

Grammar and Lexicon in the Robust Parsing of Italian

Towards a Non-Naïve Interplay

Roberto BARTOLINI	Alessandro LENCI	Simonetta MONTEMAGNI	Vito PIRRELLI
Istituto di Linguistica Computazionale CNR Area della Ricerca Via Moruzzi 1 56100 PISA (Italy)	Università di Pisa Via Santa Maria 36 56100 PISA (Italy)	Istituto di Linguistica Com- putazionale CNR Area della Ricerca Via Moruzzi 1 56100 PISA (Italy)	Istituto di Linguistica Computazionale CNR Area della Ricerca Via Moruzzi 1 56100 PISA (Italy)

{roberto.bartolini, alessandro.lenci, simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Abstract

In the paper we report a qualitative evaluation of the performance of a dependency analyser of Italian that runs in both a non-lexicalised and a lexicalised mode. Results shed light on the contribution of types of lexical information to parsing.

Introduction

It is widely assumed that rich computational lexicons form a fundamental component of reliable parsing architectures and that lexical information can only have beneficial effects on parsing. Since the beginning of work on broad-coverage parsing (Jensen 1988a, 1988b), the key issue has been how to make effective use of lexical information. In this paper we put these assumptions to the test by addressing the following questions: to what extent should a lexicon be trusted for parsing? What is the neat contribution of lexical information to overall parse success?

We present here the results of a preliminary evaluation of the interplay between lexical and grammatical information in parsing Italian using a robust parsing system based on an incremental approach to shallow syntactic analysis. The system can run in both a non-lexicalised and a lexicalised mode. Careful analysis of the results shows that contribution of lexical information to parse success is more selective than commonly assumed, thus raising the parallel issues of how to promote a more effective integration between

parsers and lexicons and how to develop better lexicons for parsing.

1 Syntactic parsing lexicons

Syntactic lexical information generally feeds parsing systems distilled in *subcategorization frames*. Subcategorization is a formal specification of a predicate phrasal context in terms of the type of arguments syntactically selected by the predicate entry (e.g. the verb *hit* selects for a subject NP and an object NP). Lexical frames commonly include: i.) number of selected arguments, ii.) syntactic categories of their possible realization (NP, PP, etc.), iii.) lexical constraints on the argument realization (e.g. the preposition heading a PP complement), and iv.) the argument functional role. Other types of syntactic information that are also found in syntactic lexicons are: argument optionality, verb control, auxiliary selection, order constraints, etc. On the other hand, collocation-based lexical information is only rarely provided by computational lexicons, a gap often lamented in robust parsing system development.

A number of syntactic computational lexicons are nowadays available to the NLP community. Important examples are LDOCE (Procter 1987), ComLex (Grishman *et al.* 1994), PAROLE (Ruimy *et al.* 1998). These lexicons are basically hand-crafted by expert lexicographers, and their natural purpose is to provide general purpose, domain-independent syntactic information, covering the most frequent entries and frames. On the other hand, parsing systems

often complement general lexicons with corpus-driven, automatically harvested syntactic information (Federici *et al.* 1998b, Briscoe 2001, Korhonen 2002). Automatic acquisition of sub-categorization frames allows systems to access highly context dependent constructions, to fill in possible lexical gaps and eventually rely on frequency information to tune the relative impact of specific frames (Carroll *et al.* 1998).

Lexicon coverage is usually regarded as the main parameter affecting use of lexical information for parsing. However, the real comparative impact of the *type* (rather than the mere quantity) of lexical information has been seldom discussed. Our results show that the contribution of various lexical information types to parse success is not uniform. The experiment focuses on a particular subset of the information available in syntactic lexicons - the representation of PP complements in lexical frames - tested on the task of PP-attachment. The reason for this choice is that this piece of information occupies a central and dominant position in existing lexicons. For instance in the Italian PAROLE lexicon, more than one third of verb frames contain positions realized by a PP, and this percentage raises up to the near totality noun-headed frames.

