The European LE-PAROLE Project: The Italian Syntactic Lexicon

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
v. della Faggiola, 32 - 56126 - Pisa - Italy
[2] Consorzio Pisa Ricerche
p. A. D'Ancona , 1 - 56126 -Pisa - Italy

Abstract

LE-PAROLE is the first EC funded project producing large, generic and re-usable written language resources in so many languages of the European Union and built according to the same design principles, same linguistic specifications and representation format.
This paper presents an overview of the PAROLE Italian Syntactic Lexicon as an instantiation of a lexicon built according to the PAROLE model. Some language-specific linguistic and lexicographic options, complying with the conceptual and representational model and according to which the syntactic encoding was performed will be illustrated. An overview of the syntactic patterns encoded for verbs, nouns and adjectives will allow the lexicon syntactic coverage to be estimated.

1. Introduction

The 1986 Grosseto workshop «On automating the lexicon» (Walker *et al.*, 1995) is usually recognised as the starting point of the increasing interest the NLP community took in re-usable language resources. The LE-PAROLE project is the follow-up of initiatives promoted on this occasion. The Council of Europe, co-sponsor of the workshop, formed a group of experts representing European institutes to explore the feasibility of harmonising their activities, in order to establish a Network of European Reference Corpora (NERC) (Calzolari *et al.*, 1996; Zampolli, 1996). This group, gradually enlarged to include members of all the Union languages, constituted the PAROLE Consortium[1] which carried out the MLAP PP-PAROLE project for the elaboration of the linguistic specifications to be followed in the LE-PAROLE project. This two-year project, which has just concluded, aimed at developing and making available large, generic and re-usable harmonised written language resources, i.e. corpora for 14 languages and electronic lexica for 12 languages of the European Union. 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.
Each of the 12 monolingual lexica consists of 20,000 entries providing morphological, syntactic and, in a few cases, semantic information. They all follow the PAROLE project linguistic specifications (Calzolari, Montemagni & Pirrelli. 1996; Flores 1996) which are based on EAGLES recommendations for morphosyntactic information and verb syntax (Sanfilippo *et al.* 1996) and on the extended GENELEX (GENEric LEXicon) model for morphology and for the handling of non-verb categories. These linguistic guidelines are implemented in the LE-PAROLE model which provides the overall lexicon architecture and the coding formalism. 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 inter-consistency.
This paper presents an overview of the PAROLE Italian Syntactic Lexicon as an instantiation of a lexicon built according to the PAROLE model. Some language-specific linguistic and lexicographic options which were taken up for Italian and according to which the syntactic encoding was performed will be illustrated. An overview of the syntactic patterns encoded for verbs, nouns and adjectives will enable the lexicon syntactic coverage to be estimated.

2. The PAROLE 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) which 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 definiteky. The object identifier is then simply assigned to all relevant lexical items sharing that pattern without need to stipulate once more in the lexical entries the pattern properties.
The modularity of the PAROLE lexical model is such that the information encoded in the morphological, syntactic and semantic descriptive levels is independent from each other although the three levels are connected. A complete entry is a progression through the levels of information encoded. A morphological unit 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

---

[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).

morphological unit it is associated with. A syntactic unit, on the other hand, is associated to 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.

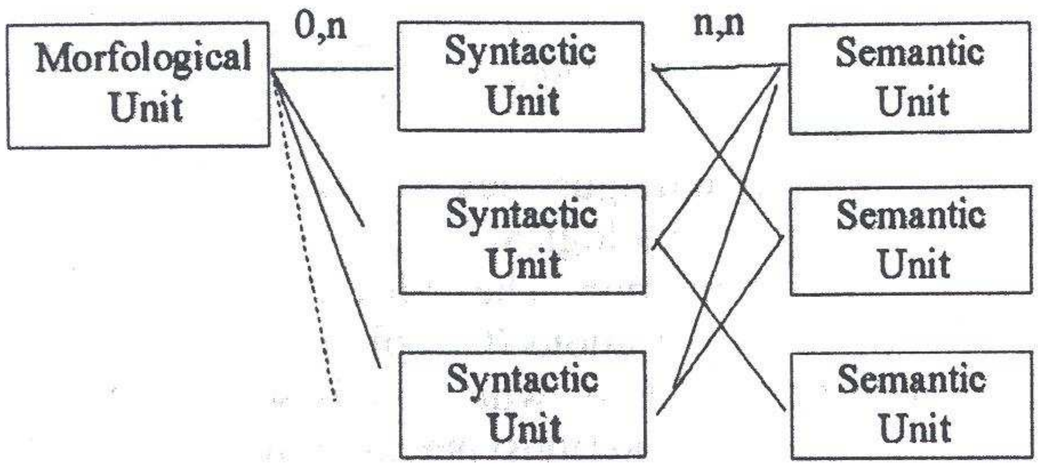


Figure 1: General architecture of the lexicon

For each linguistic level, the descriptive structure consists of an interaction of basic and complex descriptive elements, the complex elements being described by more basic ones. Most of the descriptive elements are shared by other elements of higher level. At syntactic level, a lexical entry is encoded as a (set of) Syntactics Unit(s), an entity which describes one of the syntactic behaviours of a morphological unit. A Syntactic Unit is characterized by one Base Description and, optionally, some Related Descriptions encoding closely related surface alternations of the base one. A Description, or frame, consists of both a Construction providing information about the syntactic context of the lexical entry — i.e. a list of canonically ordered[2] Positions, or frame slots, which describe the syntactic constituents and their restrictions — and a Self encoding the properties/restrictions of the entry at hand (e.g.: PoS and subcategory) in the specific subcategorization frame being dealt with (e.g.: selected auxiliary for a verb reading).

