Nilda RUIMY, Istituto di Linguistica Computazionale-CNR,
Ornella CORAZZARI, Consorzio Pisa Ricerche,
Elisabetta GOLA, Consorzio Pisa Ricerche,
Antonietta SPANU, Consorzio Pisa Ricerche,
Nicoletta CALZOLARI, Istituto di Linguistica Computazionale-CNR,
Antonio ZAMPOLLI, Istituto di Linguistica Computazionale-CNR

LE-PAROLE Project: The Italian Syntactic Lexicon

Abstract

This paper presents a large scale Syntactic Computational Lexicon, representative of Italian modern language use. The entries were in fact selected on frequency criteria from the ILC Corpus and the syntactic patterns encoded were partly inferred from their contexts of occurrence. This lexicon was elaborated in the framework of the EC LE-PAROLE project which developed large, generic and re-usable written language resources in most EU languages. The lexica linguistic specifications, based on EAGLES recommendations and on the GENELEX model, were implemented in the LE-PAROLE model. This paper aims to provide an overview of the Italian instantiation of the PAROLE syntactic lexicon and of the coverage of the syntactic structures handled. It illustrates language-specific linguistic and lexicographic choices concerning crucial issues to a lexicon building process.

Keywords: Reusable Resources, Computational Lexicology, Syntax.

1. Introduction

As a result of the complexity and elevated cost of creation of new language resources, in the last decade, the NLP community has become more and more interested in issues of reusability of linguistic resources. PAROLE, a LE project funded by CEC-DGXIII and carried out by the PAROLE Consortium[1] falls within this trend of developing generic, multifunctional textual and lexical resources. The effort has involved building monolingual harmonised corpora of at least 20 million words for 14 languages[2], and 20,000 entries monolingual lexica with morphological and syntactic information for 12 languages.[3] In this paper we present the syntactic layer of the Italian Computational Lexicon built in the framework of this project. We illustrate both the general structure of a PAROLE lexicon and the specificity of its Italian instantiation. In particular, we focus on the motivations for choosing among alternative linguistic and lexicographic options which are specific to Italian. An overview of the syntactic patterns encoded for verbs, nouns and adjectives will then allow the syntactic coverage of the Italian lexicon to be estimated.

2. Linguistic Specifications and Representational Model

The aim of favouring the reusability of the resources can be achieved by relying on the most generic lexical architecture and descriptive language. The linguistic specifications for the PAROLE lexica are not committed in fact to any particular linguistic theory nor application framework. The general linguistic guidelines of the project for the lexicon syntactic level (Calzolari et al. 1996; Flores 1996) are based on the recommendations of the EAGLES / Lexicon / Syntax group (Sanfilippo et al. 1996), that provided a general scheme for verb encoding, and on

the extended GENELEX model for handling other categories (Flores 1996). Conformity and consistency to the model of the twelve lexica are guaranteed by the use of a common software tool for data management — an adaptation of the EUREKA GENELEX project tools (AA.VV. 1993).

3. The Parole Italian Syntactic Lexicon

Building a computational lexicon implies facing specific problems which derive from the need of structuring a formally consistent lexical database and of accounting for wide-ranging and flexible linguistic phenomena. The Parole Italian Syntactic Lexicon conforms to the Italian instantiations of the project linguistic specifications (Montemagni & Pirrelli 1996a-b). However, as the encoding process went on, we felt the need of setting more precise criteria in an attempt to provide lexicographers with an explicit, coherent and straightforward solution for each linguistic problem they would face (Ruimy et al. 1997). These criteria were conceived with the aim of guaranteeing that the description of the actual usage of lemmas would comply with the constraints from the conceptual model. Some of the linguistic and lexicographic decisions adopted for verbs, nouns and adjectives will be illustrated below.

3.1. Selection of Lexical Units

In the first phase of the project, the 20,000 lemmas to be encoded at syntactic level were selected on frequency basis from the ILC Italian Reference Corpus (IRC)[4] (Bindi et al. 1991). The resulting selection consisted of 3,000 verbs, 13,000 nouns, 3,000 adjectives, 500 adverbs and 500 empty / grammatical words, belonging to the general contemporary Italian language.