2 Robust Parsing of Italian

The general architecture of the Italian parsing system used for testing adheres to the following principles: 1) modular approach to parsing, 2) underspecified output (whenever required), 3) cautious use of lexical information, generally resorted to in order to refine and/or further specify analyses already produced on the basis of grammatical information. These principles underlie other typical robust parsing architectures (Chanod 2001, Briscoe and Carroll 2002).

The system consists of i.) CHUNK-IT (Federici *et al.* 1998a), a battery of finite state automata for non-recursive text segmentation (*chunking*), and ii.) IDEAL (Lenci *et al.* 2001), a *dependency-based* analyser of the full range of intra-sentential functional relations (e.g. subject, object, modifier, complement, etc.). CHUNK-IT requires a minimum of lexical knowledge: lemma, part of speech and morpho-syntactic features. IDEAL includes in turn two main components: (i.) a Core Dependency Grammar of

Italian; (ii.) a syntactic lexicon of ~26,400 sub-categorization frames for nouns, verbs and adjectives derived from the Italian PAROLE syntactic lexicon (Ruimy *et al.* 1998). The IDEAL Core Grammar is formed by ~100 rules (implemented as finite state automata) covering major syntactic phenomena,¹ and organized into *structurally-based rules* and *lexically-based rules*. IDEAL adopts a slightly simplified version of the FAME annotation scheme (Lenci *et al.* 2000), where functional relations are head-based and hierarchically organised to make provision for underspecified representations of highly ambiguous functional analyses. This feature allows IDEAL to tackle cases where lexical information is incomplete, or where functional relations cannot be disambiguated conclusively (e.g. in the case of the argument vs. adjunct distinction). A “confidence score” is associated with some of the identified dependency relations to determine a plausibility ranking among different possible analyses.

In IDEAL, lexico-syntactic information intervenes only after possibly underspecified dependency relations have 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 non-monotonically altered in various ways (see section 3.3). This strategy minimises the impact of lexical gaps (whether at the level of lemma or of the associated sub-categorization frames) on the system performance (in particular on its coverage).

3 The Experiment

3.1 The Test Corpus (TC)

The test corpus contains a selection of sentences extracted from the balanced partition of the Italian Syntactic Semantic Treebank (ISST, Montemagni *et al.* 2000), including articles from

¹ Adjectival and adverbial modification; negation; (non-extrapolated) 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.

contemporary Italian newspapers and periodicals covering a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). TC consists of 23,919 word tokens, corresponding to 721 sentences (with a mean sentence length of 33.18 words, including punctuation tokens). The mean number of grammatical relations per sentence is 18.

3.2 The Baseline Parser (BP)

The baseline parser is a non-lexicalised version of IDEAL including structurally-based rules only. The mean number of grammatical relations per sentence detected by BP in TC is 15.

The output of the baseline parser is shallow in different respects. First, it contains underspecified analyses, resorted to whenever available structural information does not allow for a more specific syntactic interpretation: *e.g.* at this level, no distinction is made between arguments and modifiers, which are all generically tagged as “complements”. Concerning attachment, the system tries all structurally-compatible attachment hypotheses and ranks them according to a confidence score. Strong preference is given to rightmost attachments: *e.g.* a prepositional complement is attached with the highest confidence score (50) to the closest, or rightmost, available lexical head. In the evaluation reported in section 4, we consider top-ranked dependents only, *i.e.* those enforcing rightmost attachment. Moreover, in matching the relations yielded by the parser with the ISST relations in TC we make allowance for one level of subsumption, *i.e.* a BP relation can be one level higher than its ISST counterpart in the hierarchy of dependency relations. Finally, the BP output is partial with respect to those dependencies (*e.g.* a *that*-clause or a direct object) that would be very difficult to identify with a sufficient degree of confidence through structurally-based rules only.