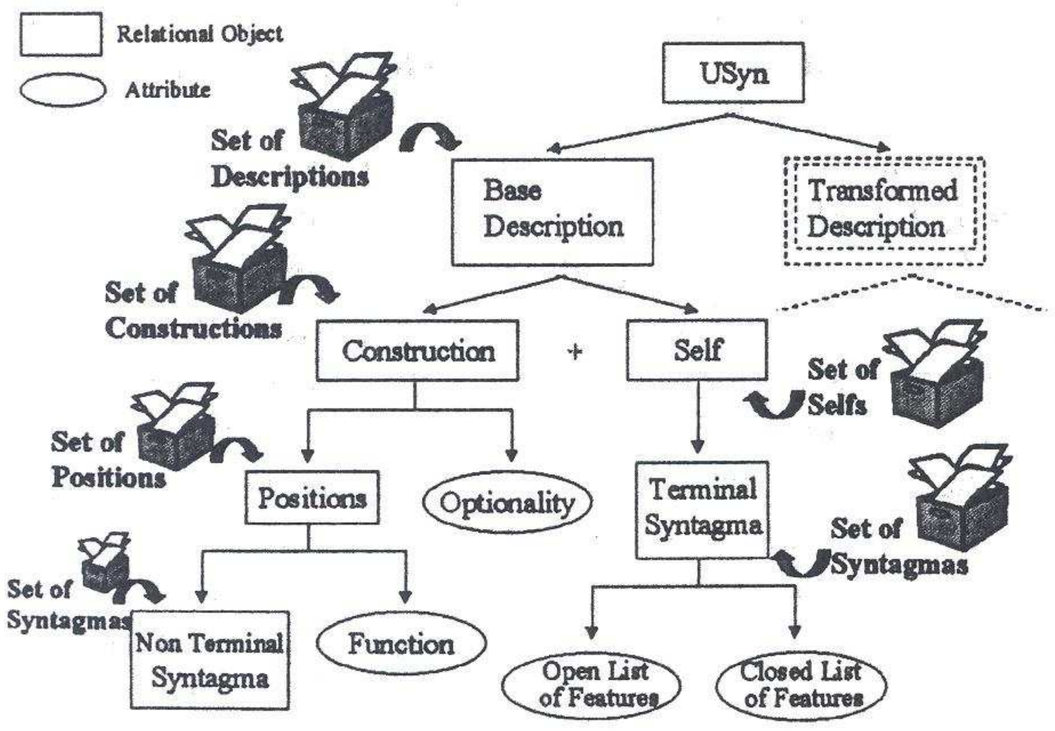


Figure 2: Main objects and attributes at syntactic level.

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

Each position filler is a constituent strongly-bound to the lexical unit and is modelled as a bundle of linguistic information ranging from syntactic function and 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. Within a Syntactic Unit, the base and related descriptions (e.g.: causative and inchoative readings of a verb) 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. The relationship between two Syntactic units encoding different parts of speech (e.g.: a verb and its nominalization) may be realized on the other hand 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 structure being 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

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

At syntactic level, on the other hand, it is composed of 20,000 one-word entries selected on frequency criteria from the ILC Italian Reference Corpus (IRC) (Bindi *et al.*, 1991) and therefore belonging to general modern language. The selected lemmas pertain to the following parts of speech: verbs (3,000), nouns (13,000), adjectives (3,000), adverbs (500), and empty words (500).

3.1 Splitting Criteria

The splitting of entries, at syntactic level, was performed avoiding both redundancy and over-powerful gatherings as a general rule. Syntactic-driven criteria were followed and semantic considerations were accounted for only in so far as they had consequences at syntactic level. Differences in arity and function assignments were therefore clearly criterial for splitting and among the many other fuzzier cases which presented less striking

dissimilarities between syntactic structures, the following ones gave rise to a split:

- optionality of a complement in one reading only[3]:

forare (una gomma) / **forare (un biglietto)*
'to burst a tyre' 'to punch a ticket'

un soldato prigioniero (dei tedeschi) / **un uomo prigioniero (delle proprie idee)*
'a soldier prisoner of the Germans' 'a man prisoner of his own ideas'

- relationship of only one reading with another Syntactic Unit, e.g. link between transitive and reciprocal verbs:

Luca affronta il pericolo ↔ ** Luca e il pericolo si affrontano*
'Luca faces danger' 'Luca and the danger face each other'

Luca affronta il nemico ↔ *Luca e il nemico si affrontano*
'Luca confronts the enemy' 'Luca and the enemy face each other'

- alternative realizations of a complement in one reading only:

1a. *il conto comprende il servizio*
'the bill includes service charge'

1b. *i genitori non comprendono*
'parents do not understand'

che i figli vogliono essere liberi 'that children want to be free'
di dover lasciar i figli liberi 'that they should let their children be free'
i figli 'their children'

- different behaviour w.r.t. nominalization for homographic verbs:

doppiare un film / *il doppiaggio di un film*
'to dub a film' 'the dubbing of a film'

vs

doppiare il Capo Horn / ** il doppiaggio del Capo Horn*
'to round the C.H.'

