The Lexicon-Grammar Balance in Robust Parsing of Italian

Roberto Bartolini[1], Alessandro Lenci[2], Simonetta Montemagni[1], Vito Pirrelli[1]

Istituto di Linguistica Computazionale – CNR[1]
Area della Ricerca, via G. Moruzzi 1, 56100 Pisa, Italy
{roberto.bartolini, simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Università di Pisa, Dipartimento di Linguistica[2]
via Santa Maria, 36, 56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Abstract

What is the role of lexical information in robust parsing of unrestricted texts? In this paper we provide experimental evidence showing that, in order to strike the balance between robustness and coverage needed for practical NLP applications, judicious use of positive lexical evidence given a text should be complemented with a battery of dynamic parsing strategies aimed at solving local constraint conflicts. Likewise, negative lexical evidence should not blindly override grammatical information. Unlike fully lexicalised approaches to parsing where cross-categorial constraints on lexicon usage apply freely, optimal results can be obtained by modulating the way subcategorisation information is brought to bear in identifying dependency relations in context.

1. Introduction

Robustness is a key issue in nowadays NLP technology and a necessary precondition for building systems able to tackle the intricacies of real-world language. It is widely acknowledged that rich computational lexicons form a fundamental component of reliable parsing architectures and that lexical information can only have a positive effect on parsing results. In this paper we intend to tackle this issue from a slightly different and often neglected perspective. To what extent should a lexicon be trusted for parsing? Are we in a position to make a careful, objective assessment of the contribution of lexical information to overall parse success? And what does this tell us about the way general purpose and domain-specific grammatical and lexical information should optimally be complemented?

With these questions in mind, we present here an incremental approach to shallow syntactic analysis of Italian where the balanced contribution of lexical and grammatical information, together with the use of underspecified syntactic annotation, provides the necessary backbone to robust parsing. In our approach, lexical information intervenes only after a number of possibly underspecified dependency relations have already been identified on the basis of structural information only. We suggest using the described parsing architecture as an operational probe into the interplay between lexical and grammatical information for robust parsing. What emerges is a more articulated picture than commonly assumed, showing that use of lexical data in robust parsing is an extremely delicate issue, strictly related to the design of both parsing systems and computational lexicons.

2. The problem

Lexical information can interact with context in a number of ways. In case-impoverished languages such as Italian, for example, preposition selection is governed by the interaction of three factors: i) the lexical unit on which the preposition syntactically depends, ii) the syntactic relation itself and iii) the lexical head introduced by the preposition. Factor i) is dominant with “strongly bound” complements, while playing second fiddle with weakly bound modifiers. It is also observed that prepositions exhibit a higher degree of autonomy in modifiers than they do in typical arguments. For example, a temporal modifier such as *on Tuesday* remains constant whatever the verb head it modifies in context. Nonetheless, for a given relation of modification (say temporal punctuality), the specific lexical head introduced by a preposition can play an important role in selecting the preposition itself. So we have ***on** Tuesday*, but ***in** the morning* and ***at** five o’clock*.

The problem is compounded with the fact that, in many cases, the distinction between strongly and weakly bound modifiers is very difficult to draw both in practice and in principle. A frequently selected modifier is eventually perceived as endowed with argument-like properties. Another notorious culprit is the case of frame-bearing nouns. First, it is not obvious to transfer criteria for argumethood (such as optionality *vs* obligatoriness) from verbs to nouns. Secondly, deverbal nouns only occasionally “inherit” the prepositions selected by their verb base. More often they syntactically realize their candidate arguments as ordinary modifiers (*e.g. to attack someone*, but *an attack against someone*). In other more complex cases they seem to acquire the preposition selected by the support verb they typically occur with (*e.g. to kiss someone*, but *the kiss to someone*). These and other related facts make the correspondence between syntactic relations and preposition selection extremely indirect and very difficult to draw automatically. In many respects, it is somewhat reminiscent of the idiosyncratic many-to-many correspondence between inflectional endings and paradigm slots in morphology. The similarity is not surprising since prepositions form, together with inflectional endings, a closed class of grammatical items, with comparatively sparse lexical properties (Beard, 1995). It would be very difficult to model this correspondence as a problem of one-way lexical selection; here it would be more appropriate to talk about a more dynamic lexical *co-selection*.

Current computational lexicons tell only part of this story. We can reasonably hope that future generation lexicons will cover points ii) and iii) above more thoroughly than they do today. Nonetheless, it is important to appreciate that however extensive the lexicon coverage will be and however expressive its coding, such a wealth of lexical information will only throw in sharper relief the complex dynamics between points i), ii) and iii) above. It is the computational treatment of this dynamics in context that makes parsing considerably difficult. The present paper intends to scratch the surface of this problem by analysing in some detail the extent to which lexical information can be brought to bear to complement information coming from more general constraints on left-to-right word order. This analysis will hopefully not only make suggestions about the most urgently needed pieces of lexical information for parsing, but also shed light on how this information should be used in context.