3.3 The Lexically-Augmented Parser (LAP)

The lexically-augmented version of IDEAL includes both structurally-based and lexically-based rules (using the PAROLE lexicon). In this lexically-augmented configuration, IDEAL first tries to identify as many dependencies as possible with structural information. Lexically-based rules intervene later to refine and/or complete structurally-based analyses. Those structurally-

based hypotheses that find support in the lexicon are assigned the highest score (60). The contribution of lexically-based rules is *non-monotonic*: old relations can eventually be downgraded, as they happen to score, in the newly ranked list of possible relations, lower than their lexically-based alternatives. Furthermore, specification of a former underspecified relation is always accompanied by a *re-ranking* of the relations identified for a given sentence; from this re-ranking, restructuring (*e.g.* reattachment of complements) of the final output may follow.

LAP output thus includes:

- a) fully specified dependency relations: *e.g.* an underspecified dependency relation such as “complement” (COMP), identified by a structurally-based rule, is rewritten, when lexically-supported, as “indirect object” (OBJI) and assigned a higher confidence value;
- b) new dependency relations: this is the case, for instance, of *that*-clauses, direct objects and other relation types whose identification is taken to be too difficult and noisy without support of lexical evidence;
- c) underspecified dependency relations, for those cases that find no lexical support.

The mean number of grammatical relations per sentence detected by LAP in TC is 16. In the evaluation of section 4, we consider top-ranked dependents only (confidence score ≥ 50), corresponding to either lexically-supported dependency relations or – in their absence – to rightmost attachments. Again, in matching the relations yielded by the parser with the ISST relations in TC we make allowance for one level of subsumption.

4 Analysis of Results

The parsing outputs of BP and LAP were compared and projected against ISST annotation to assess the contribution of lexical information to parse success. In this paper, we focus on the evaluation of how and to which extent lexico-syntactic information contributes to identification of the proper attachment of prepositional complements. For an assessment of the role and impact of lexical information in the analysis of dependency pairs headed by specific words, the interested reader is referred to Bartolini *et al.* (2002).

4.1 Quantitative Evaluation

Table 1 summarises the results obtained by the two different parsing configurations (BP and LAP) on the task of attaching prepositional complements (PC). Prepositional complements are classified with respect to the governing head: PC_VNA refers to all prepositional complements governed by V(erb)al), N(ominal) or A(dj)ectival) heads. PC_V is the subset with a V(erb)al) head and PC_N the subset with a N(ominal) head. For each PC class, precision, recall and f score figures are given for the differ-

ent parsing configurations. Precision is defined as the ratio of correctly identified dependency relations over all relations found by the parser ($\text{prec} = \text{correctly identified relations} / \text{total number of identified relations}$); recall refers to the ratio of correctly identified dependency relations over all relations in ISST ($\text{recall} = \text{correctly identified relations} / \text{ISST relations}$). Finally, the overall performance of the parsing systems is described in terms of the f score, computed as follows: $2 \text{ prec recall} / \text{prec} + \text{recall}$.

	ISST	BP			LAP		
		Prec	recall	F score	Prec	recall	f score
PC_VNA	3458	75,53	57,40	65,23	74,82	61,02	67,22
PC_V	1532	75,43	45,50	56,76	74,23	49,50	61,22
PC_N	1835	73,53	80,82	77,00	72,76	81,36	76,82

Table 1. Prepositional complement attachment in BP and LAP

	Lexicalised atts			Confirmed atts			Restructured atts		
	total	OK	prec	Total	OK	prec	total	OK	prec
PP_VNA	919	819	89,12	816	771	94,49	103	65	63,11
PP_V	289	244	84,43	201	194	96,52	88	61	69,32
PP_N	629	575	91,41	614	577	93,97	15	4	26,67

Table 2. Lexicalised attachments

To focus on the role of the lexicon in either confirming or revising structure-based dependencies, lexically-supported attachments are singled out for evaluation in Table 2. Their cumulative frequency counts are reported in the first three columns of Table 2 (“Lexicalised attachments”), together with their distribution per head categories. Lexicalised attachments include both those structure-based attachments that happen to be confirmed lexically (“Confirmed attachments”), and restructured attachments, *i.e.* when a prepositional complement previously attached to the closest available head to its left is eventually re-assigned as the dependent of a farther head, on the basis of lexicon look-up (“Restructured attachments”). Table 2 thus shows the impact of lexical information on the task of PP attachment. In most cases, 89% of the total of lexicalised attachments, LAP basically confirms dependency relations already assigned at the previous stage.