3.2 Lexical Entry Information

The lexicon describes the lexically-governed syntactic context, i.e. the peculiar properties of a lexical unit and of the syntactic constituents it specifically subcategorizes for. By contrast, general properties shared by a whole word class and derivable by virtue of the membership of a lemma to a class, are assumed to be dealt with at grammatical level. Only the idiosyncratic behaviours w.r.t. to grammatical rule's application are therefore stipulated in the lexicon. Wider context is not specified unless crucially required. 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.

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

3.3 Complements and their Optionality

In the PAROLE guidelines, the definition of frame is somewhat liberal. 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, Montemagni & Pirrelli 1996). The determination of which constituents are lexically-selected and which are not is therefore a crucial — and sometimes hard — 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 patterns against corpus evidence.

Complement optionality was assessed for verbs by considering nuclear, unmarked contexts since marked ones allow even the omission of complements usually considered as obligatory. Optionality of noun complements, which is a more controversial issue, was on the other hand less easy to determine. Noun complements are in fact often all assumed to be optional. In reality, a distinction must be made between simple and deverbal nouns. Simple nouns complements were considered as optional. As for deverbal nouns, by-phrases (realized either with *da parte di* or *di* in Italian) were encoded as optional and object-like complements as obligatory in two-argument bearing complex event nominals. Complements of both deverbal and simple nouns were encoded as obligatory in non literal meanings e.g.: *una fonte di problemi* 'a source of problems'; *la fioritura delle arti* ' the flourishing of arts'.

3.4 Syntactic Functions

The assignment of syntactic functions was not always straightforward. For example, the choice 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[4] for their assignment were therefore established in order to ensure coding consistency[5]. 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). PPs introduced by strongly-bound prepositions, as in *rinunciare a qualcosa* 'to give something up' were also attributed this function. The adverbial function, on the

[4] We thank Maria Gronostaj, from the Swedish Parole team, for her relevant contribution to the statement of these criteria.

[5] 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).

other hand, was assigned to PP complements in alternative distribution with adverbs and whose interrogative form was built with an interrogative adverb (table 2). It was assigned as well to PPs or adverbial phrases which, together with the verb, confer an idiomatic meaning to the phrase, i.e.: *accendersi d'ira* 'to explode with rage'. These complements correspond to the thematic roles (not encoded in the lexicon) reported in tables 1 and 2.

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.thing' (about what?)

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

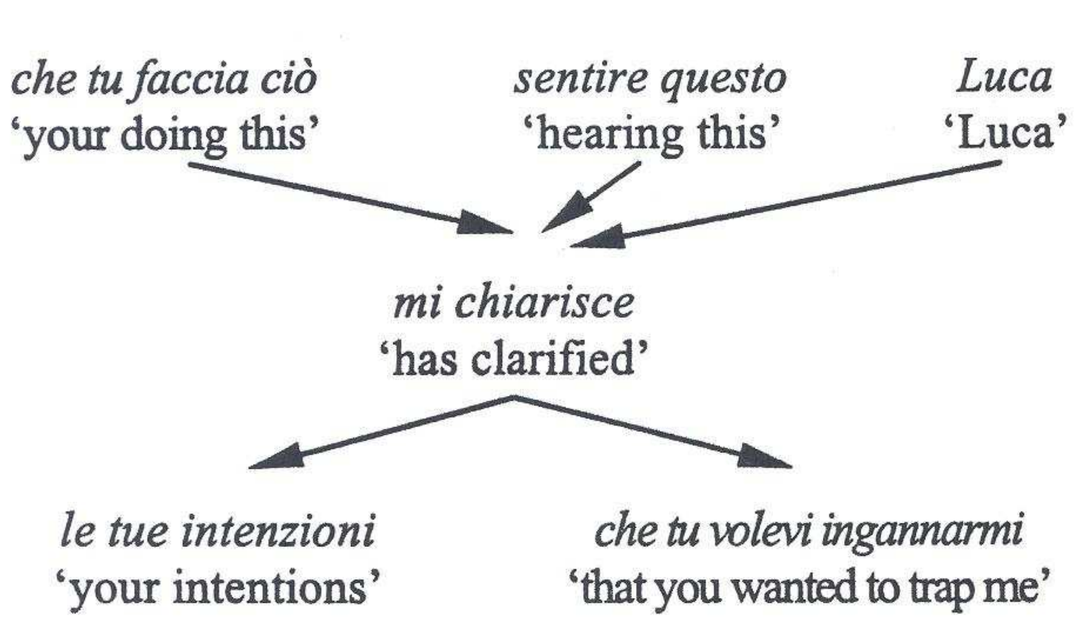
θ-role	adverbial
manner	*circolare a piedi* (come?) 'to go on foot' (how?)
measure	*allungare di un metro* (di quanto?) 'to lengthen by one metre' (how much?)
time	*iniziare presto/alle otto* (quando?) 'to start early/at 8 o'clock' (when?)
locative/ directional	*vivere a Parigi* (dove?) 'to live in Paris'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

