

[1] During the project, the Consortium was 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); Fundación Bosch Gimpera Universitat de Barcelona; Institut für Deutsche Sprache (IDS); Institut National de la Langue Française, CNRS (INaLF).

[2] CLIPS Italian national project (Corpora e Lessici di Italiano Parlato e Scritto).

1996; Flores, 1996) which are based on EAGLES recommendations (Sanfilippo, 1996) for morphosyntactic information and verb syntax, and on the extended GENELEX (GENEric LEXicon) model for morphology and for the handling of non-verbal categories (GENELEX CONSORTIUM, 1993). These guidelines are implemented in the LE-PAROLE model which provides the overall lexicon architecture and the descriptive language. The use of a common DTD for morphological and syntactic layers, of the SGML exchange format and of a common software tool for data management, an extension of the GENELEX tools, guarantees both the conformity of the twelve lexica to the model and their interconsistency. This approach, which meets the requisites of genericity, explicitness and variability of granularity, ensures a large scale reusability of the produced resources for different application purposes. The 12 PAROLE monolingual lexica consist of 20,000 entries providing morphological and syntactic information.

This paper presents an overview of the PAROLE Italian Syntactic Lexicon as an instantiation of a lexicon built according to the PAROLE model. We illustrate some language-specific linguistic and lexicographic options that were taken up for Italian and guided the syntactic encoding. An overview of the syntactic structures encoded for verbs, nouns and adjectives allows the lexicon syntactic coverage to be estimated.

2. REPRESENTATIONAL MODEL

The formal representation model adopted in PAROLE lexica is the Entity/Relationship model. Such a model enables a non-redundant and intuition-based representation of data. The Entity/Relation model is implemented in the PAROLE lexica through an SGML document type definition (DTD) that defines the structure of the different objects relevant for each representational level, their legal features and co-occurrence restrictions and the relationships holding among objects. An object describing a pattern shared by a set of entries is defined and named once and for all. The object identifier is then simply assigned to all relevant lexical items sharing that pattern, with no need to stipulate the pattern properties once more in the lexical entries.

The modularity of the PAROLE-SIMPLE lexical model is such that the information encoded on the morphological, syntactic and semantic descriptive levels is independent of each other although the three levels are connected (fig. 1). A complete entry is a progression through the levels of information encoded. A morphological unit (MU) is linked to one or more syntactic units which share the same morphological information.

A syntactic unit has thus access to its morphological information through the link to the morphological unit it is associated with. A syntactic unit, on the other hand, is associated with one or more semantic units, depending on the number of meanings which can be distinguished for a single syntactic structure of a lemma. Each semantic unit, in its turn, has access to the syntactic information of the entry it is linked with. The PAROLE model also provides for multilingual links between semantic units.

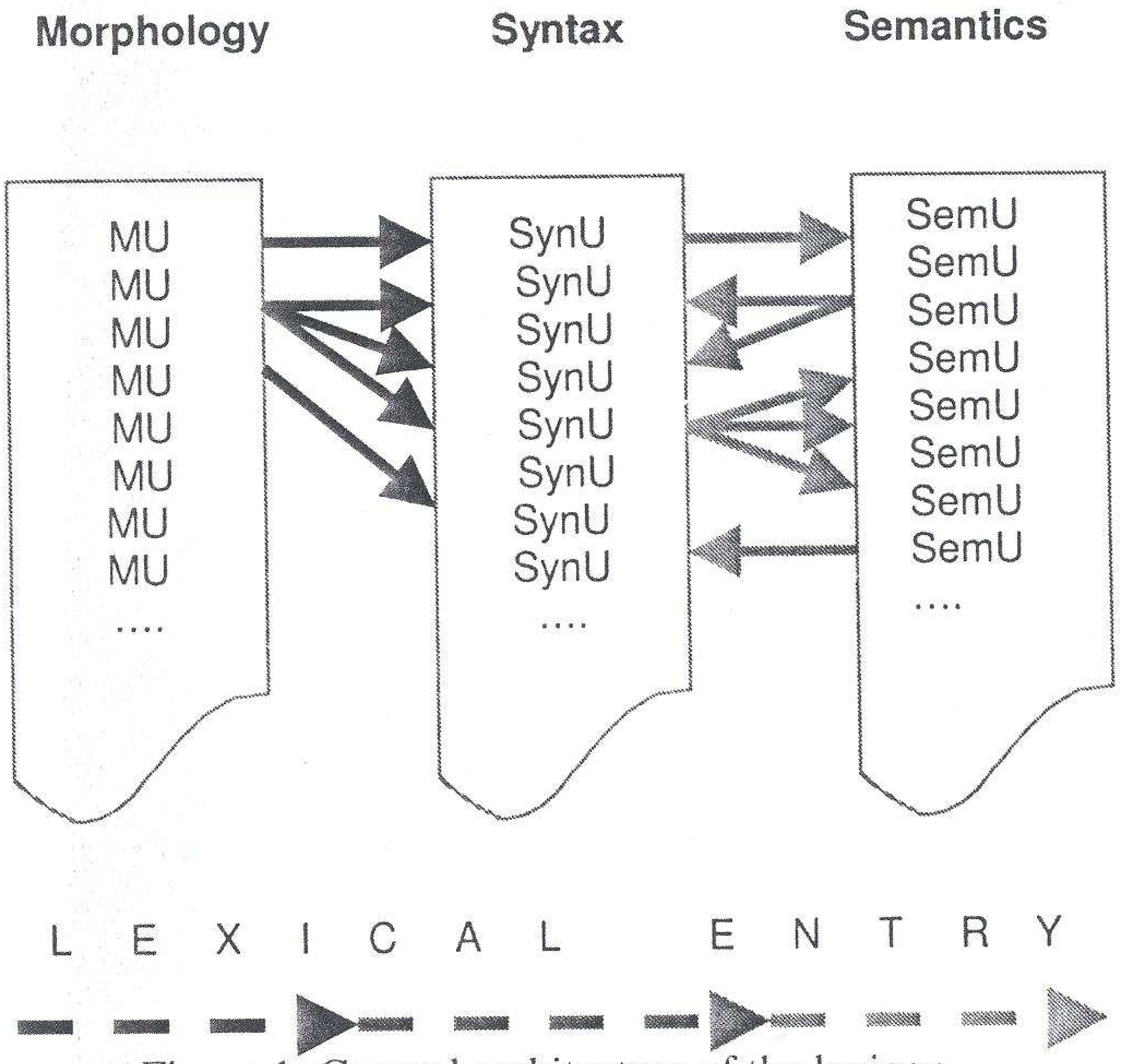

Figure 1. General architecture of the lexicon

For each linguistic level, the descriptive structure consists of an interaction of basic and complex descriptive elements, the complex elements described by more basic ones. Most of the descriptive elements are shared by various elements of higher level. Different sets of descriptive objects are available according to the linguistic level to be handled. At the syntactic level (fig. 2), the basic formal object is the *Syntactic Unit* (*SynU* or *Usyn*) defined by a *Base Description* describing one syntactic behaviour of a morphological unit and, optionally, by *Transformed Description(s)* encoding closely related surface syntactic transformations of the base structure, e.g. causative-inchoative alternation.