Newly discovered attachments, which are detected thanks to lexicon look-up and re-ranking, amount to only 11% of all lexicalised attachments, less than 3% of all PP attachments yielded by LAP.

4.3 Discussion

4.3.1 Recall and precision on noun and verb heads

Let us consider the output of BP first. The striking difference in the recall of noun-headed *vs* verb-headed prepositional attachments (on comparable levels of precision, rows 2 and 3 of Table 1) prompts the suggestion that the typical context of use of a noun is more easily described in terms of local, order-contingent criteria (*e.g.* rightmost attachment) than a verb context is. We can give at least three reasons for that. First, frame bearing nouns tend to select fewer argu-

ments than verbs do. In our lexicon, 1693 verb-headed frames out of 6924 have more than one non subject argument (24.4%), while there being only 1950 noun-headed frames out of 15399 with more than one argument (12.6%). In TC, of 2300 head verb tokens, 328 exhibit more than one non subject argument (14%). Rightmost attachment trivially penalises such argument chains, where some arguments happen to be overtly realised in context one or more steps removed from their heads. The second reason is sensitive to language variation: verb arguments tend to be dislocated more easily than noun arguments, as dislocation heavily depends on sentence-level (hence main verb-level) phenomena such as shift of topic or emphasis. In Italian, topic-driven argument dislocation in preverbal position is comparatively frequent and represents a problem for the baseline parser, which works on a head-first assumption. Thirdly, verbs are typically modified by a wider set of syntactic satellites than nouns are, such as temporal and circumstantial modifiers (Dik 1989). For example, deverbal nouns do not inherit the possible temporal modifiers of their verb base (*I run the marathon in three hours*, but **the run of the marathon in three hours*). Modifiers of this sort tend to be distributed in the sentence much more freely than ordinary arguments.

4.3.2 Impact of the lexicon on recall

Of the three above mentioned factors, only the first one has an obvious lexical character. We can provide a rough estimate of the impact of lexical information on the performance of LAP. The lexicon filter contributes a 9% increase of recall on verb complements (4% over 45.5%), by correctly reattaching to the verbal head those arguments (61) that were wrongly attached to their immediately preceding constituent by BP. This leads to an overall 49.5% recall. All remaining false negatives (about 48%) are i) either verb modifiers or ii) proper verb arguments lying out of the reach of structure-based criteria, due to syntactic phenomena such as complement dislocation, complex coordination, parenthetical constructions and ellipsis. We shall return to a more detailed analysis of false negatives in section 4.3.4. In the case of noun complements, use of lexical information produces a negligible increase of recall: 0.6% (0.5% over 80.8%). This

is not surprising, as our test corpus contains very few cases of noun-headed argument chains, fewer than we could expect if the probability of their occurrence reflected the (uniform) type distribution of noun frames in the lexicon. The vast majority of noun-headed false negatives, as we shall see in more detail in a moment, is represented by modifiers.

4.3.3 Impact of the lexicon on precision

Reattachment is enforced by LAP when the preposition introducing a candidate complement in context is found in the lexical frame of its head. Table 2 shows that ~37% of the 103 restructured attachments proposed by the lexicon are wrong. Even more interestingly, there is a strong asymmetry between nouns and verbs. With verb heads, precision of lexically-driven reattachments is fairly high (~70%), nonetheless lower than precision of rightmost attachment (~75%). In the case of noun heads, the number of lexically reattached dependencies is instead extremely low. The percentage of mistakes is high, with precision dropping to 26.6%.