3. Robust Parsing of Italian

The Italian parsing "assembly line" consists of: tokenisation of the input text; morphological analysis and lemmatisation; shallow syntactic parsing. In this paper we will focus on shallow syntactic analysis, which includes *chunking*, a process of non-recursive text segmentation, and *dependency analysis*, aimed at identifying the full range of functional relations (e.g. subject, object, modifier, complement, etc.) within each sentence. The general architecture of the parsing system adheres to the following principles:

A. modular approach – the architecture of the parsing system is highly modular, also for what concerns the internal architecture of individual components;

B. incremental monotonic analysis – the parsing flow across the different components and modules is strictly monotonic as each processing stage can only specify or further augment decisions taken at the previous steps; this is also made possible through use of underspecified syntactic categories (see C. below);

C. underspecification – underspecified output is resorted to whenever required;

D. cautious use of lexical information – lexical information is not used as a constraint on the syntactic analysis, but it is resorted to to refine and/or further specify analyses already produced on the basis of general grammatical information.

3.1. Shallow Parsing

Text chunking is carried out through a battery of finite state automata (CHUNK-IT, Federici *et al.*, 1998), which takes as input a morphologically analysed and lemmatised text and segments it into an unstructured sequence of syntactically organized text units called "chunks". Chunking requires a minimum of linguistic knowledge; its lexicon contains no other information than the entry's lemma, part of speech and morpho-syntactic features. A chunk is a textual unit of adjacent word tokens sharing the property of being related through dependency relations (es. pre-modifier, auxiliary, determiner, etc.). A chunked sentence, however, does not give information about the nature and scope of inter-chunk dependencies. These dependencies are identified during the phase of dependency analysis, carried out by IDEAL (Italian DEpendency AnaLyzer, Lenci *et al.* 2001).

IDEAL includes two main components: (i.) a core grammar of Italian; (ii.) a syntactic lexicon of ~26,400 subcategorization frames for nouns, verbs and adjectives derived from the Italian LE-PAROLE syntactic lexicon (Ruimy *et al.* 1998). The IDEAL core grammar is formed by ~100 rules covering the major syntactic phenomena.[1] The grammar rules are regular expressions (implemented as finite state automata) defined over chunk sequences, augmented with tests on chunk and lexical attributes. The rules are organized into two major modules:

- *structurally-based rules*;
- *lexically-based rules.*

A "confidence value" (PLAUS) is associated with identified dependency relations, to determine a plausibility ranking among competing analyses.

3.2. The annotation scheme

IDEAL enforces a slightly simplified version of the FAME annotation scheme, (Lenci *et al.* 1999, 2000), where functional relations are head-based and hierarchically organised to make provision for underspecified analyses. Underspecification allows IDEAL to tackle cases where lexical information is incomplete, or where ambiguous functional relations cannot be resolved (e.g. in the case of the argument vs. adjunct distinction).

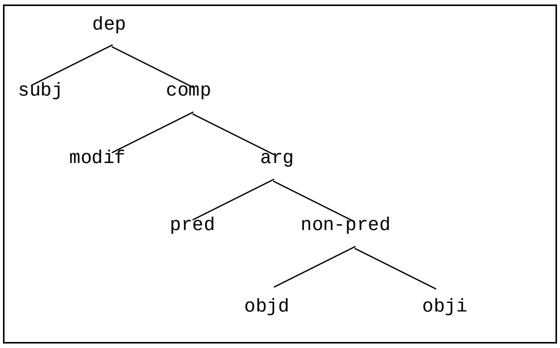

Figure 1: The FAME hierarchy

The hierarchy of the dependency relations, illustrated in Figure 1, includes:

`dep` the most general dependency relation, completely underspecified wrt its status and type;

`subj` subject;

`comp` complement, underspecified wrt its status as an argument or modifier;

`modif` modifier;

`arg` argument;

`pred` predicative argument;

[1] Adjectival and adverbial modification; negation; (non-extraposed) sentence arguments (subject, object, indirect object); causative and modal constructions; predicative constructions; PP complementation and modification; embedded finite and non-finite clauses; control of infinitival subjects; relative clauses (main cases); participial constructions; adjectival coordination; noun-noun coordination (main cases); PP-PP coordination (main cases); cliticization.

`non-pred` non-predicative argument;
`objd` direct object;
`obji` non-direct object, underspecified wrt its status as an indirect or an oblique argument.

The scheme is augmented with non head-dependent relations to annotate phenomena such as coordination and clause-internal co-referential bonds. Some extra features are associated with the participants in the dependency relation to convey, for instance, the preposition or conjunction possibly introducing the dependent, or the open/closed predicative function of a clausal dependent (to encode control information).

3.3. Lexicalised and nonlexicalised output

Consistently with the principle of incremental parsing, the nonlexicalised and lexicalised parsing stages are to be regarded as two independent and successive steps of analysis. First, IDEAL tries to identify as many dependencies as possible with no lexical information. Lexically-based rules intervene at stage 2, either to refine the output of the preceding step (e.g. by changing the ranking of identified dependencies), or to further specify types of relation based on lexicon look-up.

Le donne che dichiarano di chiedere all' uomo che modifichi il proprio comportamento vanno dal 75% al 95%.
'Women who assert that they ask the man to modify his behaviour go from 75% to 95%.'