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

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

3.5 Syntactic Realizations

Each slot in a frame is assigned a syntactic function and is instantiated by one or more paradigmatically-related alternating slot fillers. In the Italian lexicon, in order to avoid an indesirable and cumbersome splitting of descriptions, the different realizations of each position were clustered in a single description provided that all their combinations produce grammatical sentences, as illustrated in tab.3 for the example below:




```
[Construction:
P0          :[func:subject]
                [cat:np]
                [cat:clause] [synsubcat:ssinfinitive] [introd:0]
                [cat:clause] [synsubcat:thatcl] [mood:sub]
P2 [opt:no] :[func:object]
                [cat:np]
                [cat:clause] [synsubcat:thatcl] [mood:ind]
P3 [opt:yes] :[func:indirectobject]
                [cat:pp] [introd:a]]
```

Figure 3: Representation of multiple realizations of positions[6].

3.6 Constraints Information

The arguments encoded in the different positions of an entry may be constrained by means of features. Morphosyntactic information of mood, agreement, lack of determination, lexical specification of prepositional phrases and clausal complements introducers as well as control information are encoded by means of features at position filler level. Non-coreference of subjects between completive and matrix clauses, e.g.: *detesto le partenze / - partire / - che tu parta /* - che io parta* lit.: 'I hate departures / - leaving / - the fact that you are leaving / - *that I leave' and *Giovanni è lieto che tu (*egli stesso) parta* 'G. is pleased that you are (*he is) going' is also marked by means of position level features on controller and controllee. For control information in infinitive clauses, a frame level feature indicates moreover the type of construction at hand (subject control, raising, etc). Constraints on the headword, in the particular reading being described, such as e.g.: impersonal constructions for verbs, mass/count distinction for nouns, position of the adjective in NPs, on the other hand, are expressed outside the complements description, in the SELF.

3.7 Relating Lexical Information

In a lexicon, some pieces of information gain significance when related to each other. This is the case of diathesis alternations, where the expression of verb arguments alternate in two different readings which express basically the same situation. In the Italian lexicon, alternations such as causative-inchoative *rompere* 'to break', locative *caricare* 'to load', instrument subject *colpire* 'to hit', simple reciprocal alternation with both transitive *confondere* 'to mistake' and intransitive *parlare* 'to talk' verbs, were related through the Frameset mechanism which allows the slots of two frames of a complex Syntactic Unit to be linked. The verb *affondare* 'to sink' was for example encoded as a complex syntactic unit with two different descriptions and a call to a Frameset named 'causative' which relates causative readings objects to inchoative subjects. For deverbal nouns and their verbal bases, the correspondence between the set of complements they both subcategorize for — hence a relationship holding between two Syntactic Units

6 In this partial representation of an entry, the information is modelled in an internal intermediate format used to encode syntactic structures by means of macros.

encoding different PoSs — was expressed through the TransfUsyn device.

4. Encoded Linguistic Structures

Encoding 20,000 entries enabled us to deal with quite a large number of Italian syntactic structures and to build up a lexicon that is fairly representative of the grammatical behaviour of standard Italian. As a matter of fact, for the encoding of the three main categories, i.e. verbs, nouns and adjectives, a global number of 1053 syntactic structures[7] were created. For 3,000 verbs, some 781 different descriptions (complementation pattern + lexical unit properties) were identified. The 13,000 nouns required 125 different descriptions for deverbal nouns and 56 for simple nouns. As to the encoding of 3,000 adjectives, 91 different structures were detected.

4.1 Verb Patterns

The core of verb syntactic patterns 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). 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[8] 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 information on Italian syntactic structures of verbs. The arity distribution for all verb readings, illustrated in figure 4, highlights that most verbs subcategorize for two arguments.