The difference in the total number of restructured attachment may be again due to the richer complementation patterns exhibited by verbs in the lexicon. However, while in the case of verbs lexical information produces a significant improvement on restructured attachment precision, this contribution drops considerably for nouns. The main reason for this situation is that nouns tend to select semantically vacuous prepositions such as *of* much more often than verbs do. In our lexicon, out of 4157 frames headed by a noun, 4015 contain the preposition *di* as an argument introducer (96.6%). *Di* is in fact an extremely polysemous preposition, heading, among others, also possessive phrases and other kinds of modifiers. This trivially increases the number of cases of attachment ambiguity and eventually the possibility of getting false positives. Conversely, as shown by the number of confirmed attachments in Table 2, the role of lexical information in further specifying an attachment with no restructuring is almost uniform across nouns and verbs.

4.3.4 False negatives

The vast majority of undetected verb complements (80.6%) are modifiers of various kind. The remaining set of false negatives consists of

48 complements (7.7%), 30 indirect objects (4.8%) and 43 oblique arguments (6.9%). Most such complements are *by*-phrases in passive constructions which are not as such very difficult to detect but just happen to fall out of the current coverage of LAP. More interestingly, 2/3 of the remaining false negatives elude LAP because they are overtly realised far away from their verb head, often to its left. Most of these constructions involve argument dislocation and ellipsis. We can thus preliminarily conclude that argument dislocation and ellipsis accounts for about 14% of false negatives (7% over 50%). Finally, the number of false negatives due to attachment ambiguity is almost negligible in the case of verbal heads.

On the other hand, the impact of undetected modifiers of a verbal head on attachment recall is considerable. The most striking feature of this large subset is the comparative sparseness of modifiers introduced by *di* (*of*): 31 out of 504 (6.2%). At a closer scrutiny, the majority of these *di*-phrases are either phraseological adverbial modifiers (*di recente* ‘of late’, *del resto* ‘besides’ etc.) or quasi-arguments headed by participle forms. Notably, 227 undetected modifiers (45% of the total) are selected by semantically heavy and complex (possibly discontinuous) prepositions (*davanti a* ‘in front of’, *in mezzo a* ‘amid’, *verso* ‘towards’, *intorno a* ‘around’, *contro* ‘against’, *da ... a* ‘from ... to’ etc.). As to the remaining 241 undetected modifiers (48%), they are introduced by ‘light’ prepositions such as *a* ‘to’, *in* ‘in’ and *da* ‘from’. Although this 48% contains a number of difficult attachments, one can identify subsets of fairly reliable modifiers by focusing on the noun head introduced by the preposition, which usually gives a strong indication of the nature of the modifier, especially in the case of measure, temporal and locative expressions.

4.3.5 False positives

Table 2 shows a prominent asymmetry in the precision of confirmed and restructured attachments. Wrong restructured attachments are mainly due to a misleading match between the preposition introducing a PC and that introducing a slot in the lexical frame of its candidate head (~85%). This typically occurs with ‘light’ prepositions (e.g. *di*, *a*, etc.). Most notably, in a

relevant subset of these mistakes, the verb or noun head belongs to an idiomatic multi-word expression. In the case of confirmed attachments, about one third of false positives (~5%) involve multi-word expressions, in particular compound terms such as *presidente del consiglio* ‘prime minister’, where the rightmost element of the compound is wrongly selected as the head of the immediately following PP. In both restructured and confirmed attachments, the remaining cases (on average ~4%) are due to complex syntactic structures (e.g. appositive constructions, complex coordination, ellipsis etc.) which are outside the coverage of the current grammar.

Conclusion

Larger lexicons are not necessarily better for parsing. The issue of the interplay of lexicon and grammar, although fairly well understood at the level of linguistic theory, still remains to be fully investigated at the level of parsing. In this paper, we tried to scratch the surface of the problem through a careful analysis of the performance of an incremental dependency analyser of Italian, which can run in both a non-lexicalised and a lexicalised mode.