Nonlexicalised output

```
1.  PLAUS=50 Subj(DICHIARARE[14832],DONNA[14830])
2.  PLAUS=50 Subj(DICHIARARE[14832],CHE[14831])
3.  PLAUS=50 Comp(CHIEDERE[14833],UOMO[14834]<Def=1><Intro=A>)
4.  PLAUS=50 Subj(MODIFICARE[14836],CHE[14835])
5.  PLAUS=40 ObjD(MODIFICARE[14836],CHE[14835])
6.  PLAUS=40 Subj(MODIFICARE[14836],COMPORTAMENTO[14837]<Def=1>)
7.  Modif(COMPORTAMENTO[14837]<Def=1>,PROPRIO[14837]<Role=restr>)
8.  PLAUS=50 Subj(ANDARE[14838],DONNA[14830]<Def=1>)
9.  PLAUS=50 Comp(ANDARE[14838],75%[14839]<Def=1><Intro=DA>)
10. PLAUS=40 Comp(ANDARE[14838],95%[14840]<Def=1><Intro=A>)
11. PLAUS=50 Comp(75%[14839]<Def=1>,95%[14840]<Def=1><Intro=A>)
```

Lexicalised output

```
1.  PLAUS=50 Subj(DICHIARARE[14832],DONNA[14830])
2.  PLAUS=50 Subj(DICHIARARE[14832],CHE[14831])
3.  PLAUS=60 Arg(DICHIARARE[14832],CHIEDERE[14833]<Intro=DI><Status=open>)
4.  PLAUS=60 ObjI(CHIEDERE[14833],UOMO[14834]<Def=1><Intro=A>)
5.  PLAUS=60 Arg(CHIEDERE[14833],MODIFICARE[14836]<Intro=CHE><Status=close>)
6.  PLAUS=50 Subj(MODIFICARE[14836],CHE[14835])
7.  PLAUS=40 ObjD(MODIFICARE[14836],CHE[14835])
8.  PLAUS=60 ObjD(MODIFICARE[14836],COMPORTAMENTO[14837]<Def=1>)
9.  PLAUS=40 Subj(MODIFICARE[14836],COMPORTAMENTO[14837]<Def=1>)
10. Modif(COMPORTAMENTO[14837]<Def=1>,PROPRIO[14837]<Role=restr>)
11. PLAUS=50 Subj(ANDARE[14838],DONNA[14830]<Def=1>)
12. PLAUS=60 ObjI(ANDARE[14838],75%[14839]<Def=1><Intro=DA>)
13. PLAUS=60 ObjI(ANDARE[14838],95%[14840]<Def=1><Intro=A>)
14. PLAUS=50 Comp(75%[14839]<Def=1>,95%[14840]<Def=1><Intro=A>)
```

Figure 2. Nonlexicalised and lexicalised output for the same sentence.

Figure 2 reports, for the same sentence, a slightly simplified version of the output obtained through structurally-based rules only, together with the output of the lexicalised stage. The output consists of binary relations between content words, typically a head and a dependent. The features associated with both participants in the relation convey other types of information such as: the definiteness status of a noun (DEF), the semantic type of a dependent (ROLE), the preposition introducing a certain relation (INTRO), the open/closed predicative function of clausal dependents (STATUS). The sentence in Figure 2 is described by 11 functional relations in the nonlexicalised output, and by 14 relations in the lexicalised one. The 3 relations which appear in the lexicalised output only (marked in bold) include: two sentential complements (ARG), one infinitival clause and one *che*-clause governed respectively by the verbs *dichiarare* 'assert' and *chiedere* 'ask', and one direct object (OBJD). The typology of relations which are introduced at the lexicalised analysis stage is illustrated in section 4.1.2. There are also relations in the lexicalised output which further specify the corresponding ones in the nonlexicalised output. Compare, for instance, relation 3 in the nonlexicalised output with relation 4 in the lexicalised one: the underspecified relation COMP has been interpreted as OBJI (indirect object) and has been assigned a higher plausibility value. For the typology of lexically-based analysis refinements see section 4.1.1.

4. Role of lexical information in IDEAL

In IDEAL, lexico-syntactic information intervenes only after a number of possibly underspecified dependency relations have already been identified on the basis of structural information only. At this second stage, the lexicon is accessed to provide extra conditions on parsing, so that the first stage parse can be altered in two basic ways: i) new dependency relations are identified, and ii) old underspecified relations are confirmed and assigned more specific labels. As a side effect of i) old relations can eventually be downgraded, as they happen to score, in the ranked list of possible relations, lower than their lexically-based alternatives. Furthermore, ii) is always accompanied by a reranking of the relations identified for a given sentence; from this reranking, restructuring (e.g. reattachment of complements) of the final output may follow.

The strategy is justifiable both on practical and theoretical grounds. As acquisition of lexical information is a virtually never-ending process, subject to continuous revision/refinement in the light of novel evidence from new technical domains/genres, it would be unwise to trust a lexicon once and for all. In this connection, the most commonly acknowledged problem is that of “lexical gaps”, either at the level of lemma or of the associated subcategorization frame: i.e. the text can contain lexical items or frames that are simply not attested in the lexicon. However serious this problem is, lack of positive lexical evidence is only one aspect of the damaging impact that careless use of lexical information can have on parsing. In this section we intend to analyse this impact in more detail, by illustrating how IDEAL makes use of positive lexical evidence (section 4.1), how it copes with false negative evidence (i.e. lexical gaps, section 4.2) and how it deals with false positive lexical evidence (section 4.3).