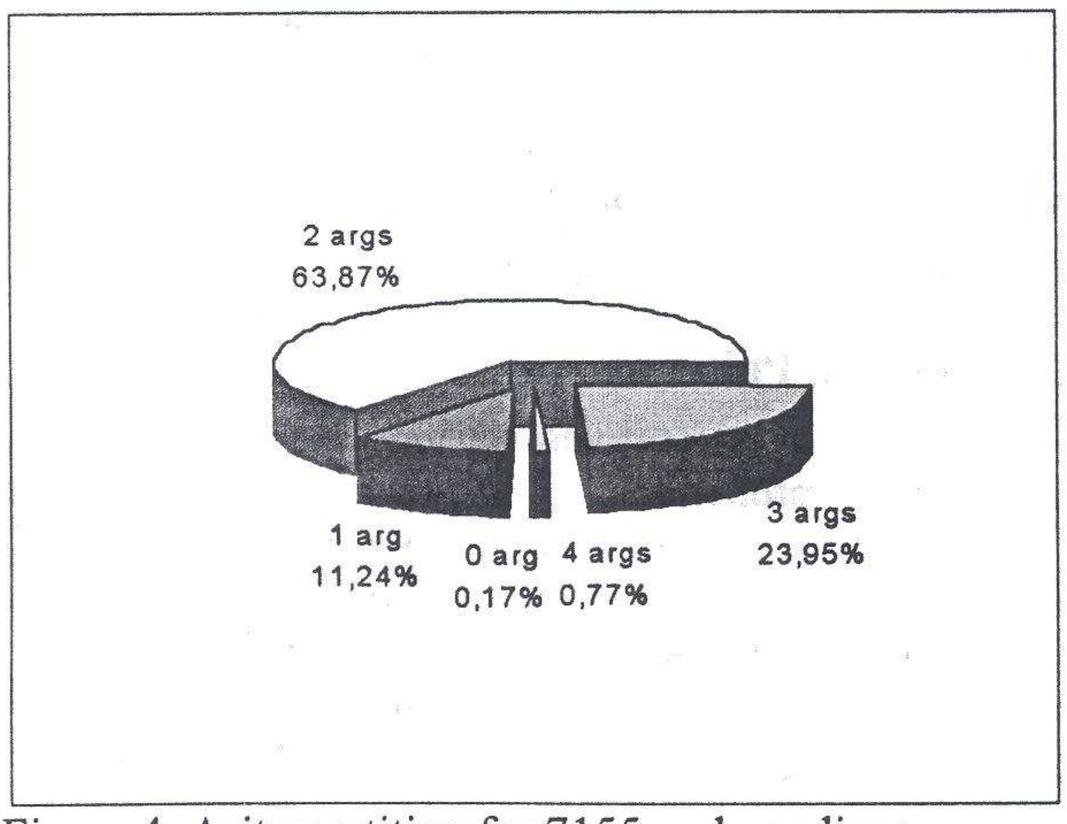


Figure 4: Arity partition for 7155 verb readings

Figure 5 illustrates the correlation between main verb classes and patterns arity. It shows for example that two-slot frames occur mainly with transitives, but are also frequent with intransitives and pronominals.

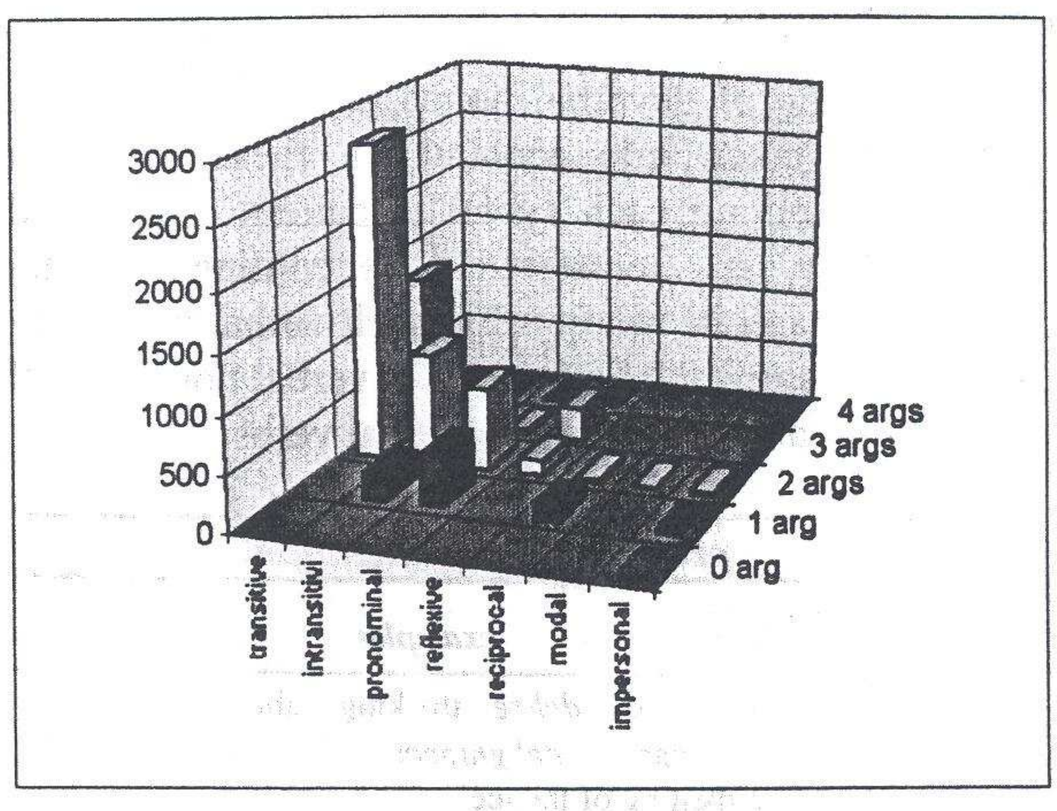


Figure 5: Arity and verb type.