3.2. Reading Distinctions: some Criteria

Before starting the actual coding, it was necessary to state explicitly when lexical entries had to be split into readings. As a general rule, both redundancy and over-powerful gatherings had to be avoided. In each particular situation, however, the final choices were guided simultaneously by linguistic requirements and by the constraints from the representational model and the coding formalism. As for the linguistic aspect, in the Italian syntactic lexicon syntactic-based criteria (a-c; 1-4) were clearly the relevant ones. Arity and function assignment differences (a-c) were in fact patently criterial for the splitting of entries:

a. *disporre i libri negli scaffali*	/	*disporre di due auto*
'to put books on the shelves'		'to have two cars at one's disposal'
b. *la leggerezza di una piuma*	/	*ha commesso una leggerezza*
'the lightness of a feather'		'he was too lenient'
c. *uomo appassionato di musica*	/	*amante appassionato*
'man who has a passion for music'		'passionate lover'

Many other cases presenting less striking dissimilarities between syntactic structures emerged. These required splitting a lexical unit into different readings:

1. The optionality of a complement was not always allowed for all readings. This phenomenon was especially observed in non-literal senses[5]:

evadere (dal carcere) 'to escape from prison'	/	**evadere (dalla realtà)* 'to escape from reality'
un uomo armato (di fucile)	/	**un uomo armato (di buone intenzioni)*
'a man armed with a gun'		'a man armed with good intentions'

2. The syntagmatic realization of a complement were different:
i.a. *Luca evita Maria* 'Luca avoids Maria'

i.b. *Luca ha evitato* 'Luca has avoided' → *che Maria si ferisse* 'that M. has injured' / *una sciagura* 'a disaster' / *di dover partire* 'leaving'

3. The linking to another entry was not relevant for all readings, e.g. in case of the relationship between transitive and reciprocal verbs:

I romani combatterono i greci 'the romans fought the greeks'	↔	*I romani e i greci combatterono* 'the romans and the greeks fought each other'
Luca combatte le ingiustizie 'Luca fights the injustices'	↮	* *Luca e le ingiustizie si combattono* 'Luca and the injustices fight each other'

4. Nominalization did not occur with all verb readings:

rialzare i prezzi 'to increase the prices' / *il rialzo dei prezzi* 'the increment of the prices'
vs *rialzare la testa* 'to lift up the head' / * *il rialzo della testa*

Semantic considerations were therefore accounted for only in so far as they had consequences at syntactic level, as in cases 2 and 4 above.

3.3. A PAROLE Lexical Entry

3.3.1. Representational Structure

The structure of lexical representation is shared by all PAROLE lexica. Different sets of descriptive objects are available according to the linguistic level to be described. From a syntactic point of view, a lexical unit is encoded as a (set of) Syntactic Unit(s) which describes the syntactic behaviour of a Morphological Unit in a particular context. It consists of a Base Description and, optionally, of Transformed Description(s) encoding closely related syntactic surface alternations (e.g.: causative alternation). A Description comprises two main objects: a Self where the lemma's properties — in the specific reading described — are stored and a Construction which encodes the entry's subcategorization pattern. A Construction consists of a canonically ordered list of Positions or frame slots. Each Position is provided with the linguistic information identifying the position occupant complement. The PAROLE model makes also provision for relating lexical information throughout the lexicon by means of two descriptive devices: a) within a Syntactic Unit, the different positions of a Base and a Transformed Descriptions may be linked to each other through the Frameset mechanism; b) Syntactic Units encoding different parts of speech, but having some kind of relationship, (e.g.: a deverbal noun and its verbal base) may be related through the TransfUsyn device.

3.3.2. Information Content

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 lexica. For Italian, a syntactic entry encodes the specific properties / restrictions of a lemma and of its sub-categorizing elements in a given syntactic structure: it describes the lexically-governed syntactic context. By contrast, 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 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 the idiosyncratic behaviours w.r.t. to grammatical rule's application are therefore stipulated in the lexicon. As shown in figure 1[6], 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, its optionality, any relevant morphosyntactic or lexical constraint and any link, whenever relevant, to other slot fillers. Constraints on the headword, in the particular reading being described — i.e.: auxiliary selection for verbs, mass/count distinction for nouns, pre or postnominal position for adjectives, etc. — are expressed outside the complements' description, in the SELF.