4.1. Positive lexical evidence

IDEAL makes use of positive lexical evidence to further specify already identified grammatical relations (section 4.1.1) or to detect new relations (section 4.1.2).

4.1.1. Refining structurally-based analyses

Consider the sentence fragment *si addentrò nella foresta* ‘(she) entered into the forest’. (1) and (2) below show the dependency relation identified between the verbal head *addentrarsi* and the prepositional complement *nella foresta* at the two parsing stages:

```
1.  PLAUS=50
    Comp(ADDENTRARSI,FORESTA.<Intro=IN>)
2.  PLAUS=60
    ObjI(ADDENTRARSI,FORESTA.<Intro=IN>)
```

(1) is the output of structurally-based rules only. At this stage, an underspecified dependency relation (COMP) is identified between the verb and the prepositional complement. This relation is assigned a confidence value (PLAUS) equal to 50, which is the highest confidence value assigned on the basis of structural information only. In fact, at this stage preference is given to rightmost attachments: e.g. a prepositional complement is attached with the highest confidence value to the closest, or rightmost, available lexical head. Note that this does not prevent the system from forming other attachment hypotheses, if structurally possible; however, these other hypotheses are assigned lower confidence values.

(2) is the output of the lexicalised parsing stage which refines the analysis produced at the previous stage. The verb *addentrarsi* subcategorises for a prepositional complement with the preposition *in*: the dependency relation is thus rewritten as OBJI and the assigned confidence value is 60.

Similar observations hold in the case of constructions with nominal or adjectival heads subcategorising for specific complements. The nominal construction *le ricerche di Gabriella* ‘the research of Gabriella’ is assigned the following two analyses:

```
3.  PLAUS=50 Comp(RICERCA,GABRIELLA.<Intro=DI>)
4.  PLAUS=60 Arg(RICERCA,GABRIELLA.<Intro=DI>)
```

The lexicalised interpretation further specifies the structurally-based one (COMP > ARG) with assigned an higher confidence value.

Depending on the whole sentence structure, the refinements of the analysis illustrated above may be accompanied by a restructuring of the whole dependency structure assigned to the sentence. Let us take as an example the sentence fragment *appoggiare la testa sulla sua spalla* ‘to rest the head on his shoulder’:

```
5.
PLAUS=50 ObjD(APPOGGIARE, TESTA)
PLAUS=50 Comp(TESTA, SPALLA<Intro=SU>)
PLAUS=40 Comp(APPOGGIARE, SPALLA<Intro=SU>)
6.
PLAUS=50 ObjD(APPOGGIARE, TESTA)
PLAUS=60 ObjI(APPOGGIARE, SPALLA<Intro=SU>)
PLAUS=50 Comp(TESTA<Def=1>, SPALLA<Intro=SU>)
```

In the nonlexicalised output in (5), the prepositional complement *sulla spalla* is analysed as a possible dependent of both the verbal head *appoggiare* and the nominal head *testa*. In both cases, the same underspecified COMP relation is assigned, but with different confidence values. The most likely interpretation corresponds to the rightmost attachment: i.e. COMP(*testa*, *spalla*) is assigned the score PLAUS = 50 whereas COMP(*appoggiare*, *spalla*) the score PLAUS = 40. The lexicalised output in (6) revises the analysis in (5) on the basis of the subcategorisation properties of *appoggiare*, which is a transitive verb with a prepositional complement introduced by *su*. Refinement involves here the following aspects: a) the COMP interpretation is further specified into OBJI; b) from the different confidence values assigned to the identified relations, the most likely governing head of the prepositional complement *sulla spalla* is now the verb (PLAUS = 60).

4.1.2. Introducing new dependency relations

In IDEAL lexical information is also used to enrich the output with new relations, which could not be identified with a sufficient confidence value by structurally-based rules.

The analysis of that-clauses may be quite difficult without lexical information available. Consider the sentence fragment *chiedere all' uomo che modifichi il proprio comportamento* lit. ' to ask to the man that (he) modifies his behaviour' meaning to ask the man to modify his behaviour (whose lexicalised and nonlexicalised analyses are reported in Figure 2 above). In this specific context, without specific lexical evidence the sentence fragment *che modifichi il proprio comportamento* can only be analysed as a relative clause with *uomo* as antecedent. Only knowledge of the fact that *chiedere* is a verb subcategorising for a clausal complement introduced by *che* licenses this second analysis for the same sentence fragment. In fact, IDEAL introduces the analysis as a clausal complement only at the lexicalised analysis stage (see 7 below).

```
7.  PLAUS=60
   Arg(CHIEDERE,MODIFICARE<Intro=CHE>
        <Status=close>)
```

Another relation type which is often introduced at the lexicalised stage is direct object (OBJD). In fact, due to the free constituent order of Italian we cannot exclude that a postverbal nominal constituent be a subject, if it agrees with the verb. In cases like this one transitivity information becomes crucial for licensing the hypothesis of a direct object relation, even if the subject interpretation cannot be in principle excluded. An example of this type can be found in relation 8 in the lexicalised output reported in Figure 1 above where the relation OBJD(*modificare, comportamento*) is added on the basis of lexical evidence.