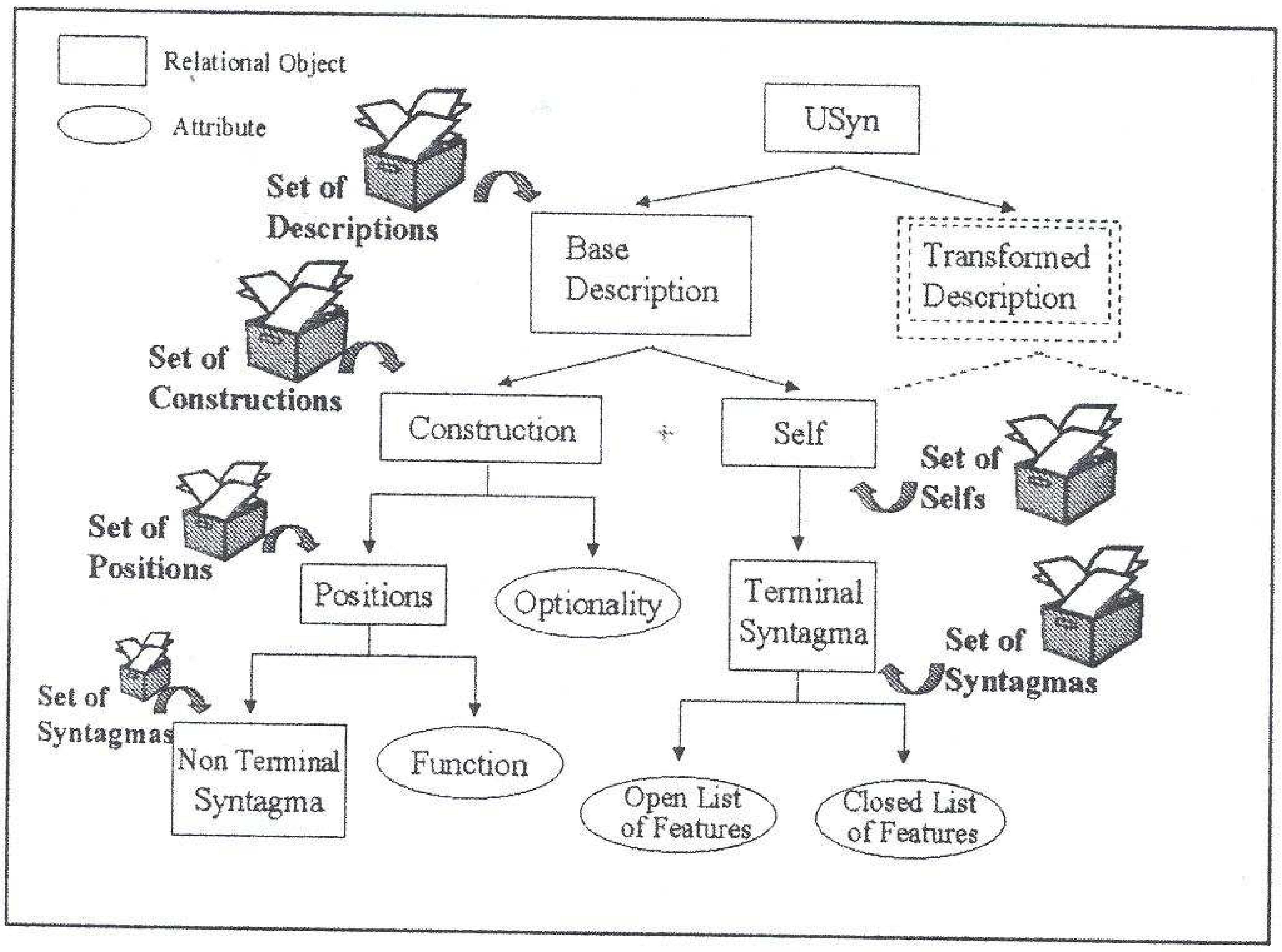

Figure 2. Main objects and attributes at the syntactic level

A *Description*, or frame, consists of two main objects: a *Construction* providing information about the syntactic context of the lexical entry - i.e. a canonically ordered[3] list of Positions, or frame slots, which describe the syntactic constituents and their restrictions - and a *Self* where the lemma properties/restrictions, in the specific subcategorization frame described (e.g. selected auxiliary for a verb reading), are stored.

[3] The syntactic function is criterial for position ordering, which may therefore be different from surface order.

Positions are provided with linguistic information identifying the position occupant. Each position filler is a syntactic constituent strongly-bound to the word-entry and is modelled as a bundle of linguistic information ranging from syntactic function and syntactic realization (expressed in terms of non-terminal or terminal syntactic category) to morphosyntactic or lexical inherent properties as well as any link, whenever relevant, to other position fillers.

The PAROLE model also makes provision for relating lexical information throughout the lexicon by means of two descriptive devices. Within a *SynU*, the different positions of a Base and a *Transformed* description may be linked to each other through the *Frameset* mechanism which relates slot fillers of a frame alternation shared by a large number of entries. On the other hand, the relationship between Syntactic units encoding different parts of speech (e.g.: a verb and its nominalization) may be captured through the *TransfUsyn* device.

The PAROLE model enables a very fine-grained description to be performed. However, beyond a core set of mandatory information which guarantees a high level of uniformity in the information content of all lexica, the level of descriptive granularity is at the discretion of each partner in so far as the description performed meets the model requirements.

The morphological level provides information concerning morphosyntactic category and sometimes subcategory, morphological features of gender, number, etc., as well as inflectional pattern and variants. Derivation, affixes and compound units, abridged forms and usage values may optionally be encoded. The syntactic level allows for the encoding of basic and more refined information concerning the syntactic behaviour of a morphological unit. As regards basic, and hence mandatory information, subcategorization pattern and relevant properties of slot fillers (syntactic realization, function and basic control information) as well as the properties/restrictions of the word entry in the particular syntactic construction described are to be provided. Lexical alternations, refined control, insertion context, slot fillers thematic roles and semantic restrictions are on the other hand optional.

3. THE PAROLE ITALIAN SYNTACTIC LEXICON

3.1. *Selection of lexical units*

The Italian PAROLE Lexicon was mainly built, at the morphological level, through the conversion of pre-existing ILC resources (inflectional models and encoded lemmas). It consists of 60,000 lemmas encoded with all the relevant mandatory information.

At the syntactic level, it is composed of 20,051 one-word entries selected among the most frequent words of the ILC Italian Reference Corpus (IRC)[4] and therefore belonging to general modern language. The selected lemmas belong to the following parts of speech: verbs (3,120), nouns (13,212), adjectives (2,997), adverbs (562), and empty words (160).

3.2. *Encoding and reading distinction criteria*

As was demonstrated by experiments performed in the field of semantic disambiguation, likewise in a lexicon encoding process there is always a part of the lexicographer's subjective assessment. Although this phenomenon is less relevant to syntax than to semantics, it is important 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 the Italian language (Montemagni and Pirrelli, 1996a-b) which provided the general orientation to be followed, and as the encoding process went on, we worked out finer-grained criteria in order to control as much as possible the encoding task (Ruimy *et al.,* 1998a). 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 have expected according to grammar and dictionary indications. In these cases, we tried to keep a balance between encoding attested patterns only and providing an exhaustive description of all theoretically possible

[4] A textual corpus available at the Pisa Istituto di Linguistica Computazionale. This corpus consists of 12,750,000 word tokens from newspapers, magazines, novels, short stories, technical reports, handbooks and scientific texts (Bindi *et al.*, 1991).

structures, even those not likely to be realized. Corpus data was also used to check and tune intuition-based descriptions.

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. As a general rule, both redundancy and over-powerful gatherings were avoided. At the syntactic level, reading distinction is clearly syntactic-driven and semantic considerations were therefore accounted for only in so far as they had consequences at the syntactic level. Arity and function assignment differences (fig. 3) were criterial for the splitting of entries.

Disporre i libri negli scaffali (to put books on the shelves)	/	*Disporre di due auto* (to have 2 cars at one's disposal)
La leggerezza di una piuma (the lightness of a feather)	/	*Ha commesso una leggerezza* (he was too lenient)
Uomo appassionato di musica (man who has a passion for music)	/	*Amante appassionato* (passionate lover)

Figure 3

Other syntactic structure[5] variations, such as the following ones, gave rise to a split:

- optionality of a complement in one reading only[6]: 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. This phenomenon is especially observed in non-literal senses which generally obligatorily require a complement (fig. 4):

[5] By syntactic structure, we intend the complete information about the behaviour of a lemma in a given reading, i.e. both its properties/restrictions and its complementation pattern including the lexical specification of complement introducers, control, agreement and mood restrictions.

[6] Round brackets, in these examples, indicate the optionality of a complement.