Figure 6 illustrates the relationship between the different verb classes and the syntactic patterns: it shows that transitive and intransitive verbs display a similar number of different patterns but that they use them quite differently. The intransitive readings encoded through these patterns are about one third of the transitive ones.

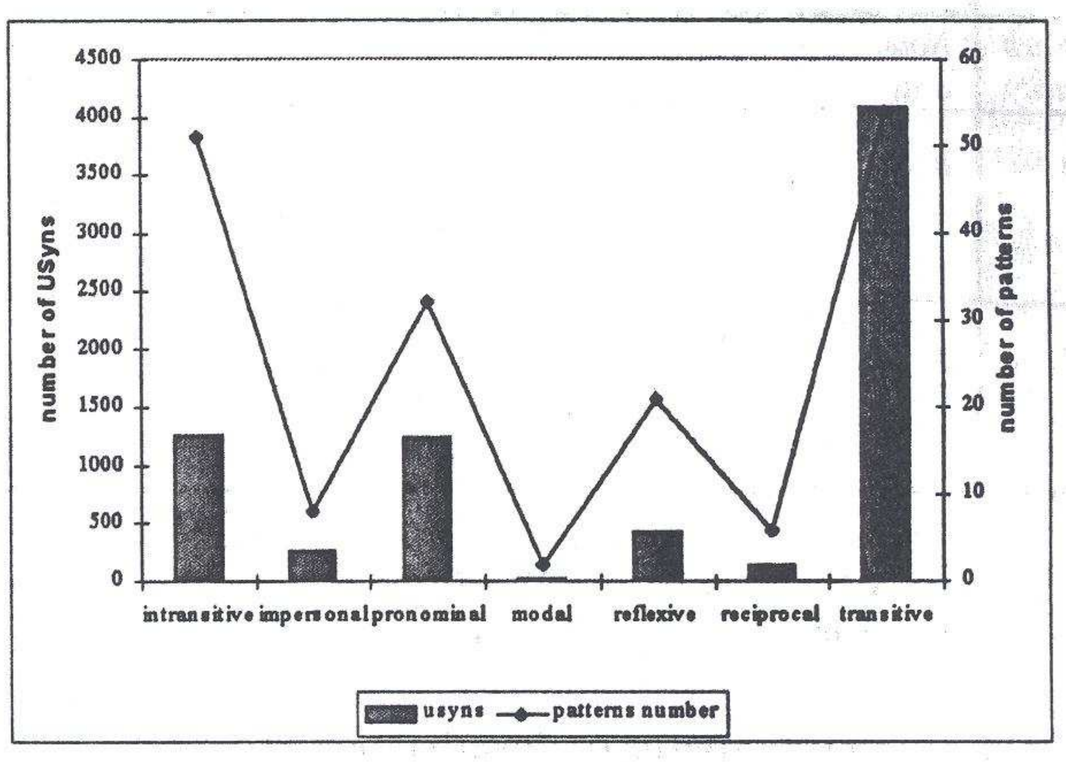


Figure 6: The different patterns for each verb class

Figure 7 indicates that transitive verbs occur mainly in basic (subject/object) constructions while intransitive verbs occur more frequently in complex rather than in basic (subject) structures.