4.2. Lexical gaps

Lexical gaps can either be at the level of lemma or, more often, of the associated subcategorisation frames. In this section we briefly illustrate the IDEAL strategy for minimising the impact of lexical gaps on the system performance.

Let us consider the verb *discutere* 'discuss' first: in our reference lexicon it is associated with a number of different subcategorisation frames, among which the transitive frame with a prepositional complement introduced by the preposition *su* 'about' is missing. This entails that there is no lexical evidence to guide the analysis of a sentence like *discutere su argomenti finanziari* 'discuss about financial topics' which would thus remain the same as the analysis assigned to it during the nonlexicalised parsing stage, i.e.:

```
8.  PLAUS=50
    Comp(DISCUTERE, ARGOMENTO<Intro=SU>)
```

A more complex case can be illustrated with the noun *colpo* 'stroke'. In the reference lexicon, *colpo* subcategorises for two different complements, respectively introduced by *di* 'of' (referring to the instrument used for hitting, e.g. a hammer) and by *a* 'on' (referring to the target, e.g. the head). Yet, in Italian the subcategorised preposition varies depending on the semantic properties of the lexical head introduced by the preposition (in the case at hand, the hit area). Consider as an example the complex nominal construction *un colpo in bocca e un colpo al cuore* lit, 'a stroke on the mouth and a stroke to the heart'. The lexicalised output appears as follows:

```
9.  PLAUS=50 Comp(COLPO, BOCCA<Intro=IN>)
10. PLAUS=60 Arg(COLPO, CUORE<Intro=A>)
```

Due to the lexical gap, the two dependencies do not receive a symmetric dependency analysis. However, it is important to point out that, regardless of the lexical gap, the system is able to detect in any case a dependency relation holding between *colpo* and *bocca*.

4.3. False positive evidence

For the sake of concreteness, consider the following example:

11. *discutere con maggior serenità*
lit. 'discuss with more peace'

```
11.a    PLAUS=60
        ObjI(DISCUTERE,SERENITA'<Intro=CON>)
11.b    PLAUS=50
        Comp(DISCUTERE,SERENITA'<Intro=CON>)
```

In the lexicalised output in 11.a, a manner modifier (*con serenità*, 'with peace, peacefully') is erroneously parsed as an argument of *discutere* ('discuss') in the light of the lexical information that *discutere* takes a *con*-argument (as in *discuss something **with** somebody*). Selectional preferences should be instrumental in filtering out the argument interpretation. For lack of this information it seems more reasonable to resort to an underspecified analysis, whereby *con serenità* is considered a simple complement of *discutere* (see the nonlexicalised output in 11.b).

However helpful, selectional preferences are not always decisive in filtering out structural ambiguities engendered by false positive evidence. Consider for example the following more complex case:

12 *Non è ancora certo se ad accompagnare il presidente del Consiglio a Mosca vi sarà anche Andreotti*

```
12.a
Modif(ESSERE, NON<Role=neg>)
Modif(CERTO, ANCORA)
PLAUS=50 ObjD(ACCOMPAGNARE,PRESIDENTE)
PLAUS=40 Comp(ACCOMPAGNARE,CONSIGLIO<Intro=DI>)
PLAUS=60 Arg(PRESIDENTE, CONSIGLIO<Intro=DI>)
PLAUS=40 Comp(PRESIDENTE, MOSCA<Intro=A>)
PLAUS=60 Arg(CONSIGLIO, MOSCA<Intro=A>)
```

In the lexicalised parse of 12.a, *a Mosca* 'to Moscow' is interpreted as an *a*-argument of *consiglio* 'advice'. In fact, in 12 *consiglio* forms part of the term *Presidente del Consiglio* 'prime minister' and takes no argument here. In turn, *a Mosca* is an argument of *accompagnare* 'accompany'. Note that selectional preferences would be of little avail here, as the wrong attachment is mainly ruled out on the basis of the information that *Presidente del Consiglio* is a multi-word expression.

Another interesting case is given below:

13 *La gendarmeria francese , impegnata in indagini sul patrimonio di Tapie*

```
13.a
PLAUS=50 Comp(IMPEGNATO, INDAGINE<Intro=IN>)
PLAUS=40 Comp(IMPEGNATO, PATRIMONIO <Intro=SU>)
PLAUS=50 Comp(INDAGINE, PATRIMONIO<Intro=SU>)
PLAUS=40 Comp(INDAGINE, TAPIE<Intro=DI>)
PLAUS=50 Comp(PATRIMONIO, TAPIE<Intro=DI>)
13.b
PLAUS=50 Comp(IMPEGNATO, INDAGINE<Intro=IN>)
PLAUS=40 Comp(IMPEGNATO, PATRIMONIO<Intro=SU>)
PLAUS=60 Arg(INDAGINE, PATRIMONIO<Intro=SU>)
PLAUS=60 Arg(INDAGINE, TAPIE<Intro=DI>)
PLAUS=50 Comp(PATRIMONIO, TAPIE<Intro=DI>)
```

Contrast the nonlexicalised parse (13.a) with the lexicalised one (13.b). Note that lexical information correctly helps to give higher preference to the hypothesis of *patrimonio* being dependent on *indagine* (rather than on *impegnato* 'engaged'). However, the same information supports a dependency between *Tapie* and *indagine*, thus blurring the correct preference of the non-lexicalised parse for a dependency of *Tapie* on *patrimonio* 'asset'. The example shows the role of parsing principles such as "rightmost attachment" in overriding lexical evidence while responding to functional needs of ease and speed of parsing.

