Comparing Italian parsers on a common treebank: the Evalita experience

C. Bosco*, A. Mazzei*, V. Lombardo*,
G. Attardi†, A. Corazza◇, A. Lavelli‡, L. Lesmo*, G. Satta•, M. Simi†

* Università di Torino, Italy, {bosco,mazzei,vincenzo,lesmo}@di.unito.it
‡ FBK-irst - Trento, Italy, lavelli@fbk.eu,
◇ Università "Federico II" - Napoli, Italy, corazza@na.infn.it
• Università di Padova, Italy, satta@dei.unipd.it
† Università di Pisa, Italy, {attardi,simi}@unipi.it

Abstract

The Evalita '07 Parsing Task has been the first contest among parsing systems for Italian. It is the first attempt to compare the approaches and the results of the existing parsing systems specific for this language using a common treebank annotated using both a dependency and a constituency-based format.

The development data set for this parsing competition was taken from the Turin University Treebank, which is annotated both in dependency and constituency format. The evaluation metrics were those standardly applied in CoNLL and PARSEVAL. The results of the parsing results are very promising and higher than the state-of-the-art for dependency parsing of Italian. An analysis of such results is provided, which takes into account other experiences in treebank-driven parsing for Italian and for other Romance languages (in particular, the CoNLL X & 2007 shared tasks for dependency parsing). It focuses on the characteristics of data sets, i.e. type of annotation and size, parsing paradigms and approaches applied also to languages other than Italian.

1. Introduction

By providing a very large set of syntactically annotated sentences, the Penn Treebank has played an invaluable role in enabling the development of state-of-the-art parsing systems (Ratnaparki, 1997; Charniak, 1997; Collins, 1999). But the strong focalization on Penn Treebank, and more specifically on the Wall Street Journal portion of this treebank, has left open several questions on parsers' portability. The application of parsing methods to different languages and treebanks is currently considered a crucial and challenging task, and system porting across text genres, languages and annotation formats should be a research problem in itself. The validation of existing parsing models, in fact, strongly depends on the possibility of generalizing their results on corpora other than those on which they have been originally trained and tested.

For constituency-based parsing, strong empirical evidence demonstrates that results obtained on a particular treebank are not portable to other corpora. For instance, Gildea (2001) shows that the results obtained on the Wall Street Journal section of the Penn Treebank are not reproducible on the Brown Corpus, which is annotated according to the same format but contains texts featured by different genre. Other works showed the difficulty of replicating the performance achieved on English when applying statistical parsing to other languages (e.g. (Collins et al., 1999) on Czech, (Dubey and Keller, 2003) on German, (Levy and Manning, 2003) on Chinese, (Corazza et al., 2004) on Italian). While, e.g. (Kübler, 2005; Kübler and Prokič, 2006; Maier, 2006) on Negra and TübaD/Z treebanks show that parsing results vary according to the features of the annotation schema applied to the same corpus of sentences, i.e. dependency or constituency-based.

For dependency parsing, the results of the CoNLL X and CoNLL 2007 multilingual shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007a) together with those reported in (Nivre et al., 2007b; Chanev, 2005), showed that it is as robust as the constituency parsing, but equally affected by the problem of irreproducibility of results across corpora and languages.

The aim of the EVALITA '07 Parsing Task (EPT), whose outcome was presented in Frascati (Rome) in September 2007, was to assess the current state of the art in parsing Italian by encouraging the application of existing parsing models to this language, and to contribute to the investigation on the causes of this irreproducibility (Bosco et al., 2007). It allowed to focus on Italian by exploring both different paradigms, i.e. constituency and dependency, and different approaches, i.e. rule-based and statistical. In fact, the task consisted of subtasks with separate development and evaluation data sets for both constituency and dependency parsing. Indeed, the EPT can be considered as the first outlook of the problems to be faced for parsing Italian and of the kind of work required to adapt existing parsing models to this language.

The paper presents an analysis of the results that goes beyond the limits of the event. The next section presents the development and test data sets, and the evaluation metrics applied in EPT. The following section presents the results obtained by the participating parsing systems. The final section presents an analysis of these results including comparisons with parsing experiences on other languages and in similar contests.