Forare (una gomma)	/	**forare (un biglietto)*[7]
(to burst a tyre)		(to punch a ticket)
attraversare (la strada)	/	**attraversare (un periodo difficile)*
(to cross (the road))		(to go through a difficult period)
Uomo prigioniero (dei nemici)	/	** uomo prigioniero (delle proprie idee)*
(man prisoner of the enemies)		(man prisoner of his own ideas)

Figure 4

- alternative realization of a complement in one reading only (fig. 5, 1.b.):

1.a *Luca evita Maria*
(Luca avoids Maria)

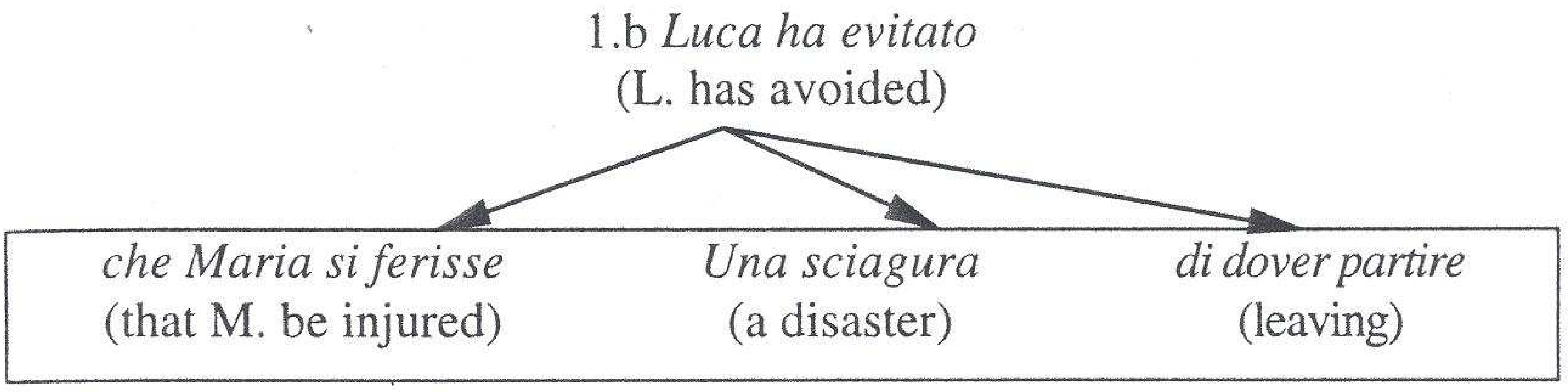

Figure 5

- different behaviour w.r.t. nominalization for homographic or polysemic verbs: for some verbs, two different polysemic or homographic readings sharing the same syntactic structure were nonetheless split into two SynUs since the verb could be nominalized in only one meaning (fig. 6):

[7] The RHS examples are not acceptable without a complement.

Doppiare un film		_il doppiaggio di un film_
(to dub a film)	/	(the dubbing of a film)
rialzare i prezzi		_il rialzo dei prezzi_
(to raise prices)		(the rise in prices)

vs.

Doppiare il Capo Horn		* _il doppiaggio del Capo Horn_
(to round the Cap Horn)	/	
rialzare la testa		*_il rialzo della testa_
(to lift up one's head)		

Figure 6

4. A PAROLE LEXICAL ENTRY: INFORMATION CONTENT

A PAROLE syntactic entry encodes the specific properties / restrictions of a lemma and of its subcategorizing elements in a given syntactic construction: it describes the lexically-governed syntactic context. By contrast, all the general properties shared by whole word classes (e.g. for verbs, passivization, pro-drop, subject and object pronominalization and postposed subject), and which can be derived 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 idiosyncratic behaviours with respect to grammatical rule application are therefore stipulated in the lexicon.

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 European lexica (LE-PAROLE, 1995). In the Italian lexicon, all mandatory information (i.e.: subcategorization pattern with function, PoS and relevant features for each argument as well as constraints on the word entry) has been encoded. As regards optional information, diathesis alternations for verbs, derivational links between verbs and nouns, mass/count feature for nouns and insertion context for adjectives were handled.

As shown in figure 7[8], for frame-bearing units, each slot in the subcategorization frame is associated with a bundle of information about the syntagmatic realization and syntactic function of the argument and its optionality. Besides, any argument may be constrained at the morphosyntactic, syntactic or lexical levels by means of features. Such features may indicate clause type, mood, number, agreement, lack of determination, lexical specification of clausal or phrasal complement introducers, and any link, whenever relevant, to other slot fillers. Control information is also encoded by means of features at the position filler level while for control on infinitive clause an additional frame level feature specifies the type of construction at hand, i.e.: subject control, raising, etc. On the other hand, constraints enforced on the headword, in the particular reading described, namely auxiliary selection or impersonal construction for verbs, mass/count distinction for nouns, pre- or postnominal position for attributive readings of adjectives, etc., are expressed outside the complement description, in the SELF.

```
[SynU: chiarire (to clarify)
 [Description:
 [Construction:
 [Syntlabel:Clause]
P0          :[function:subject]
                    [cat:np]
                    [cat:cl] [synsubcat:infcl] [introd:0]
                    [cat:cl] [synsubcat:thatcl] [ mood:sub]
P1[opt:no]:[function:object]
                    [cat:np]
                    [cat:cl] [syn_sbcat:thatcl] [mood:ind]
                    [cat:cl] [synsubcat:infcl] [introd:di] [coreference:I]
P2[opt:yes]:[function:indirectobject]
                    [cat:pp] [introd:a] [coreference:I]]
[SELF: Intervconst: V [func:head] [morphsubcat:main] [aux:avere]]].
```

Figure 7. Partial representation of a verb entry in a working format

[8] In this partial representation of an entry, the information is modelled in an internal intermediate format worked out in Pisa, which was used to encode syntactic structures by means of macros.

4.1. *Lexically-governed syntactic context*

Predicate arity has been considered as language-specific parameterizable information within the PAROLE project. In the Italian lexicon, it has been limited to four arguments. As to which elements should be considered as subcategorized for, the PAROLE Linguistic Specifications propose a somewhat liberal definition of frame. A distinction is in fact drawn between lexically-governed and non lexically-governed syntactic contexts rather than between arguments and adjuncts. A position filler is considered as syntactically strongly-bound provided that it is lexically selected by the head. This excludes for example the specification of adverbial phrases in verb frames unless they are specifically required by these verbs. On the other hand, syntactically strongly-bound elements may be either arguments or adjuncts: they are referred to as complements as long as they are lexically governed (Calzolari *et al.*, 1996). The determination of which constituents are lexically-selected and which are not is therefore a crucial, and sometimes tricky, task to the assignment of the adequate arity. Cases of questionable complements emerged for which no consensual solution was found on the basis of our linguistic intuition. Those cases were solved by checking the candidate syntactic frames against corpus evidence. An element occurring 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.

4.2. *Complement optionality*

Once identified, complements are marked with regard to obligatoriness. Complement optionality was assessed for verbs by considering nuclear, unmarked contexts since marked ones allow even the omission of complements usually considered obligatory. For dubious cases, we referred to corpus data. Optionality of noun complements, which is a more controversial issue, was on the other hand less easy to determine. In fact, noun complements are often assumed to be all optional. In reality, a distinction is to be made between simple and deverbal nouns. Following the linguistic tradition, simple noun complements were generally considered as optional. As for deverbals, by-phrases - which occur quite rarely in the IRC corpus - were encoded as optional, while object-like complements in complex event nominals were marked as obligatory. Deverbal and simple noun complements were encoded as obligatory in figurative meanings, e.g.: *la chiave del problema* (the key to the problem); *la fioritura delle arti* (the flourishing of arts) in order to stress the different syntactic behaviour of literal and non-literal senses of a lexical unit.

4.3. *Syntactic functions*

The assignment of syntactic functions to each position occupant was not always straightforward. The decision of assigning to some verb complements an oblique/prepositional object or an adverbial syntactic function was sometimes problematic since a clear-cut borderline between these two functions is often hard to draw. Some criteria[9] for their assignment were therefore established in order to ensure coding consistency[10]. As documented in Ruimy *et al.* (1997), the prepositional object function was ascribed to PPs not substitutable by adverbs and whose interrogative form was built with personal pronouns (table 1). The semantic role of these complements may be beneficiary, instrument or cause. PPs introduced by strongly-bound prepositions, as in *dedicarsi a qualcosa* (to devote oneself to something) were also attributed this function. The adverbial function, on the other hand, was assigned to PP complements in alternative distribution with adverbs and whose interrogative form was built with an interrogative adverb (table 2). These complements bear a semantic information of manner, measure, time, location or direction. This syntactic label

[9] Maria Gronostaj, from the Swedish Parole team, gave a most relevant contribution to the statement of these criteria.