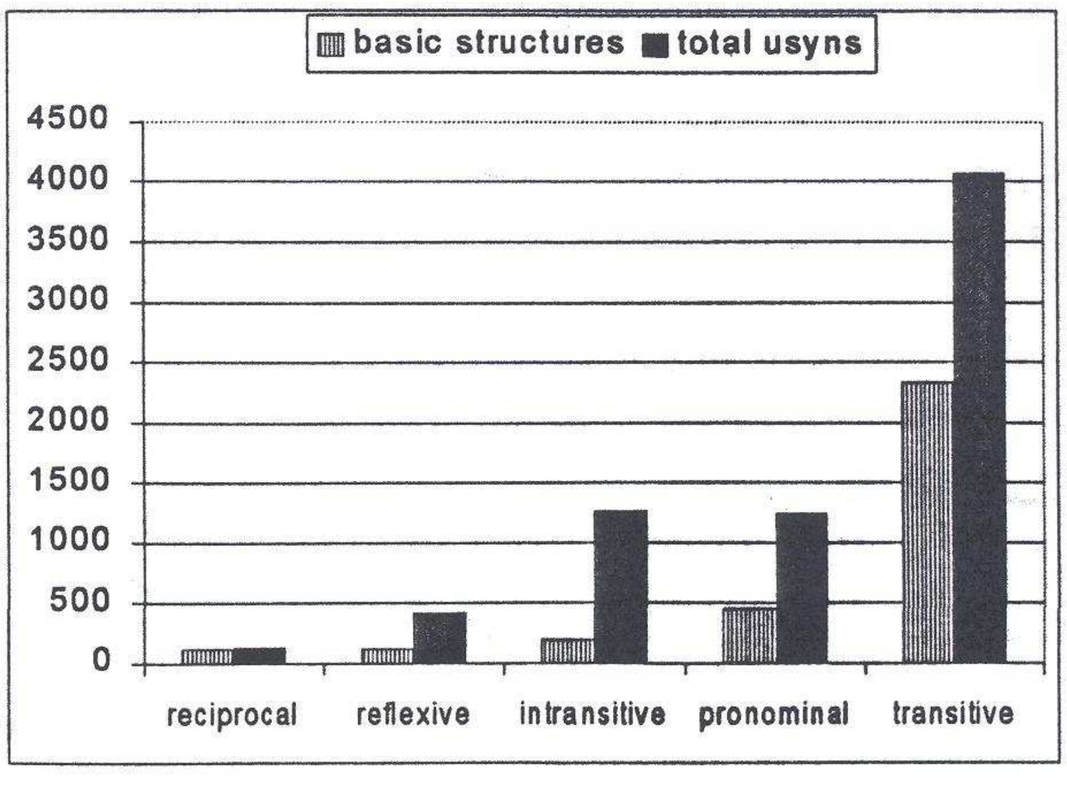


Figure 7: Incidence of basic structures readings.

---

[7] 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.
[8] In the Italian lexicon, the number of complements we encoded has been limited to four.

4.2 Noun Patterns

For the encoding of deverbal nouns, tables 3-6 illustrate the structures taken into account. In these tables an example representing each syntactic pattern is reported. The examples are clustered according to the deverbals'arity and its relationship with verbal base arity. Tables 3 and 4 show the patterns of deverbal nouns which derive from intransitive and pronominal verbs.

	predicate nominalization	
Verb arity	*Noun arity*	*examples*
1	*1*	• *l'abdicazione del re* 'the king's abdication' • *la liquefazione del ghiaccio* 'the melting of the ice'
2	*2*	• *la partecipazione di Luca al progetto* 'Luca's participation in the project' • *l'esitazione di Luca a partire* 'Luca's hesitation in leaving'

Table 3: Predicate nominalization of deverbal nouns derived from intransitive and pronominal verbs.

	argument nominalization	
Verb arity	*Noun arity*	*examples*
1	*0*	◊ subject nominalization: nomina agentis & instruments • *un viaggiatore* 'traveller' • *un frullatore* 'a blender'
2	*1*	◊ subject nominalization: nomina agentis • *un aspirante al successo* 'so. aspiring to being famous' • *un intenditore di vini* 'a connoisseur of wines'

Table 4: Argument nominalization of deverbal nouns derived from intransitive and pronominal verbs.

In tables 5 and 6 patterns of deverbal nouns deriving from intransitive and pronominal verbs are illustrated.

	predicate nominalization	
Verb arity	*Noun arity*	*examples*
2	*2*	*la constatazione da parte di Luca della propria ignoranza / - di aver sbagliato / - che tutto era finito* 'Luca's awareness of his own ignorance / - of being wrong / - that everything was over' *il ricatto di Luca a Maria* 'Luca's blackmailing of Mary'
3	*3*	*l'educazione dei bambini al rispetto dell'ambiente da parte dei genitori* 'the parents' education of children about respect for the environment' ◊ object predicate structures *l'elezione di Luca a sindaco da parte dei cittadini* 'Luca's election as mayor by the citizens'
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 5: Predicate nominalization of deverbal nouns derived from intransitive and pronominal verbs.

	argument nominalization	
Verb arity	*Noun arity*	*examples*
2	*0*	◊ object nominalization *l'accusato* 'accused' ◊ object nominalization: result *l'acquisto* 'a purchase'
	1	◊ subject nominalization: nomina agentis *un venditore di libri* 'book seller'
3	*1*	◊ indirect object nominalization *il destinatario di un regalo* 'addressee of a gift'

Table 6: Argument nominalization of deverbal nouns derived from transitive verbs.

Deadjectival and non-deverbal predicative nouns were also assigned an argument structure, similar to the one ascribed to deverbals, e.g.: *la possibilita' per Luca di fare qualcosa* The possibility for Luca to do s.thing'; *l'obiettivo di Leo di vincere* 'Leo's objective to win'; *la paura di Luca del buio* 'Luca's fear of the dark'.

As for simple nouns, the general principle is that concrete nouns are generally not frame-bearing. Abstract nouns, on the other hand, may take a complement when the lexical unit denotes some an inherent property *la grandezza della casa* 'the size of the house', a relationship *l'amico di Luca* 'Luca's friend', a dimension *una distanza di 3 metri* 'a 3-metre distance', an interval *una pausa di dieci minuti* 'a 10-minute pause', when it indicates a group or collection *un'assemblea di persone* 'a gathering of people' or sometimes when it is used in a metaphorical sense *la chiave del problema* 'the key of the problem'. Some nouns also require a complement which specifies their meaning *un sacco di farina* 'a bag of flour', *la citta' di Roma* 'the city of Rome'. By contrast, PPs denoting possession *il libro di Leo* 'Leo's book', free relation *la scuola di Leo* 'Leo's school', kind of constituency *tavolo di legno* 'wooden table', part of *pagina di un libro* 'page of a book' were not considered as subcategorized for by the lexical entry.