2. Task, data sets and evaluation metrics

The EPT is defined as the task of assigning a syntactic structure to a given Part of Speech (PoS) tagged Italian sentence using a fully automated parser. The syntactic structure has to be expressed according to one of two annotation schemes presented in the development set, one for the dependency and one for the constituency parsing subtask. The annotation schemes, data sets and standard evaluation metrics applied in EPT are described in the rest of this section.

2.1. Development and test data sets

The reference treebank for EPT is the Turin University Treebank (TUT). The full TUT data set (see the TUT web site for a free download: *http://www.di.unito.it/˜tutreeb*) was provided to the EPT participants as development corpus. It currently consists in 2,000 sentences that correspond to about 58,000 annotated tokens. In order to allow for comparison of results across text genres, the treebank is organized in two subcorpora of one thousand sentences each, i.e. the Italian legal Code (47.5% of tokens) and Italian newspapers (52.5% of tokens).

The test set consists instead of 200 new sentences (100 from newspapers and 100 from Italian legal Code), in order to represent a text genre balance similar to that in the development set, and thus to allow for separate evaluations on the different genres.

The TUT collection has been available since several years both in dependency and constituency format. For EPT, in order to make the data more similar to those used in previous parsing contests, such as CoNLL and PARSEVAL, the organizers have generated new formats without non-standard features. The rest of this section describes the details of the development data for the EPT dependency and constituency parsing subtasks.

2.1.1. Development data for dependency parsing subtask

The native annotation schema of TUT is dependency-based (see Figure 1). It follows the major tenets of Hudson's dependency grammar (Hudson, 1984), but includes null elements for the representation of particular phenomena, such as non-projective structures and pro-drops. For instance, in the example of Figure 1, the subject of the verb ’investire’ (to invest) which is not lexically realized because of the equi phenomenon, is annotated as the null element ’t’ on line 11.10, and co-referenced with ’l’Italia’ by using the index ’[6f]’ and the same PoS tags of the sixth word of the sentence.

```
1 Davanti (DAVANTI PREP POLI LOC) [8;PREP-RMOD-LOC+METAPH]
2 all' (A PREP MONO) [1;CONTIN+PREP]
2.1 all' (IL ART DEF F SING) [2;PREP-ARG]
3 emergenza (EMERGENZA NOUN COMMON F SING) [2.1;DET+DEF-ARG]
4 umanitaria (UMANITARIO ADJ QUALIF F SING) [3;ADJC+QUALIF-RMOD]
5 , (#\, PUNCT) [8;SEPARATOR]
6 l' (IL ART DEF F SING) [8;VERB-SUBJ]
7 Italia (ITALIA NOUN PROPER F £STATE) [6;DET+DEF-ARG]
8 decise (DECIDERE VERB MAIN IND REMPAST TRANS 3 SING) [0;TOP-VERB]
9 comunque (COMUNQUE ADV INTERJ) [8;ADVB-RMOD-CONJTEXT]
10 di (DI PREP MONO) [8;VERB-OBJ]
11 investire (INVESTIRE VERB MAIN INFINITE PRES TRANS) [10;PREP-ARG]
11.10 t [6f] (IL ART DEF F SING) [11;VERB-SUBJ]
12 in (IN PREP MONO) [11;PREP-RMOD-LOC+IN]
13 Albania (ALBANIA NOUN PROPER F £STATE) [12;PREP-ARG]
14 . (#\. PUNCT) [8;END]
```

Figure 1: The sentence ’Davanti all’emergenza umanitaria, l’Italia decise di investire in Albania’ (In front of the humanitarian emergency, Italy decided to invest in Albany.) in the native TUT format.

Moreover, the treebank features a rich set of grammatical relations (i.e. around 250 relations) developed according to the Augmented Relational Structure (Bosco and Lombardo, 2004). Each of these relations can, in fact, include three different components, i.e. morpho-syntactic, functional-syntactic and syntactic-semantic. For instance, in the example of Figure 1, in the relation PREP-RMOD-LOC+METAPH, annotated on the first word, i.e. ’Davanti’ (in front of), PREP is the morpho-syntactic component, RMOD the functional-syntactic component, and LOC+METAPH represents the syntactic-semantic component consisting of two features (one indicating the type as a location and the other further specifying the location as metaphorical).