[10] Nonetheless, the assignment of the 'adverbial' or 'prepobj' function was sometimes quite problematic. In fact, in many cases, the complement at hand fulfilled the requisites for the assignment of both function labels, i.e. that the complement be introduced by a strongly-bound preposition (for 'prepobj' label) and that the phrase convey a semantic content of Manner, Measure, Time or Location (for 'adverbial' label).

was also assigned to PPs or adverbial phrases which, together with the verb, confer an idiomatic meaning to the phrase, i.e.: *saltare agli occhi* (to jump out at someone).

θ-role	oblique/prepositional object
beneficiary	*Fare qualcosa per qualcuno* (per chi?) (to do something for s.o.) (for who?)
instrument	*Colpire con un pugno* (con che cosa?) (to strike s.o. with a blow) (with what?)
cause	*Lamentarsi per qualcosa* (per che cosa?) (to complain about s.th.) (about what?)

Table 1. Assignment of the 'prepositional object' function to verb complements

θ-role	adverbial
manner	*Circolare a piedi* (come?) (to go on foot) (how?)
measure	*Allungare di un metro* (di quanto?) (to lengthen by one metre) (how much?)
time	*Iniziare presto/alle otto* (quando?) (to start early/at 8 o'clock) (when?)
locative/ directional	*Vivere a Parigi* (dove?) (to live in Paris) *andare a Pisa* (to go to Pisa) (where?)

Table 2. Assignment of the 'adverbial' function to verb complements

As far as nouns and adjectives are concerned, no specific syntactic function was assigned either to simple noun or adjective complements. By contrast, deverbal noun complements were implicitly ascribed syntactic functions, through the linking of slots in their frame to the corresponding verbal base frame slots.

4.4. *Paradigmatically-related position occupants*

Each frame position may be instantiated by one or more paradigmatically-related alternating slot fillers, each member of the distribution paradigm being a potential syntagmatic realization of the function associated to that position. Splitting syntactic descriptions in order to encode separately each alternative realization of an argument may be regarded as an advantageous and easy solution for maintaining the syntactic frames as simple as possible. However this would on the one hand increase dramatically the lexicon size and, on the other hand, prevent from keeping trace of linguistically-relevant distributional equivalences occurring in real language use. In the Italian lexicon, the clustering of different realizations of each position in a single description (fig. 8), insofar as all their combinations produce grammatical sentences, as in the example in figure 9, was therefore adopted as a linguistically sounder solution.

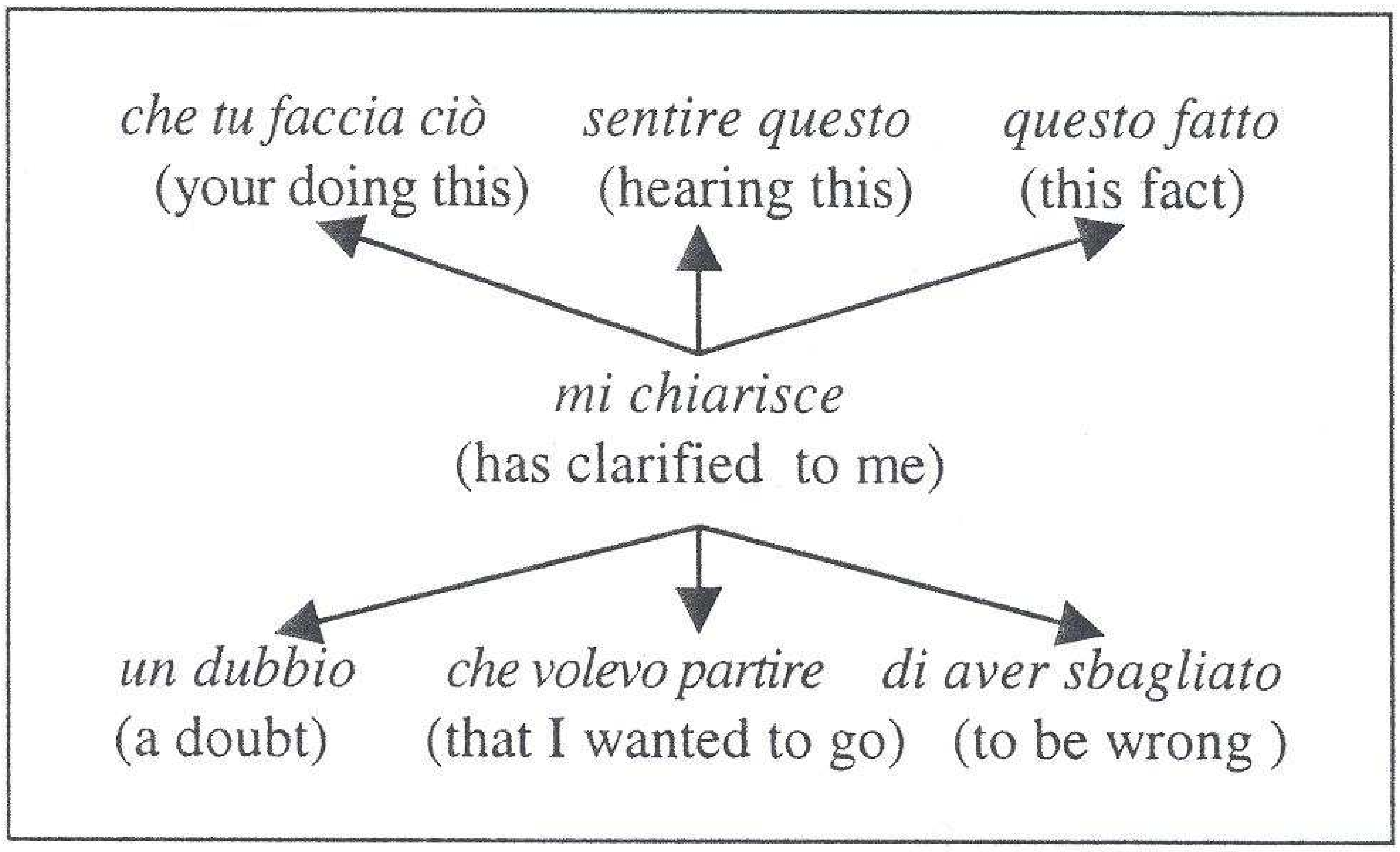

Figure 8. Multiple realizations of complements

```
[Construction:
P0                :[function:subject]
                   [cat:np]
                   [cat:cl] [synsubcat:infcl] [introd:0]

P1[opt:no]        :[function:object]
                   [cat:np]
                   [cat:cl] [syn_sbcat:thatcl] [mood:ind]
                   [cat:cl] [synsubcat:infcl][introd:di]
                                      [coref:COI]
P2[opt:yes]       :[function:indirectobject]
                   [cat:pp] [introd:a] [coref:COI]]
```

Figure 9. Representation of multiple realizations of positions

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, 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 is in this case a mere exemplification of all possible combinations.

5. ENCODED LINGUISTIC STRUCTURES

The number of lexical units handled during the lexicon building process being rather large, we daresay that most of the syntactic structures relevant in modern Italian have been identified and that a lexicon that is fairly representative of the grammatical behaviour of standard Italian has thus been built up.

An overview of the encoded patterns allows the lexicon coverage to be estimated. For the encoding of the three main categories, i.e. verbs, nouns and adjectives, a global number of 1070 syntactic structures were created. For 3,120 verbs, some 776 different descriptions (complementation pattern + lexical unit properties) were identified. The 13,212 nouns required 206 different descriptions for deverbal and simple nouns. As to the encoding of 2,997 adjectives, 88 different structures were detected.

Some observations on the linguistic data, in particular on verb and adjective behaviour, were drawn from the encoding phase. They are illustrated in the following sections.

5.1. *Verb patterns*

The core of verb syntactic structures encoded in the PAROLE lexicon consists of the set of standard structures studied in the framework of the European MLAP project 'COnstraint-based Linguistic Specifications for ITalian' (COLSIT) (Allegranza *et al.*, 1995). This core set has then been gradually enlarged by extracting from the IRC the contexts of occurrence of the most frequent verbs. In the PAROLE Italian lexicon, where the subject is considered as an argument, zero to tetravalent structures of 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.