In the following of this section, the representation in an Italian entry of the information regarding the headword's subcategorization frame will be illustrated.

3.3.2.1. Paradigmatically-related position occupants

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 and easy solution for maintaining the syntactic patterns as simple as possible. However this would, on 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. The clustering of the different realizations of each position in a single description (fig. 1), insofar as all their combinations produce grammatical sentences, as in the example in figure 2, was therefore adopted as a linguistically sounder solution.

```
[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]
                                        [coref:COI]
P3[opt:yes]:[function:indirectobject]
                    [cat:pp] [introd:a] [coref:COI]]
[SELF: Intervconst: V[func:head][morphsubcat:main]
[aux:avere]]
```

Figure 1: Representation of multiple realizations of positions.

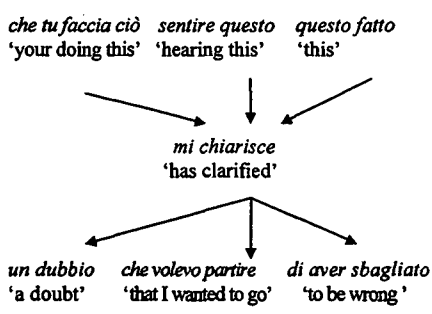

Figure 2: Multiple realizations of complements.

3.3.2.2. Predicate arity and complements' description

Predicates' arity have been considered as a language-specific parameterizable information within the PAROLE project. In the Italian lexicon, it has been limited to four arguments. As for which elements should be considered as subcategorized for, the PAROLE guidelines propose a rather liberal frame definition. A distinction is in fact drawn between lexically-governed and non lexically-governed syntactic contexts rather than between arguments and adjuncts. A position filler (even though traditionally regarded as an adjunct) is considered as syntactically strongly-bound — and is referred to as a complement — provided that it is lexically-governed

by the head (Calzolari, Montemagni & Pirrelli. 1996). 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.

Once identified, complements are marked as for their obligatoriness. For verbs, complement optionality was assessed by considering nuclear, unmarked contexts and by referring to corpus data for dubious cases. The optionality of noun complements, which is a more controversial issue, was on the other hand less easy to determine. Following the linguistic tradition, simple nouns 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 nouns complements were encoded as obligatory in non-literal meanings, e.g.: *la chiave del problema* 'the key of the problem'; *la fioritura delle arti* 'the flourishing of arts'.

The assignment of a syntactic function to each position occupant was sometimes tricky. In particular, distinguishing betwen oblique / prepositional object and adverbial functions for verb complements was not always straightforward. Some criteria for their assignment were therefore established[7] in order to avoid coding discrepancies. The adverbial function was assigned to Pp complements in alternative distribution with adverbs and whose interrogative form was built with interrogative adverbs, e.g.: *arrivare alle dieci / tardi (quando?)* 'to arrive at ten / late (when?)'. These complements bear a semantic information of manner, measure, time, location or direction. The adverbial label was also assigned to Pps or Advps which, together with the verb, convey an idiomatic meaning, i.e.: *saltare agli occhi* 'to jump out at someone'. The prepositional object function, on the other hand, was ascribed to Pps not substitutable by adverbs and whose interrogative form was built with personal pronouns, e.g.: *dare la vita per i figli (per chi?)* 'to give one's life for one's children (for whom?)'. These complements denote a beneficiary, an instrument or a cause. Pps introduced by strongly-bound prepositions, as in *dedicarsi a qualcosa* 'to devote oneself to sth.' were also attributed this function. As far as nouns and adjectives are concerned, no specific syntactic function was assigned either to simple nouns[8] or adjectives complements. By contrast, deverbal nouns complements were implicitly ascribed functions, through the derivational links established by means of the *TransfUsyn* device which enables deverbal nouns and verb frame slots to be related.

In addition to these types of information, constraints enforced on arguments were described through position level features. They encode morphosyntactic information of mood, agreement, lack of determination, lexical specification of complements introducers, control information in infinitive clauses and non-coreference of subjects between completive and matrix clauses.