This provides a scalable representation at different degrees of specificity. For instance, by selecting only the functional-syntactic component of each relation, we can reduce the cardinality of the relation set from 250 (fully-specified) to 74 (specified only from the functional-syntactic point of view) items. In Figure 2, you can see the same example of Figure 1 annotated with this reduced relation set; here, e.g., the above mentioned relation PREP-RMOD-LOC+METAPH is reduced to RMOD. For EPT, this annotation with a reduced set of relations, including only the functional-syntactic component, has been considered as the more adequate since its relation cardinality is closer to that of treebanks used in the CoNLL competitions. In fact, in the treebanks used in the CoNLL X Shared Task, the number of dependency relations ranged from 82 (in the Chinese treebank) to 7 (in the Japanese treebank), and in CoNLL 2007 ranged from 69 (in the Chinese treebank) to 20 (in the English treebank), with an average of 39 relations per treebank.

```
1 Davanti (DAVANTI PREP POLI LOC) [8;RMOD]
2 all' (A PREP MONO) [1;CONTIN+PREP]
2.1 all' (IL ART DEF F SING) [2;ARG]
3 emergenza (EMERGENZA NOUN COMMON F SING) [2.1;ARG]
4 umanitaria (UMANITARIO ADJ QUALIF F SING) [3;RMOD]
5 , (#\, PUNCT) [8;SEPARATOR]
6 l' (IL ART DEF F SING) [8;SUBJ]
7 Italia (ITALIA NOUN PROPER F £STATE) [6;ARG]
8 decise (DECIDERE VERB MAIN IND REMPAST TRANS 3 SING) [0;TOP]
9 comunque (COMUNQUE ADV INTERJ) [8;RMOD]
10 di (DI PREP MONO) [8;OBJ]
11 investire (INVESTIRE VERB MAIN INFINITE PRES TRANS) [10;ARG]
11.10 t [6f] (IL ART DEF F SING) [11;SUBJ]
12 in (IN PREP MONO) [11;RMOD]
13 Albania (ALBANIA NOUN PROPER F £STATE) [12;ARG]
14 . (#\. PUNCT) [8;END]
```

Figure 2: The same sentence of Figure 1 in TUT format with reduced relations.

Moreover, in order to further increase the comparability with other works and the adequateness for the application of standard measures for the evaluation of parsing results, a format without null elements has been also produced for EPT. In this format amalgamated words are also annotated slightly differently than in the native TUT[1] (see Figure 3). The training data for the dependency parsing subtask has also been provided in the standard CoNLL format with the information split into ten columns (see Figure 4) that respectively represent the identifier (i.e. position) of the word in the sentence, the word form, the word lemma, the coarse-grained and the fine-grained PoS of the word, morphological features, the head word, the dependency relation, the projective head and the projective dependency relation (the last two are not present in the TUT since TUT adopts null elements to annotate non-projective structures). In conclusion, the development set for EPT dependency subtask has been made available in the formats below:

- in native TUT (Figure 1)
- in TUT with a reduced relation set (Figure 2)
- in a null elements free annotation without sub-indexes (Figure 3)
- in the 10-column standard CoNLL format (with a relation reduced set, without sub-indexes and null elements) (Figure 4).

```
1 Davanti (DAVANTI PREP POLI LOC) [9;PREP-RMOD-LOC+METAPH]
2 all' (A PREP MONO) [1;CONTIN+PREP]
3 all' (IL ART DEF F SING) [2;PREP-ARG]
4 emergenza (EMERGENZA NOUN COMMON F SING) [3;DET+DEF-ARG]
5 umanitaria (UMANITARIO ADJ QUALIF F SING) [4;ADJC+QUALIF-RMOD]
6 , (#\, PUNCT) [9;SEPARATOR]
7 l' (IL ART DEF F SING) [9;VERB-SUBJ]
8 Italia (ITALIA NOUN PROPER F £STATE) [7;DET+DEF-ARG]
9 decise (DECIDERE VERB MAIN IND REMPAST TRANS 3 SING) [0;TOP-VERB]
10 comunque (COMUNQUE ADV INTERJ) [9;ADVB-RMOD-CONJTEXT]
11 di (DI PREP MONO) [9;VERB-OBJ]
12 investire (INVESTIRE VERB MAIN INFINITE PRES TRANS) [11;PREP-ARG]
13 in (IN PREP MONO) [12;PREP-RMOD-LOC+IN]
14 Albania (ALBANIA NOUN PROPER F £STATE) [13;PREP-ARG]
15 . (#\. PUNCT) [9;END]
```