From the lexical data encoded, we derived the following figures regarding Italian syntactic structures of verbs. Figs. 10 and 11 allow making an inference on both verb type and arity: they clearly indicate the large prevalence of bivalent transitive constructions. On the other hand, intransitive verbs occur much more frequently in complex rather than in basic, monovalent structures.

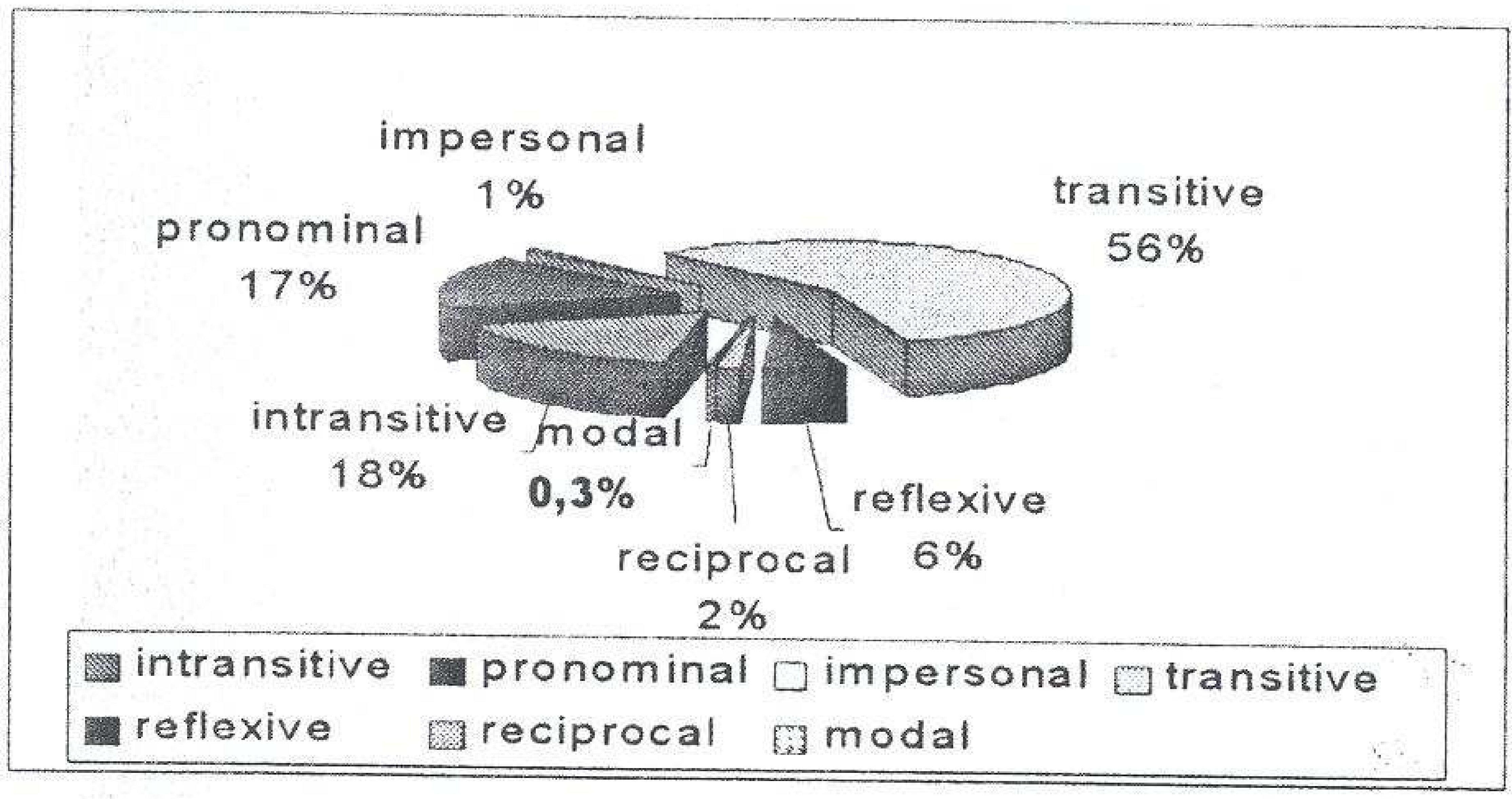

Figure 10. Verb type partition for 7364 verb readings

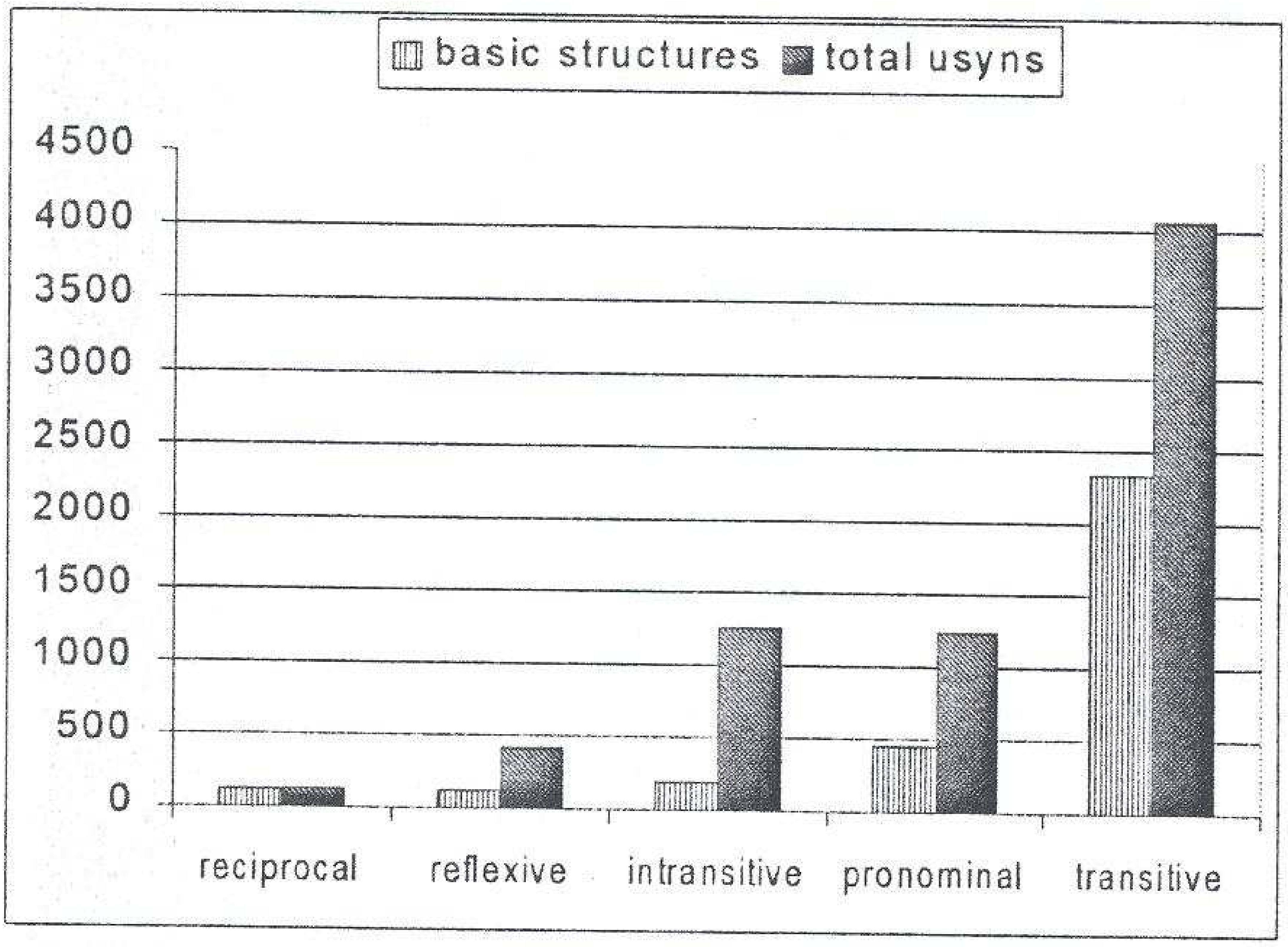

Figure 11. Incidence of basic structures readings

This is confirmed by fig. 12 illustrating the relationship between different verb classes and syntactic structures: it reveals that transitive and intransitive verbs display a similar number of different patterns but that the intransitive readings encoded by means of these patterns are about one third of the transitive ones.