3.3.2.3. Relating lexical information

In the Italian lexicon, information was also provided on some diathesis alternations such as causative-inchoative, e.g.: *chiudere* (to close), locative *schizzare* (to splutter), instrument subject *scrivere* (to write), simple reciprocal alternation with both transitive *unire* (to unite) and intransitive verbs *coincidere* (to coincide). Verbs undergoing these frame alternations were therefore encoded, using the Frameset mechanism, as complex syntactic units with two different descriptions. For the verb *rompere* (to break), for example, a Frameset named 'causative' was invoked which relates the slots of the transitive and intransitive frames described, linking thus causative readings objects to inchoative subjects.

4. Encoded Linguistic Structures

The number of lexical units handled during the lexicon building process being quite large, we daresay that most of the syntactic structures relevant in modern Italian have been identified. Following is an overview of the encoded patterns which allows the lexicon coverage to be estimated. Some observations on the linguistic data, in particular on verbs and adjectives' behaviour, were drawn from the encoding phase. They are illustrated in the following sections.

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 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. Zero to tetravalent structures[9] 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. Out of the whole number of verb readings encoded, the large prevalence of transitive constructions is evidenced in fig. 3.

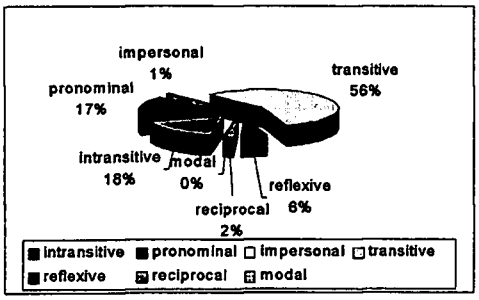

Figure 3: Verb type partition for 7155 verb readings

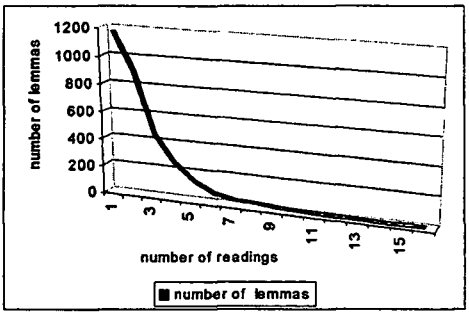

Figure 4: Reading distinction for 3090 verbs.

On the basis of the syntactic-based criteria adopted (3.2) an average number of 2,5 readings, ranging from 1 to 16, was distinguished for encoding verbs. Figure 4 illustrates verbs' complexity by showing that from a total number of 3090, 1170 verbs only (i.e. 38%) could be described by means of a single syntactic structure. Two or three patterns were used to describe other 1347 verbs (about 44%). 82% of verbs were therefore encoded by means of 1 to 3 readings. The description of 32 highly frequent verbs such as *dare* 'to give', *fare* 'to do', *mettere* 'to put', etc. required a number of syntactic structures ranging from 10 to 16.

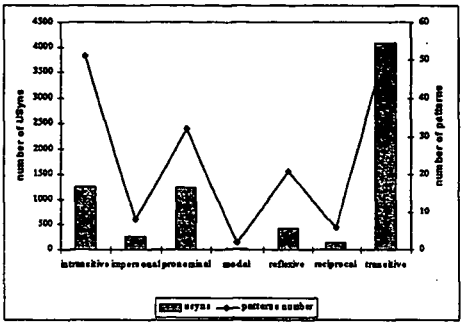

Figure 5: The different patterns for each verb class

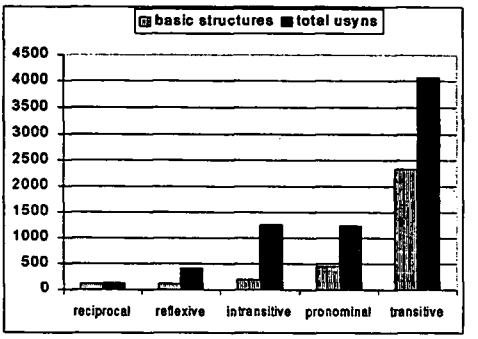

Figure 6: Incidence of basic structures readings.

Figure 5 illustrates the relationship between the different verb classes and the syntactic patterns: 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 6 indicates that transitive verbs occur mainly in basic (subject/object) constructions while intransitive verbs occur more frequently in complex rather than in simple (subject) structures.