The contribution of lexical information to parse success is unevenly distributed over both part of speech categories and frame types. For reasons abundantly illustrated in section 4, the frames of noun heads are not quite as useful as those of verb heads, especially when available information is only syntactic. Moreover, while information on verb transitivity or clause embedding is crucial to filter out noisy attachments, information on the preposition introducing the oblique complement or the indirect object of a verb can be misleading, and should thus be used for parsing with greater care. The main reason is that failure to register in the lexicon all possible prepositions actually found in real texts may cause undesired over-filtering of genuine arguments (false negatives). In many cases, argument prepositions are actually selected by the lexical head of the subcategorised argument, rather than by its subcategorising verb. Similarly, while information about argument optionality *vs* obligatoriness is seldom confirmed in real language use, statistical preferences on the order of argument realisation can be very useful.

Most current lexicons say very little about temporal and circumstantial modifiers, but much more can be said about them that is useful to parsing. First, some prepositions only occur to introduce verb modifiers. These semantically heavy prepositions, often consisting of more than one lexical item, play a fundamental role in the organization of written texts, and certainly deserve a special place in a parsing-oriented lexicon. Availability of this type of lexical information could pave the way to the development of specialised “mini-parsers” of those satellite modifiers whose structural position in the sentence is subject to considerable variation. These mini-parsers could benefit from information about semantically-based classes of nouns, such as locations, measure terms, or temporal expressions, which should also contain indication of the preposition they are typically introduced by. Clearly, this move requires abandoning the prejudice that lexical information should only flow from the head to its dependents. Finally, availability of large repertoires of multi word units (both complex prepositions and compound terms) appears to have a large impact on improving parse precision.

There is no doubt that harvesting such a wide range of lexical information in the quantity needed for accurate parsing will require extensive recourse to bootstrapping methods of lexical knowledge acquisition from real texts.

References

- Bartolini R., Lenci A., Montemagni S, Pirrelli V. (2002) *The Lexicon-Grammar Balance in Robust Parsing of Italian*, in Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria.
- Briscoe, E.J. (2001) *From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change*, in *Proceedings of Corpus Linguistics 2001*, Lancaster University, pp. 79-89.
- Briscoe T., Carroll J., (2002) *Robust Accurate Statistical Annotation of General Text*, in Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria.
- Carroll, J., Minnen G., Briscoe E.J. (1998) *Can sub-categorisation probabilities help a statistical parser?*, in Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora, Montreal, Canada. 118-126.
- Chanod J.P. (2001) *Robust Parsing and Beyond*, in J.C. Junqua and G. van Noord (eds.) *Robustness in Language and Speech Technology*, Dordrecht, Kluwer, pp. 187-204.
- Federici, S., Montemagni, S., Pirrelli, V. (1998a) *Chunking Italian: Linguistic and Task-oriented Evaluation*, in Proceedings of the LREC Workshop on “Evaluation of Parsing Systems”, Granada, Spain.
- Federici, S., Montemagni, S., Pirrelli, V., Calzolari, N. (1998b) *Analogy-based Extraction of Lexical Knowledge from Corpora: the SPARKLE Experience*, in Proceedings of the 1st International Conference on Language resources and Evaluation, Granada, Spain.
- Grishman, R., Macleod C., Meyers A. (1994) *COMLEX Syntax: Building a Computational Lexicon*, in Proceedings of Coling 1994, Kyoto.
- Jensen K. (1988a) *Issues in Parsing*, in A. Blaser (ed.), *Natural Language at the Computer*, Springer Verlag, Berlin, pp. 65-83.
- Jensen K. (1988b) *Why computational grammarians can be skeptical about existing linguistic theories*, in Proceedings of COLING-88, pp. 448-449.
- Lenci, A., Bartolini, R., Calzolari, N., Cartier, E. (2001) *Document Analysis*, MLIS-5015 MUSI, Deliverable D3.1,.
- Lenci, A., Montemagni, S., Pirrelli, V., Soria, C. (2000) *Where opposites meet. A Syntactic Metascheme for Corpus Annotation and Parsing Evaluation*, in Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.
- Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Paziienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. (2000) *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.
- Procter, P. (1987) *Longman Dictionary of Contemporary English*, Longman, London.
- Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., Zampolli, A. (1998) *The European LE-PAROLE Project: The Italian Syntactic Lexicon*, in Proceedings of the 1st International Conference on Language resources and Evaluation, Granada, Spain, 1998.