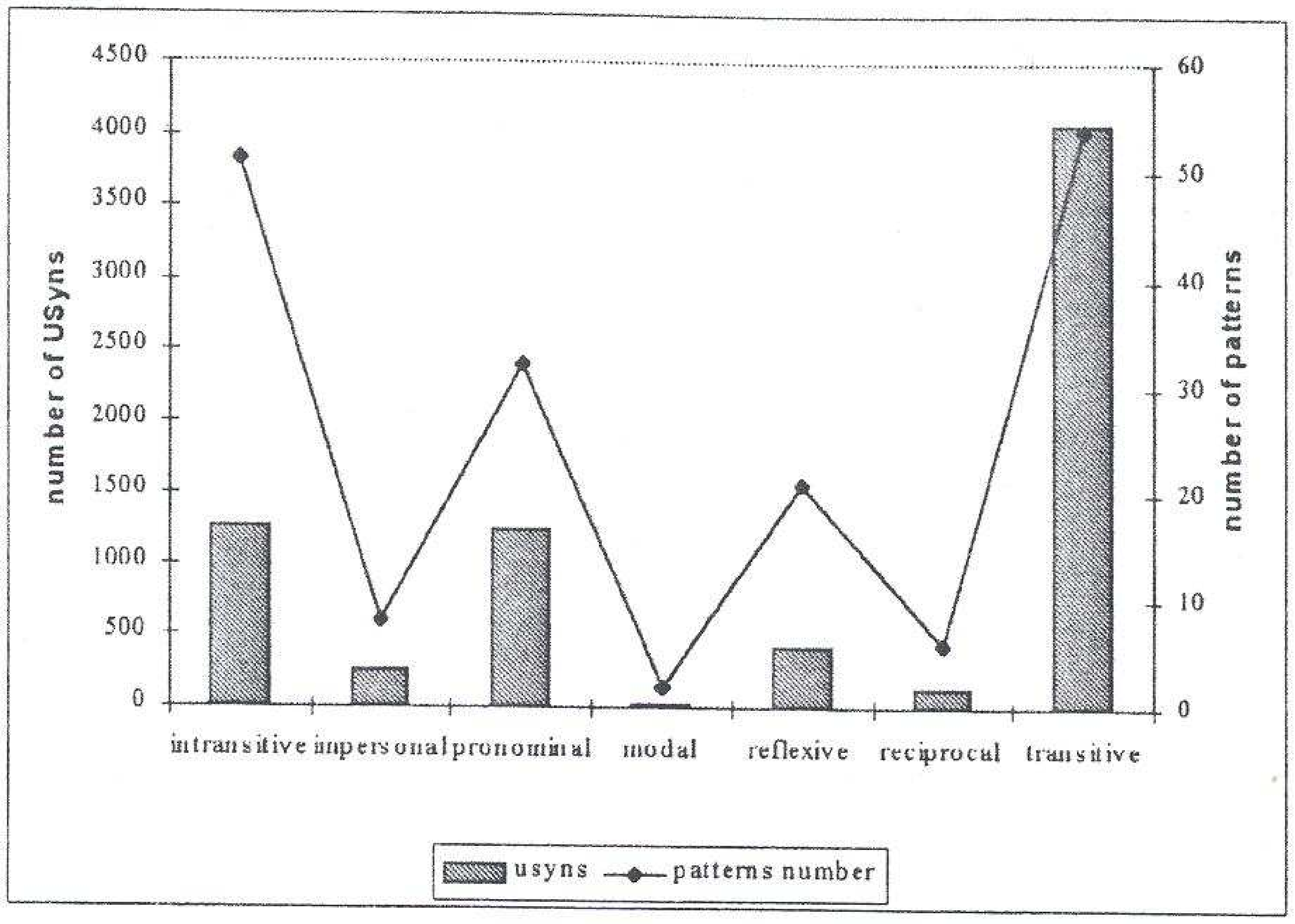

Figure 12. The different patterns for each verb class

5.2. *Noun patterns*

A total number of 13,212 concrete and abstract simple nouns as well as deverbal nouns with up to 4 clausal or phrasal arguments were encoded. From their encoding what emerges is that their syntactic frames present a different level of complexity according to the semantic classes they belong to: the more concrete the meaning, the simpler the syntactic structure. In fact, concrete nouns which were encoded as countable (objects, animals, people, etc.) are generally non frame-bearing. Abstract nouns, on the other hand, may take a complement when the lexical unit denotes one of the properties listed in table 3. Other simple nominals, both abstract and concrete, require a complement specifying them (table 4).

reading	Examples	
inherent/abstract	grandezza/bellezza di	(largeness/beauty of)
relation	amico/zio/capo di	(Luca's friend/uncle/boss)
dimension	distanza/lunghezza di	(a 3-metre distance/length)
interval	pausa/intervallo di	(a 20-minute pause/interval)
group	combriccola/gruppo di	(a gang/group of)
collection	campionario/carnet di	(a sample/booklet of)

Table 3. Simple nouns requiring a phrasal argument

Complement	Examples	
Lemma specification	il centenario di	(one hundred years of)
Content of	un sacco di farina	(a bag of flour)
Topic	un libro di geografia	(a geography book)
Apposition	il fiume Po'/la città di	(the river Po'/city)

Table 4. Simple noun complements

Prepositional phrases denoting possession, e.g.: *la casa di Maria* (Mary's house), free relation *il libro di Luca* (Luca's book), kind of constituency *tubo di acciaio* (steel tube), part of *la gamba del tavolo* (the table leg) were not considered as subcategorized for by the lexical entry. On the other hand, corpus evidence suggested that some nouns take an obligatory complement when used in a metaphorical sense. For example, the lemma *chiave* (key) always occurs with a PP complement in readings such as *la chiave del problema* (the key to the problem).

As to deverbal nouns, tabs. 5 and 6 account for all patterns of deverbal nouns handled in the Italian lexicon.

	Predicate nominalization		***Argument nominalization***	
Verb arity	***Noun arity***	***Examples***	***Noun arity***	***Examples***
1	*1*	• *l'arrivo di Luca* (Luca's arrival) • *il pentimento dello assassino* (the murderer's repentance)	*0*	◊ subject nominalization: nomina agentis *un viaggiatore* (a traveller)
2	2	• *la partecipazione di Luca al progetto* (Luca's participation in the project) • *il desiderio di Luca di viaggiare* (Luca's desire to travel)	*1*	◊ subject nominalization: nomina agentis *un intenditore di vini* (a connoisseur of wines)

Table 5. Deverbal nouns derived from intransitive and pronominal verbs

	Predicate nominalization		*Argument nominalization*	
Verb arity	*Noun arity*	*Examples*	*Noun arity*	*Examples*
2	2	• *l'affermazione da parte di Luca della propria innocenza / di essere innocente / che Maria era innocente)* (Luca's statement of his own innocence / to be innocent / that Mary was innocent)	*0*	◊ subject nominalization: instrument *un frullatore* (a blender) ◊ object nominalization *l'invitato* (the guest) ◊ object nominalization: result *un acquisto* (a purchase)
2	2	• *l'ammirazione di Luca per Maria* (Luca's admiration for Mary)	*1*	◊ subject nominalization: nomina agentis *uno scrittore di romanzi* (a novel writer)
3	*3*	• *l'educazione dei bambini alla tolleranza da parte dei genitori* (parents' education of children about tolerance) ◊ obj. predicate structures *la designazione di Luca a presidente da parte dei soci* (Luca's nomination to president by the members)	*1*	◊ indirect object nominaliz. *il destinatario di un pacco* (the addressee of a parcel)
4	*4*	• *il trasferimento di lire da Pisa a N.Y. da parte della banca* (the transfer of lira from Pisa to N.Y. by the bank)		

Table 6. Deverbal nouns derived from transitive verbs

De-adjectival and non-deverbal predicative nouns were also assigned an argument structure similar to the one ascribed to deverbals, e.g.: *il diritto di Luca di votare (*Luca's right to vote)*; la paura di Luca del buio* (Luca's fear of the dark).