4.3 Adjective Patterns

A peculiarity of adjectives is the relevance of their distributional properties to their syntactic structure. Adjectival phrases may in fact be used both predicatively and attributively depending on their position with respect to the nominal phrase. Information about their function, prenominal or postnominal position in attributive uses[9] and (non-)gradability are therefore stipulated in adjective lexical entries.

Since the possibility of being used with both functions is a feature shared by most Italian adjectives, it seemed to us wiser to avoid the redundant and labour intensive encoding of both syntactic behaviours for each entry and

9 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 adjective position is to be understood, in most cases, as an indication of preferential behaviour rather than as an absolute constraint.

to describe such items in a unique, frameless Syntactic Unit, with the specification of their double function. 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 *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: *e' opportuno partire* 'leaving is necessary' or *e' 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.

5. Concluding Remarks

The creation of language resources is largely acknowledged as being both a complex and expensive task. The need to reuse existing data and to avoid the costly and time consuming effort of constructing a new lexicon for each system is therefore more and more pressing. However, language resources are often created from too specialized approaches which render the resulting data inadequate for further uses. Language 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.

The lexical resources developed in the framework of the PAROLE project share both a common architecture and a managment software tool. Thanks to their background, they are declarative, theory and application independent and multifunctional. The PAROLE resources are able to be easily enriched with the integration of other levels of information: the third step of the PAROLE project consists in fact of the addition of a semantic layer, to be performed in the framework of the SIMPLE project which has just started.

They may be enlarged in coverage and size or — in virtue of their uniformity — become multilingual. These characteristics which answer the requisite of genericity, explicitness, and variability of granularity confer a considerable value to the produced resources. They ensure an easy maintenance of data and a large scale reusability in different theoretical and application frameworks. 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 variety of languages used. Besides an internal use within the PAROLE Consortium, the PAROLE resources, which will be broadly available trough ELRA, will find practical applications in NLP systems development, information retrieval, language learning and machine translation systems. The customization of resources guaranteed by the granularity of the PAROLE model will allow application-dependent data and applicative lexica to be derived from the generic lexica — in the desired application format and according to user defined criteria — through dedicated mappers. The quality level of the PAROLE resources which have been internally validated by each partner through an integrity checking procedure will undergo an external validation performed, through ELRA, by external industrial users.

The Italian instantiation of the PAROLE syntactic lexicon presents a two fold advantage. On one hand it maintains its own specificity regarding the treatment of a large number of linguistic phenomena, such as the splitting of entries, the arity of frame-bearing lexical units, the assignment of syntactic functions, the clustering of syntactic realizations and more generally in the handling of language-specific phenomena as well as in some encoding decisions. On the other hand, 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 application 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 the Shallow PARsing and Knowledge extraction for Language Engineering project (LE-SPARKLE). It will also constitute the initial nucleus of a larger lexicon, based on PAROLE specifications, to be developped in the framework of a national project.

References

AA. VV. (1990). The EUROTRA Reference Manual, 7.0, Luxembourg: Commission of the EC.

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

Allegranza, V., Mazzini, G., Ruimy, N. (1995). MLAP93-08B Project: COnstraint-based Linguistic Specifications for ITalian (COLSIT), final report, December.

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

Calzolari, N., Baker, M. and Kruyt, T.(eds., 1996). Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study, in Linguistica Computazionale XI, Giardini, Pisa.

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

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

Grimshaw, J.(1990). Argument Structure, The Mit Press, Cambridge, MA.

Montemagni, S. & Pirrelli, V. (1996a). Verb Subcategorization in Italian, Blueprint of PAROLE

Guidelines to the Encoding of Syntactic Information in the Lexicon, Pisa.
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.
Renzi, L. - Salvi G. (Eds.) (1988). Grande Grammatica italiana di consultazione, voll. I-III, 1988/91/95, Il Mulino, Bologna, I.
Ruimy, N., Battista, M., Corazzari, O., Gola, E., Spanu A. (1997) Italian Lexicon Documentation, LE-PAROLE, WP3.11, Pisa.
Sanfilippo, A. et al. (1996). Subcategorization Standards, Report of the Eagles/Lexicon/Syntax Group.
Schwarze, C. (1995). Grammatik der italienischen Sprache, vol. 2., verbesserte Auflage, Niemeyer Verlag, Tuebingen.
Walker, D., Zampolli, A., and Calzolari N, (eds., 1995). Automating the Lexicon: Research and Practice in a Multilingual Environment, Proceedings of a Workshop held in Grosseto, Oxford University Press, Oxford.
Zampolli, A. (1996). "Introduction", in N. Calzolari, M. Baker, and T. Kruyt (eds.), *ibid*.