Figure 3: The same sentence of Figure 1 in TUT format free of null elements and sub-indexes.

```
1   Davanti     DAVANTI    PREP  PREP  POLI|LOC   9   RMOD_   _
2   all'   A    PREP  PREP  MONO       1   CONTIN+PREP    _   _
3   all'   IL   ART   ART   DEF|F|SING 2   ARG  _   _
4   emergenza   EMERGENZA NOUN       NOUN     COMMON|F|SING  3   ARG  _   _
5   umanitaria  UMANITARIO      ADJ  ADJ   QUALIF|F|SING    4   RMOD_   _
6   ,      #\,  PUNCT       PUNCT      _    9   SEPARATOR _   _
7   l'     IL   ART   ART   DEF|F|SING 9   SUBJ _   _
8   Italia  ITALIA NOUN     NOUN      PROPER|F|£STATE  7   ARG _   _
9   decise DECIDERE  VERB  VERB  MAIN|IND|REMPAST|TRANS|3|SING   0   TOP  _   _
10  comunque    COMUNQUE ADV  ADV   INTERJ9    RMOD_   _
11  di     DI   PREP  PREP  MONO       9   OBJ  _   _
12  investire   INVESTIRE  VERB  VERB  MAIN|INFINITE|PRES|TRANS   11  ARG  _   _
13  in     IN   PREP  PREP  MONO       12  RMOD_   _
14  AlbaniaALBANIA   NOUN      NOUN     PROPER|F|£STATE  13  ARG  _   _
15  .      #\.  PUNCT       PUNCT      _    9   END  _   _
```

Figure 4: The same sentence of Figure 1 with reduced relations, free of null elements and sub-indexes in CoNLL standard 10columns format.

2.1.2. Development data for constituency parsing subtask

In recent years, by applying automatic procedures to the native annotation, as described in (Bosco and Lombardo, 2006; Bosco, 2007), the TUT dependency treebank has been converted to a Penn-like format called TUT-Penn. The conversion process consists of various steps corresponding to the kinds of information annotated in the dependency TUT, i.e. morphological, structural or relational syntactic. The main step consists in the translation of dependency structures in X-bar-like constituency structures and is based on Xia's algorithm (Xia, 2001). The result, which goes beyond a simple conversion in Penn format, is the generation of a set of a cascade of three parallel treebanks. Such treebanks feature formats which implement constituency structures progressively flatter and with less information about relations, with the Penn format being the result of the final step.

The TUT-Penn format (see Figure 5) includes, as usual for constituency-based annotations, null elements. While the structure is the same as the Penn corpora, TUT-Penn uses a different PoS tag set. In fact, as in other cases of treebank conversion (Collins et al., 1999), the use of a specific set of PoS tags, which are derived by reduction from the TUT original PoS tags, has been preferred to the original Penn PoS tags since they better represent the inflectional richness of Italian. This is the format of the development data for EPT constituency subtask.

```
( (S
    (PP-LOC (PREP Davanti)
            (PP (PREP all')
                (NP (ART~DE all') (NOU~CS emergenza) (ADJ~QU umanitaria))))
    (, ,)
    (NP-SBJ-633 (ART~DE l') (NOU~PR Italia))
    (VP (VMA~RA decise)
        (ADVP (ADVB comunque))
        (PP (PREP di)
            (S
                (NP-SBJ (-NONE- *-633))
                (VP (VMA~IN investire)
                    (PP-LOC (PREP in)
                        (NP (NOU~PR Albania))))))
    (. .)) )
```

Figure 5: The same sentence of Figure 1 in TUT-Penn format.

[1] In native TUT, the almagamated words are annotated using sub-indexes. Compare e.g., in Figure 1 the second and third line to the corresponding lines in Figure 3, where the word 'all' (to the) has been duplicated in order to provide separate annotations about the preposition and the article.

2.2. Evaluation metrics

The evaluation of dependency results is based on the three metrics used in the CoNLL X Shared Task (Buchholz and Marsi, 2006):

- Labeled Attachment Score (LAS), the percentage of tokens with correct head and relation label;
- Unlabeled Attachment Score (UAS), the percentage of tokens with correct head;
- Label Accuracy (LA), the percentage of tokens with correct relation label.