5.3. *Adjective patterns*

A peculiarity of adjectives is the relevance of their distributional properties to their syntactic structure. Adjectives may in fact be used both predicatively and attributively depending on their position with respect to the nominal phrase.

As shown in figure 13, this is a feature shared by most of the Italian adjectives and therefore, we adopted the solution of describing such lexical items in a unique, frameless Syntactic Unit, with the specification of their double function. Besides the function, information about pre- or post-nominal position in attributive uses[11] and (non)gradability are stipulated in adjective lexical entries. As for their position, Italian adjectives are used either in postnominal, e.g.: *uomo ammalato* (ill man) (**ammalato uomo*), in prenominal position *altre cose* (other things) (**cose altre*) or in free position, e.g.: *crescente interesse / interesse crescente* (growing interest) (see figs. 13 and 14).

Besides adjectives occurring indifferently pre- or postnominally, a group of adjectives whose position confers a different meaning to the head noun, e.g.: *alto ufficiale / ufficiale alto* (high-ranking officer / tall officer) have been encoded in two different readings.

[11] It is now a widely acknowledged fact that some linguistic phenomena are hard to describe by means of clear-cut statements since language is quite a flexible entity. In this context, we need hardly say that the information about the adjectival position is to be understood, in most cases, as an indication of preferential behaviour rather than as an absolute constraint.