5. Experiments and results

In this section, we describe an experiment carried out to evaluate the impact and role of lexical syntactic information on parsing results. The experiment was done on a selection of sentences extracted from the Italian Syntactic Semantic Treebank (ISST, Montemagni *et al.* 2000), consisting of about 300,000 word tokens of contemporary Italian. For our specific purposes, we extracted, from the balanced partition of the ISST corpus, a subcorpus of 23,919 word tokens, corresponding to 721 sentences (with an average sentence length of 33.18 tokens per sentence). Selection of sentences was lexically-based: we created a lexical test suite of 15 nouns, 10 verbs and 10 adjectives[2] and all sentences containing at least one occurrence of the lexical items in the test suite were selected.

The experiment consists of a two-stage run of IDEAL:

STAGE 1 (*lexicalized run*) – the test corpus has been parsed using the complete set of IDEAL automata, including the phase of lexicon look-up;

STAGE 2 (*non-lexicalized run*) – the parser has been run on the test corpus, excluding the rules that identify syntactic dependencies on the base of the lexicon look-up.

Two types of analysis have been performed on the parsing output. In the first one, we have focused on the contribution of the lexicon to increase the number of dependency relations identified. The second analysis has been aimed at evaluating the role of lexical syntactic information to determine correct prepositional complement attachments for nouns and verbs.

5.1. Analysis 1

Parsing data have been first analyzed to evaluate the increment in the number of the identified dependency relations determined by the use of syntactic lexical information. The analysis has been carried out on the subset of output dependency pairs, in which the head is one of the verbs of the test suite and the dependent is either a noun or a verb.[3] The following two tables provide the results of the analysis; the rightmost column reports the parsing precision in the two runs for each of the selected verbs.

Head	Subj	ObjD	ObjI	Arg	Comp	Total	% Prec.
Capire	9	12	1	13	1	**36**	**67**
Chiamare	25	18	4	1	9	**57**	**86**
Colpire	9	9	2	-	9	**29**	**83**
Contare	11	4	2	2	1	**20**	**85**
Dichiarare	11	10	4	5	2	**32**	**66**
Discutere	5	4	4	-	3	**16**	**75**
intervenire	6	1	3	-	1	**11**	**82**
Rivelare	8	9	-	4	4	**25**	**88**
Scoprire	5	6	1	2	7	**21**	**90**
Trovare	40	38	17	5	14	**114**	**85**

Table 1: Relations identified in the lexicalized run

Head	Subj	ObjD	ObjI	Arg	Comp	Total	% Prec.
Capire	10	9	-	-	1	**20**	**60**
Chiamare	19	12	-	-	11	**42**	**79**
Colpire	8	4	-	-	11	**23**	**87**
Contare	9	3	-	-	3	**15**	**93**
Dichiarare	11	1	-	-	5	**17**	**55**
Discutere	6	3	-	-	7	**16**	**69**
intervenire	5	1	-	-	5	**11**	**64**
Rivelare	8	3	-	1	3	**15**	**87**
Scoprire	5	3	-	-	7	**15**	**87**
Trovare	37	21	-	4	25	**87**	**80**

Table 2: Relations identified in the non-lexicalized run

Although, the figures *prima facie* show a neat degradation of the number of identified dependencies in the non-lexicalized run, a closer inspection to the data reveal a more complex situation. First of all, notice that the average precision of the non-lexicalized run is 76%, which is still a high result if compared with the average precision obtained in the lexicalized run (80%). Secondly, verbs greatly differ with respect to the impact of lexical information. In the case of *trovare* ‘find’, *capire* ‘understand’ or *chiamare* ‘call’ the lack of lexical information produces a reduction of almost 1/3 of the overall number of identified relations. Conversely, with other verbs the difference between the lexicalized and the non-lexicalized runs is almost minimal. Interestingly, this difference seems to be at least partially independent of the “richness” of the syntactic descriptions associated to the lexical entries. For instance, in the IDEAL lexicon *intervenire* ‘intervene’ and *discutere* ‘discuss’ have the same number of syntactic descriptions as *colpire* ‘hit’, but these verbs neatly differ for the ratio of unidentified dependencies when lexical information is not used. This could suggest that the differences in the results might be especially related to the “type” of the verb syntactic description, with particular concern to parameters such as frequency, prototypicality of the subcategorization pattern, etc.

The most interesting aspect of this analysis concerns the identification of subject and direct object dependencies. Actually, the figures above show that blocking the access to the lexicon has a really minimal impact on subject identification, while in the case of direct objects the difference is much higher (up to the double). Besides, these results are distributed among all the ten verbs in an almost uniform way. This contrast can be explained by taking under consideration both the specific features of Italian grammar and the principles of incremental robust parsing to which IDEAL adheres. In fact, in Italian there are non-lexical clues allowing for the identification of the verb subject, which are

[2] Nouns: *accordo, acqua, aumento, calcio, capo, centro, colpo, conto, controllo, crisi, fondo, intervento, mano, posto, rischio*. Verbs: *capire, chiamare, colpire, contare, dichiarare, discutere, intervenire, rivelare, scoprire, trovare*. Adjectives: *aperto, buono, capace, comune, disponibile, diverso, economico, libero, necessario, pronto*.