For constituency parsing, the evaluation is based on standard PARSEVAL measures (Black et al., 1991):

- Bracketing Precision (Br-P), the percentage of found brackets which are correct;
- Bracketing Recall (Br-R), the percentage of correct brackets which are found;
- Bracketing F_1 (Br-F), the composition of the previous two measures calculated by the following formula:

$$\frac{2 * (Br\text{–}P * Br\text{–}R)}{(Br\text{–}P + Br\text{–}R)}$$

3. Participants and results

In this section, we describe the systems that participated in EPT and their results.

3.1. Submissions and results

Test runs were submitted to EPT by 8 participants[2], among which 5 are from Italy and the others are from India, Germany, USA, all belong to academic institutes. Six submissions concern dependency parsing and two constituency parsing. Nobody participated to both subtasks. In the tables with results, one for dependency and one for constituency, systems are identified by the institution name and by the last name of the first team member separated by underscore, like (Bosco et al., 2007).

3.1.1. Dependency subtask

The participating systems to the dependency parsing subtask are the following.

The parser UniTo_Lesmo includes chunking followed by attachment of verb dependents driven by both rules manually developed and data about verbal subcategorization. It is a rule-based parser developed in parallel with the TUT and tuned on the data set.

The parser UniPi_Attardi (of the team composed by Attardi and Simi), called DeSR, is a multilingual deterministic shift-reduce dependency parser that handles non-projective dependencies incrementally and learns by means of a second-order multiclass averaged perceptron classifier.

The IIIT_Mannem is an online large margin based training framework for deterministic parsing using Nivre's shift-reduce parsing algorithm.

The UniStuttIMS_Schiehlen uses Eisner's bottom-up chart-parsing algorithm for inference and online passive aggressive algorithms for learning; it produces non-projective labelled trees.

The UPenn_Champollion system (by the team composed by Champollion and Robaldo) is a bidirectional dependency parser which does a greedy search over the sentence and picks the relation between two words with the best score each time and builds the partial tree.

The UniRoma2_Zanzotto, called CHAOS, implements a modular and lexicalised approach based on the notion of eXtended Dependency Graph.

Table 1 describes the details of the results according to the above defined standard measures[3] and show that the best scores for this task were obtained by the UniTo_Lesmo.

LAS	UAS	LA	Participant	Total
86.94	90.90	91.59	UniTo_Lesmo	1-1-1
77.88	88.43	83.00	UniPi_Attardi	2-2-2
75.12	85.81	82.05	IIIT_Mannem	3-4-3
74.85	85.88	81.59	UniStuttIMS_Schiehlen	4-3-4
*	85.46	*	UPenn_Champollion	*-5-*
47.62	62.11	54.90	UniRoma2_Zanzotto	5-6-5

Table 1: Dependency parsing subtask evaluation

3.1.2. Constituency subtask

Two teams participated to the EPT for constituency parsing. The team composed by Corazza, Lavelli, and Satta participated with a parser, i.e. UniNa_Corazza, which is an adaptation to Italian of Collins' probabilistic parser (as implemented by Dan Bikel). It achieved the best result for this task.

The FBKirst_Pianta is instead a left corner parser for Italian, based on explicit rules manually coded in a unification formalism.

The details of their results are described in Table 2.

Br-R	Br-P	Br-F	Errors	Participant
70.81	65.36	67.97	26	UniNa_Corazza
38.92	45.49	41.94	48	FBKirst_Pianta

Table 2: Constituency parsing subtask evaluation. Errors are due the wrong treatment of multiword expressions. As a consequence the number of tokens in the parser output is different from the one in the gold-standard sentence.

4. Analysis and discussion of results

In this section, the results obtained in EPT will be compared with those obtained for Italian by other data-driven parsing systems applied on it. We will present an analysis for each subtask and in the comparison, we will focus on the effects

[2] Among participants, five are single authors, while the others are teams.