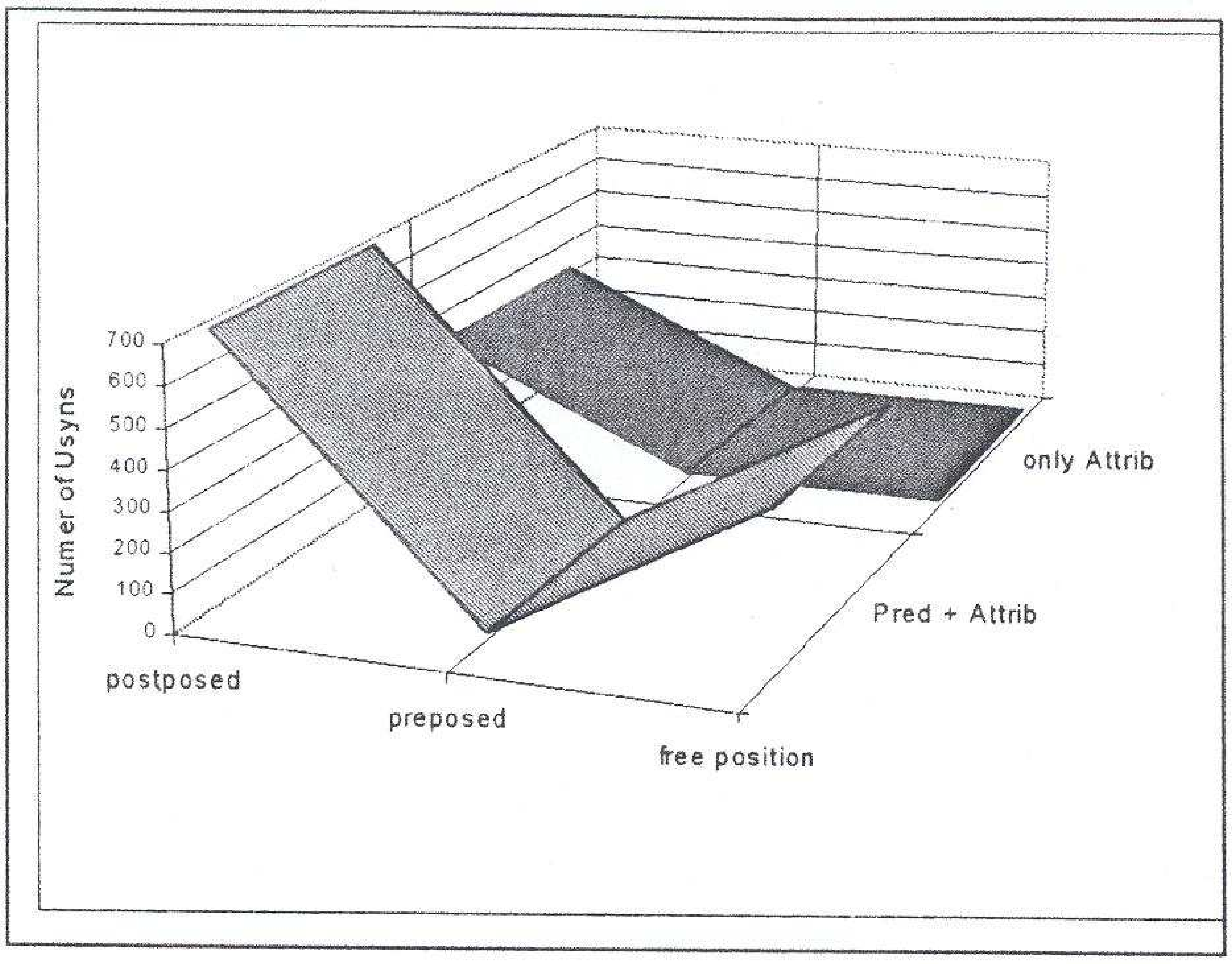

Figure 13. Position and function of adjectives

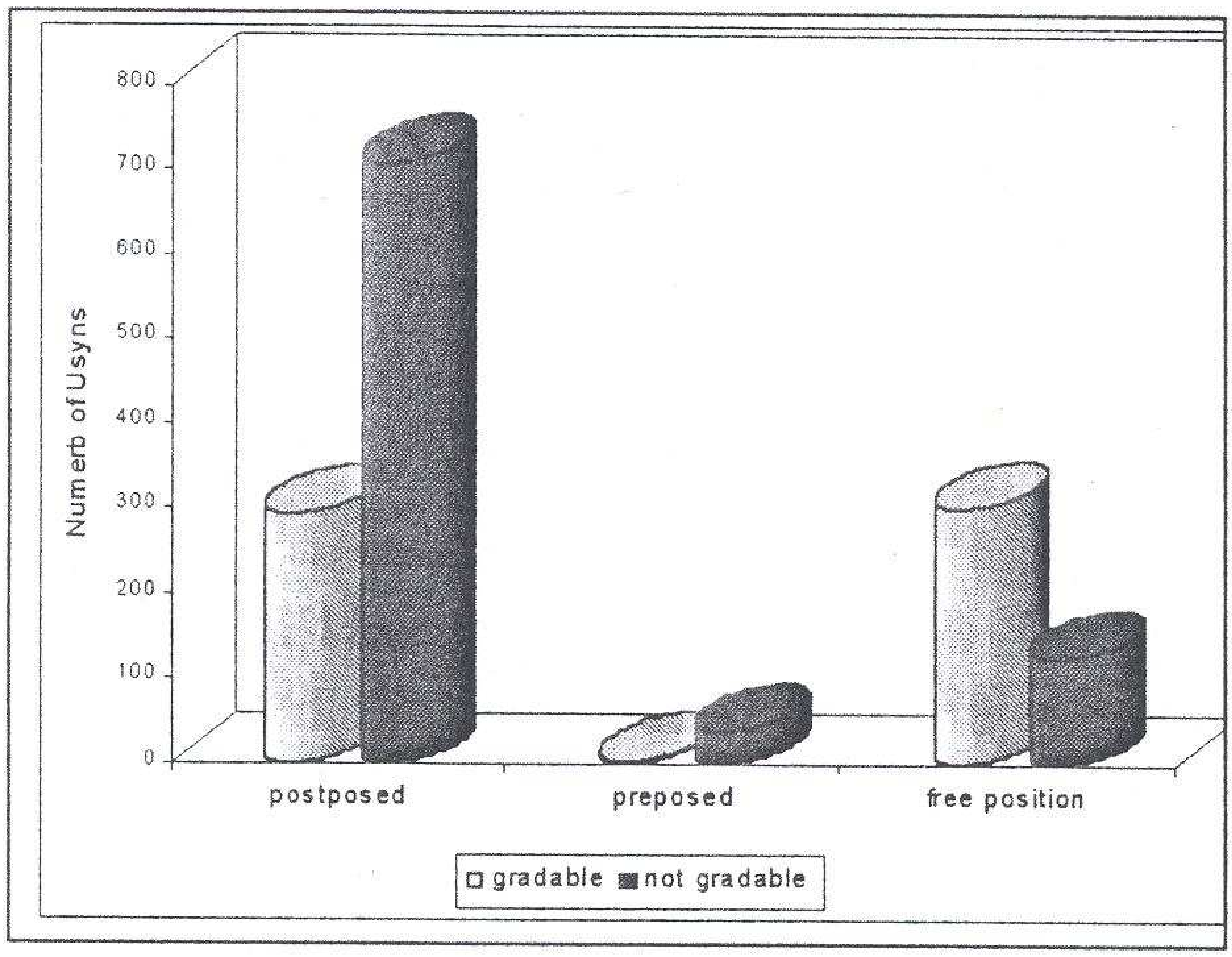

Figure 14. Position and gradability of adjectives

Non-predicative adjectives, most of which belong to the relational class, were also assigned non-valent structures. Valent adjective complements were encoded as optional or obligatory depending on whether their presence or absence affected or not the adjective's meaning. In the case of *un lettore abbonato (ad una rivista)* (lit., a reader subscribed to a magazine), the complement was marked as optional, while two entries were created for the adjective *abile* used in contexts such as *una persona abile al lavoro* (a person able to work) and *un abile politico* (*PP complement) (a clever politician). Adjectives subcategorizing for an infinitive clause required the description of a wider context in the lexical entry in order to mark coreference in the infinitive clause, as e.g. *questo lavoro è difficile da fare* (this work is difficult to do), where the complement clause object is coreferent with the main clause subject. Adjectives entering in impersonal constructions such as: *è opportuno partire* (leaving is necessary) or *è opportuno che tu parta* (it is necessary that you leave) were described by means of a one-position frame. The type of construction at hand, which implies morphological constraints on the copula, is specified among the head properties.

6. FINAL REMARKS

The complexity and elevated cost of creation of language resources has induced the scientific community to pay more attention to the issue of reusability of existing data. Unfortunately, language resources are often created from too 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 granular 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. These lexica are in fact all based on the modular and flexible PAROLE model, they share the same theory and application-independent linguistic specifications, a global architecture, a core set of information, a descriptive language, a management tool and SGML exchange format. Moreover, 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 requisites 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, a straightforward enlargement or refinement of the lexical information with no need for overall restructuring and large scale reusability in different theoretical and application frameworks, among which NLP systems development, information retrieval, language learning and MT applications. Partial knowledge, relevant for specific NLP application-dependent models and applicative dictionaries can be derived from this repository of information, mapping the application model from the generic one. Owing to their uniformity, these resources also lend themselves to multilingual applications.

These standardized lexical resources offer a significant contribution to the development of the Language Engineering Industry. Their creation is particularly critical for Europe which needs to lower its high communication costs that are due to the large number of languages used.

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

The Italian instantiation of the PAROLE syntactic lexicon presents many interesting aspects. First of all, it has been based on corpus data, as regards both the procedure of lemma acquisition and the identification or check of attested syntactic structures. It encodes therefore a broad-coverage, general and modern language. Secondly, its computational nature, which enables the handling of a very large amount of entries has permitted a coherent and standardized structuring of information, which paper dictionaries usually lack of. Thirdly, the level of quality of its data has been internally validated through an integrity checking procedure (Battista, 1998) which controlled both the completeness and consistency of the information encoded[12]. Lastly, whilst preserving its own specificity through the choice of descriptive granularity, coding strategy and a corpus evidence guided treatment of a large

[12] The data have then undergone an external validation performed, through ELRA, by external industrial users.

number of language-specific phenomena, it presents the advantage of being part of a network of European lexica sharing an approach to a conceptual and representational model, a core set of information encoded and a representation type. This membership in a network of European monolingual lexica, which thus implies the possibility of comparison, creation of multilingual links, and use in multilingual NLP applications contributes undoubtedly to increase its value.

The PAROLE Italian lexicon has been already used as a gold standard for the evaluation of Italian data in other EU projects, such as shallow parsing and knowledge extraction for the language engineering project (LE-SPARKLE). It also constitutes the initial nucleus of a larger lexicon, based on PAROLE and SIMPLE specifications, which is being developed in the framework of the Italian national project 'Corpora e Lessici di Italiano Parlato e Scritto' (CLIPS).

REFERENCES

AA. VV., *The EUROTRA Reference Manual*, 7.0., Commission of the EC, Luxembourg, 1990.

ALLEGRANZA V., MAZZINI G., RUIMY N., MLAP93-08B Project: *COnstraint-based Linguistic Specifications for ITalian (COLSIT),* Final report, December 1995.

BATTISTA M., *The PAROLE Pisa lexicon integrity checker,* ILC-CNR, Internal Report, Pisa, 1998.

BINDI R., MONACHINI M., ORSOLINI P., *Italian Reference Corpus. General Information and Key for Consultation*, ILC-TLN-1991-1, ILC-CNR, Pisa, 1991.

CALZOLARI N., BAKER M., KRUYT T. (eds.), *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study*, "Linguistica Computazionale", Pisa, 1995.

CALZOLARI N., MONTEMAGNI S., PIRRELLI V., *Verb Subcategorization, Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa, 1996.

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

GENELEX CONSORTIUM, *EUREKA Project GENELEX - Report of Syntactic Layer,* 4.0., 1993.

GRIMSHAW J., *Argument Structure*, The Mit Press, Cambridge, MA, 1990.

LE-PAROLE, *Technical and Financial Annex,* LE-4017, 1995.

MONTEMAGNI S., PIRRELLI V., *Verb Subcategorization in Italian, Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa, 1996a.

MONTEMAGNI S., PIRRELLI V., *Noun, Adjective, Adverb and Preposition Subcategorization in Italian, Blueprint of PAROLE Guidelines to the Encoding of Syntactic Information in the Lexicon*, Pisa, 1996b.

RENZI L., SALVI G. (eds.), *Grande Grammatica italiana di consultazione*, vol. I-III, 1988/91/95, Il Mulino, Bologna, I, 1988.

RUIMY N., *The Lexicon Structure,* in *Constraint-based Linguistic Specifications for Italian (COLSIT),* Final report, MLAP93-08B Project, December 1995.

RUIMY N., BATTISTA M., CORAZZARI O., GOLA E., SRANU A., *Italian Lexicon Documentation*, P-WP3.11-WP-PSA-1, Midterm Review Documentation, LE2-4017, PAROLE EC project, March 1997.

RUIMY N., BATTISTA M., CORAZZARI O., GOLA E., SPANU A., *Italian Lexicon Documentation*, P-WP3.11-WP-PSA-2, Final Review Documentation, LE2-4017 PAROLE EC project, 1998a.

RUIMY N., CORAZZARI O., GOLA E., SPANU A., CALZOLARI N., ZAMPOLLI A., *The European LE-PAROLE project: The Italian Syntactic* LEXICON - First International Conference on Language Resources and Evaluation - ELRA Proceedings, Granada, 1998b, vol. 1, 241-248.

RUIMY N., CORAZZARI O., GOLA E., SPANU A., CALZOLARI N., ZAMPOLLI A., *The European LE-PAROLE Project and the Italian Lexical Instantiation* - ALLC/ACH Proceedings, Debrecen, 1998c, 149-153.

RUIMY N., CORAZZARI O., GOLA E., SPANU A., CALZOLARI N., ZAMPOLLI A., *LE-PAROLE project: The Italian Syntactic Lexicon,* EURALEX '98, Proceedings, Université de Liège, vol. I, 259, 1998d.

SANFILIPPO A. *et al., Subcategorization Standards,* Report of the Eagles/Lexicon/Syntax Group, 1996.

SCHWARZE C., *Grammatik der italienischen Sprache*, Verbesserte Auflage, Niemeyer Verlag, Tuebingen, vol. 2., 1995.

WALKER D., ZAMPOLLI A, CALZOLARI N. (eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment*, in *Proceedings of the Grosseto Workshop*, Oxford University Press, Oxford, 1995.

ZAMPOLLI A., *MLAP PAROLE Technical Annex,* Pisa, 1994.

ZAMPOLLI A., *Introduction*, in N. CALZOLARI, M. BAKER, T. KRUYT (eds.), "Linguistica Computazionale", (1995), Giardini Editori e Stampatori, Pisa, 1995, xi-xxxix.