4.2. Noun Patterns

In literature, less attention is devoted to noun patterns than to verb one. From the encoding of 13,000 nouns what emerges is that their syntactic patterns present a different level of complexity according to the semantic classes they belong to: the more concrete the meaning, the simpler the syntactic pattern. 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 fig. 9. Other simple nominals, both abstract and concrete, require a complement specifying them (fig. 10). By contrast, Pps denoting possession *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. Corpus evidence suggested that some nouns take an obligatory complement when used in a metaphorical sense. For example, the lemma *fonte* 'source' always occurs with a Pp complement in readings such as *fonte dell'errore* 'source of the mistake'. By contrast, it is mainly used without argument in the literal sense (85%).

reading denotation	examples	
inherent/abstract property	la grandezza / bellezza della casa	'the largeness / beauty of the house'
relation	l'amico / zio / capo di Luca	'Luca's friend / uncle / boss'
dimension	una distanza / lunghezza di 3metri	'a 3-metre distance / length'
interval	una pausa / intervallo di 20 minuti	'a 20-minute pause / interval'
group	una combriccola / assemblea di ladri	'a gang / gathering of thieves'
collection	un campionario / carnet di disegni	'a sample / booklet of drawings'

Figure 9: simple nouns requiring a phrasal argument.

complement denotation	examples	
lemma specification	il centenario di / l'abbondanza di qlco	'one hundred years of / the abundance of sth.'
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 Roma	'the river Po' / the city of Rome'

Figure 10: simple nouns complements.

Event or state denoting nouns as *arrivo* 'arrival' or *desiderio* 'desire' usually derive from verbs and share their complex structures, e.g.: *il giudice legge il verdetto / la lettura del verdetto da parte del giudice* 'the judge reads the verdict / the reading of the verdict by the judge'. Fig. 11 and 12 account for all patterns of deverbal nouns handled in 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 dell'assassino* 'the murderer's repentance'	0	◊ subject nominalization: nomina agentis & instruments • *un fumatore* 'a smoker' • *un galleggiante* 'a float'
2	2	• *la collaborazione di Luca al lavoro* 'Luca's collaboration in the work' • *il desiderio di Luca di viaggiare* 'Luca's desire to travel'	1	◊ subject nominalization: nomina agentis • *un affiliato al club* 'an affiliate to a club' • *un partecipante al congresso* 'a congress participant'

Figure 11: 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 Maria was innocent' *il consiglio di Luca a Maria* 'Luca's advice to Mary'	0	◊ object nominalization *l'invitato* 'the guest' ◊ object nominalization: result *un'invenzione* 'an invention'
			1	◊ subject nominalization: nomina agentis *uno scrittore di romanzi* 'a novel writer'
3	3	*l'istigazione degli operai allo sciopero da parte dei sindacati* 'the incitement of the workers to the strike by the trades-union' ◊ object predicate structures *la designazione di Luca a presidente da parte dei soci* 'Luca's nomination to president by the members'	1	◊ indirect object nominalization *l'affittuario della casa* 'the tenant of the house'
4	4	*il trasporto di merce da Pisa a Lucca da parte della ditta* 'the transport of goods from P. to L. by the compagny'		

Figure 12: deverbal nouns derived from transitive verbs

Deadjectival and non-deverbal predicative nouns were assigned an argument structure similar to the one ascribed to deverbals, e.g.: *il diritto di Luca di votare* 'Luca's right to vote'; *l'odio di Luca per le bugie* 'Luca's hatred for lies'.

4.3. 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 fig. 13, this is a feature shared by most Italian adjectives.[10] Information about their function, prenominal or postnominal position in attributive uses and (non-)gradability are therefore 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 *prima pagina* 'first page' or in free position, e.g.: *crescente interesse / interesse crescente* 'growing interest' (for frequency data, see fig. 14). Besides adjectives occurring indifferently pre or postnominally, a group of adjectives whose position confer 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.

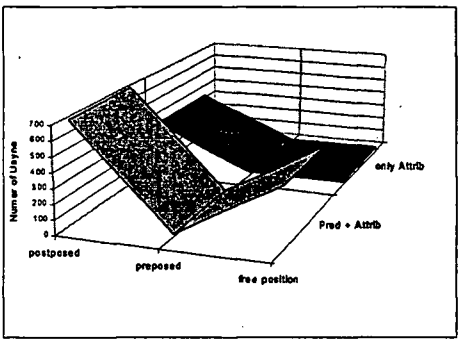

Figure13: Position and function of adjectives.