[3] For the sake of this experiment we ignore the Modif and Pred relations.

instead lacking in the case of direct object recognition. For instance, the presence of the auxiliary *essere* ‘be’ is by itself a strong evidence that the nominal head of a noun chunk can be interpreted as the subject of a finite verb. In fact, *essere* is the typical auxiliary of unaccusative intransitive, passive and reflexive verbs. Conversely, the fact that a nominal constituent appears post-verbally can not be taken as real evidence for its being a direct object, given the almost free possibility of subject inversion in Italian.

5.2. Analysis 2

The output has also been analyzed to evaluate how and to which extent lexical syntactic information contributes to identify the proper attachment of prepositional complements. In particular, we have explored the role of the lexicon in “attachment restructuring”. As already mentioned, one of the possible uses of the lexicon in IDEAL is to produce a reranking of the most plausible identified dependencies. In this case, an attachment relation is restructured, *i.e.* a prepositional complement that has been formerly attached to the closest verbal or nominal head to its left is instead re-assigned as the dependent of a farther head, on the basis of lexicon look-up. The results of the data analyses are reported in the tables below.

Total number of lexicalized attachments	953
Total number of *restructured* lexicalized attachments	104

Table 3: Lexicalized attachments

Wrong attachments	53
Correct attachment	51
Total restructured attachments	104

Table 4: Restructured attachments

Total attachments to nouns	15
Wrong attachments	10

Table 5: Restructured attachments to nouns

Total attachments to verbs	89
Wrong attachments	29

Table 6: Restructured attachments to verbs

Total un-restructured attachments to nouns	638
Wrong attachments	27

Table 6: Un-restructured attachments to nouns

Total un-restructured attachments to verbs	211
Wrong attachments	12

Table 7: Un-restructured attachments to verbs

The figures in Table 4 show that in the case of prepositional complement attachment the impact of lexical information is by and large concentrated in confirming a formerly assigned dependency relation, changing its label, rather than restructuring a non-lexicalized attachment. The latter case only applies ~10% of times. The crucial issue is however to evaluate whether even in these few cases the lexicon is able to improve the precision of the system in identifying the correct attachments.

Table 4 shows that more than a half of the 104 restructured attachments proposed by the lexicon are totally wrong. Even more interestingly, there is a strong asymmetry between nouns and verbs (Tables 5 and 6). In the latter case, the number of lexically restructured dependencies is extremely low, and above all the percentage of mistakes is particularly high (over the 66%). Vice versa, in the case of verbs, the lexicon is able to propose a bigger number of restructured attachments, and mistakes drastically reduce (32%).

The difference in the total number of restructured attachment may be due to the richer complementation patterns exhibited by verbs in the lexicon. Note however that, while in the case of verbs lexical information seems to be able to produce a significant improvement in the restructured attachment precision, this contribution drops considerably for nouns. Conversely, as shown by Tables 6 and 7, the role of lexical information in specifying an attachment dependency with no restructuring is almost uniform across nouns and verbs. One of the reasons for this situation is the fact that the vast majority of Italian nouns have an argument structure that is syntactically realized by prepositional complements headed by the preposition *di* ‘of’. *Di* is an extremely polysemous preposition, heading, among others, also possessive phrases. This produces a considerable amount of noise. Take for instance the case of the word *accesso* ‘access’, which appears in the lexicon with a two-complement frame, one headed by the preposition *di* and the other with the preposition *a* ‘to’. This corresponds to the syntactic pattern found in *l’accesso di Gianni all’Università* ‘John’s access to the University”. The problem is that, given the almost free order of complements in Italian the *di*-complement can also appear after the one headed by a, e.g. *l’accesso all’Università di Gianni*. However, because of the ambiguity of the preposition *di*, the latter pattern is perfectly compatible with a reading in which John is the owner of the University (with *Gianni* attached to *Università*). The test corpus contains a similar case: *regolare l’accesso alle risorse del Fondo*, ‘to regulate the access to the Fund’s resources’. The parser, lacking semantic information, on the base of syntactic lexicon look-up, wrongly restructures the attachment of the *di*-complement, by attaching it to *accesso*. One way to smooth the consequences of this problem - widespread in Italian – would be to impose order constraints on the complement position, but this would surely impact on the system recall in a very negative way.

6. Discussion and concluding remarks

Despite the attraction of having richly subcategorised lexical entries for parsing, the neat contribution of lexical information to parse success is still an open issue. Our data show that available lexical information cannot be relied upon blindly for parsing. First, careless use of false negative evidence (lexical gaps) may boil down to a failure to detect contextually appropriate dependency relations. This may strike the reader as a comparatively trivial remark, but we should not be too hasty to dismiss it and conclude that the easy solution is building larger or domain-specific lexicons. In fact, we are confronted here with two problems. The first one is an issue of content: what information should we strive to look for and encode in a lexicon that is instrumental for parsing? The second one is a problem of information source and methodology of acquisition: where should we look for it and how? Neither question is trivial.