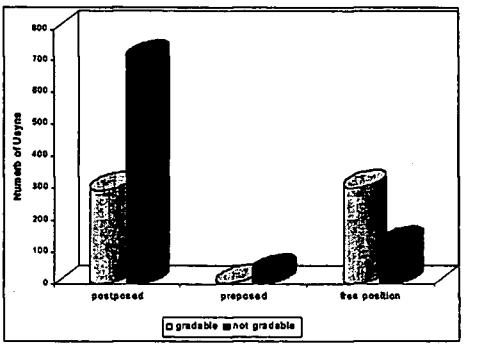

Figure 14: Position and gradability of adjectives.

5. Concluding remarks

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, descriptive language, management tool and SGML exchange format. PAROLE lexica, conceived as generic lexica easily usable by both humans and language processing systems, encode the basic information required by most NLP applications. The modularity and flexibility of the PAROLE model enables an easy maintenance of data and a straightforward enlargement or refinement of the lexical information without overall restructuration. The inclusion of semantic information, to be performed in the framework of the SIMPLE project which has just started, will in fact enrich PAROLE lexica. All these characteristics, which answer the requisite of genericity, explicitness and variability of granularity, confer to the PAROLE lexical resources a considerable value. They ensure their intra and inter consistency and a large scale reusability in NLP systems development, information retrieval, language learning and machine translation applications.

The Italian instantiation of the PAROLE syntactic lexicon is a large lexical database which presents quite interesting characteristics. First of all, it has been based on corpus data, as regards both the make-up of the entry list 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, it presents the two-fold advantage of being part of a network of European lexica, whilst preserving its specificity through the choice of the descriptive granularity and of the coding strategy as well as the treatment of a large number of language-specific phenomena. Lastly, the level of quality of its data has been validated by an integrity checking procedure (Battista, 1998) which controlled both the completeness and consistency of the information encoded.

The PAROLE Italian lexicon will constitute the initial nucleus of a larger lexicon, based on PAROLE specifications, to be developed in the framework of a national project. Encoding 20,000 entries enabled us to deal with a large number of Italian syntactic structures and to build up a lexicon that is fairly representative of the grammatical behaviour of standard Italian. The concrete experience acquired during these last two years will turn out to be precious to perform this task. It highlighted the problems a lexicographer is constantly confronted to, such as on one hand diverging opinions concerning the grain-size of lexical description and, on the other hand, the need of precise guidances which allow him to decide in a swift, easy and consistent way on the handling of phenomena for which different solutions could be appropriated. A fundamental factor determining the success of a large lexicon building process and which permits to reduce considerably the number of problems — which remain nonetheless numerous, given the versatility of language — is the availability of precise guidelines dealing with the major number of aspects of each linguistic phenomenon to be handled. The specifications elaborated a priori are undoubtedly crucial to providing the general orientation for the encoding of each category. However, they cannot focus on (and foresee a solution for) each particular aspect of all the linguistic phenomena the lexicographer will handle. The initial core of guidelines need therefore to be expanded and deepened as the lexicon building process goes on. Thanks to the experience acquired, we are now in a position to provide lexicographers with such indications as well as to propose a different treatment for some phenomena wich we felt were perhaps too constrained by the representational model within the PAROLE project.

6. Notes

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

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

[3] Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish 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.

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

[6] 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.

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

[8] Cf. Montemagni & Pirrelli (1996b:4) for discussion. However, since syntactic function is a mandatory information in the PAROLE project, generic labels such as 'Ncomp', 'Acomp', 'Aclauscomp' were used.

[9] In the Italian lexicon, the verb subject is considered as an argument.

[10] It therefore seemed to us wiser to avoid the redundant and labour intensive encoding of both syntactic behaviours for each entry and to describe such items in a unique, frameless Syntactic Unit, with the specification of their double function.

7. 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.

Battista, M. (1998) *The Parole Pisa Integrity Checker*, ILC Internal report, Pisa.

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., 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.

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, Tübingen.

Veronis J. and Ide N. (1996) *Considerations for the Reusability of Linguistic Software*, http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD1.html.