What our data seems to suggest is that the lexicon-grammar balance in parsing is a problem of typology rather than quantity of the lexical information provided. Existing computational lexicons give a wealth of information including, i.) complement lists, ii.) syntactic category of the possible slot-fillers (e.g. NP, PP, etc.) iii.) distinction of complements vs. adjuncts, iv.) obligatoriness of complements, etc. The effective contribution of each of these pieces of syntactic information needs to be assessed carefully. For instance, our results suggest that information about iii) and iv) is not so crucial for parsing real texts. Moreover, the preposition introducing a complement should better be described in terms of its semantic class (e.g. temporal, locative, etc.) rather than by crude preposition lists, since preposition selection is often dependent on the noun governed by the preposition. Finally, corpus-based preferences on complement order information could have been beneficial in avoiding parse failures on our test data.

Another crucial problem in directly projecting lexical information on context is the ubiquitous presence of false positive evidence. Again, this may prompt the suggestion that false negatives are to be filtered out by using selectional restrictions. Even if we ignore the immediate non-trivial problem of collecting these restrictions and encoding them in a lexicon, this suggestion, however, strikes us as too simplistic. Our results witness a more complex dynamics between potentially contradictory constraints on parsing. Finding an optimal solution to this problem involves a number of possible strategies. For example, spotting multi-word terminology in context, without analysing the entire sentence, can, in many cases, help the parser to identify "atomic syntactic islands", thus significantly reducing the number of potential lexical heads in context. This requires, in turn, the availability of vast repositories of multi-word terms. Other variously inspired parsing principles such as "rightmost attachment" or "longest match" do play a role here. Far from being *ad hoc* heuristics, they bear witness to the need for incorporating functional principles of ease and speed of parsing into problems concerning the most likely constituent order in real texts. It should be appreciated that a principle such as "longest match" cannot be captured by any manipulation of the probabilities in a stochastic context-free grammar. "Longest match" and other related strategies involve a comparison across competing analyses, and can be expressed in terms of the context, but not in terms of context-free rewrite probabilities (Abney 1997). This may also help us to understand why many recent approaches to lexicalised stochastic parsing have achieved comparatively little success. Once more, this fact should not persuade us to prematurely dismiss use of stochastic lexicalised parsing. It should only make us more acutely aware that traditional stochastic parsing methods do a *global optimisation*. If we have a very good model of certain lexical dependencies but a poor model of other such dependencies, the ability of a parser to solve a local conflict in contexts where well- and poorly-modelled dependencies are found simultaneously will inevitably suffer. As a prospective line of investigation, we intend to explore the adaptive use of "syntactic memories" (buffers of already parsed texts) to solve local constraint-conflicts dynamically, along the lines of Data-Oriented Language Processing (Bod and Scha 1997).

7. References

ABNEY, S. P., *Part-of-SpeechTagging and Partial Parsing*, in S. Young and G. Bollthooft (eds.), *Corpus-based Methods in Language and Speech Porcessing*, Kluwer, Dordrecht, 1997, pp. 118-136.

BEARD, R., *Lexeme-Morpheme Base Morphology*, New York, SUNY Press, 1995.

BOD, R., R. SCHA, *Data-oriented Language Processing*, in S. Young and G. Bollthooft (eds.), *Corpus-based Methods in Language and Speech Porcessing*, Kluwer, Dordrecht, 1997, pp. 137-173.

FEDERICI, S., MONTEMAGNI, S., PIRRELLI, V., *Chunking Italian: Linguistic and Task-oriented Evaluation*, in Proceedings of the LREC Workshop on "Evaluation of Parsing Systems", Granada, Spain, 1998.

LENCI, A., BARTOLINI, R., CALZOLARI, N., CARTIER, E., *Document Analysis*, MLIS-5015 MUSI, Deliverable D3.1, 2001.

LENCI A., MONTEMAGNI S., PIRRELLI V., SORIA C., *FAME: a Functional Annotation Meta-scheme for Multimodal and Multi-lingual Parsing Evaluation*, in Proceeding of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation in NLP, University of Maryland, June 22nd 1999.

LENCI, A., MONTEMAGNI, S., PIRRELLI, V., SORIA, C., *Where opposites meet. A Syntactic Meta-scheme for Corpus Annotation and Parsing Evaluation*, in Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, 31 May – 2 June 2000.

MONTEMAGNI S., BARSOTTI F., BATTISTA M., CALZOLARI N., CORAZZARI O., ZAMPOLLI A., FANCIULLI F., MASSETANI M., RAFFAELLI R., BASILI R., PAZIENZA M.T., SARACINO D., ZANZOTTO F., MANA N., PIANESI F., DELMONTE R., *The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation*, in Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000)", Luxembourg, 6 August 2000, pp. 18-27.

RUIMY, N., CORAZZARI, O., GOLA, E., SPANU, A., CALZOLARI, N., ZAMPOLLI, A., *The European LE-PAROLE Project: The Italian Syntactic Lexicon*, in Proceedings of the First International Conference on Language resources and Evaluation (LREC-1998), Granada, Spain, 1998, pp. 241-248.