[3] The results were computed using the PERL script `evalp07.pl`, provided by the CoNLL 2007 Shared Task organizers: we thank the organization that publicly released this resource.

on parsing results of various parameters, but, in particular, of differences in the data set size and annotation. Therefore, among the scores for Italian, we will take into account those based both on TUT and on another existing treebank for the same language, namely the Italian Syntactic Semantic Treebank (ISST) (Montemagni et al., 2003)[4]. The ISST treebank uses a different annotation schema than TUT, with a syntactic annotation distributed over two levels, the constituent structure and the functional relation level where 22 dependency relations are attested.

Since ISST was used in the multilanguage task of the CoNLL 2007 shared task, we can somehow compare the performance of the three parsers that participated in both the parsing tasks.

The last part of this section focuses, instead, on the parsing approaches applied in EPT.

4.1. Dependency subtask

It is interesting to compare the results in EPT with those for Italian in the CoNLL 2007 multilingual dependency parsing shared task (Nivre et al., 2007a). We will also try a comparative analysis of the effects of annotation styles on parsing accuracy.

Italian was considered among the parsed languages with highest accuracy scores, i.e. achieving LAS between 84.40% and 89.61%, together with Catalan (a Romance language like Italian), Chinese and English. In the CoNLL 2007 shared task, the training corpus for Italian was a portion of ISST, and included a larger amount of sentences than in EPT, namely 3,100 which correspond to around 71,000 annotated tokens rather than 2,000 (i.e. 58,000 tokens).

The best scores for Italian were 84.40% for LAS, obtained by the parser described in Hall et al. (2007), and 87.91% for UAS achieved by the parser described in Nakagawa (2007). These scores are still lower than those obtained in EPT, and, moreover, they were both obtained by systems exploiting a combination of several parsers.

A more fair comparison should therefore refer to the best performing single parser system, i.e. the IDP parser by Titov and Henderson (Titov and Henderson, 2007), which achieved yet lower scores, 82.26% for LAS and 86.26% for UAS.

The differences in the scores achieved in the CoNLL 2007 shared task and in EPT arise both from the use of different parsing models and from the different data sets used.

Analyzing the results from parsers that participated both to CoNLL and to EPT can shed some light on the reasons for these differences. Table 3 shows the results achieved by UniPi_Attardi, IIIT_Mannem and UniStuttIMS_Schiehlen, in both CoNLL and EPT.

All three systems obtained higher scores for LAS in CoNLL than in EPT, i.e. 81.34% versus 77.88% for UniPi_Attardi, 78.67% versus 75.12% for IIIT_Mannem, and 80.46% versus 74.54% for UniStuttIMS_Schiehlen. This can be interpreted as a confirmation of the trivial fact that the performance of parsing systems is influenced by the type of annotation in the reference treebanks.

In particular, the higher number of relations in TUT with respect to ISST (74 versus 22 relations), together with the smaller data set size (42,000 versus 71,000 tokens), may account for the score differences.

On the contrary, all three systems obtained a higher UAS in EPT than in CoNLL. This can be interpreted as a confirmation of the fact that pure dependency annotation schemes, like that of TUT, appear more adequate for representing the structure of the Italian language than the ISST annotation, as suggested previously in Chanev (2005) from experiments on TUT and ISST.

An evidence for the key role of the pure dependency annotation derives from the experiments described in Attardi and Simi (2007) where the same parser DeSR applied in the EPT, i.e. UniPi_Attardi, is used in the same EPT task but exploiting a smaller set of (less specialized) TUT relations (31 versus 74): a better LAS was achieved (83.27%) than the official score (77.88%), which is also higher than that obtained by the same parser in CoNLL on a larger size of data set of the ISST.

We tested Maltparser, which had been previously used on TUT by Nivre et al. (2007b), using the same settings of the official CoNLL 2007 run for Italian, and achieved a LAS of 74.63% and a UAS of 88.90%. We also applied IDP, kindly made available to us by Ian Titov, to the EPT task and obtained the following scores: LAS 76.79%, UAS: 88.13%. Recent experiments using DeSR, configured to use SVM as learning algorithm, achieved even slightly better scores: LAS 77.95%, UAS of 88.50%. These can hence be considered as the best scores that a single statistically trained parser can currently achieve on the EPT dependency parsing task. One must remark though that the accuracy of statistical parsers increases with the size of the training corpus and the TUT corpus is quite smaller than corpora like the English Penn WSJ Treebank (1 million tokens) or the Catalan CESS Treebank (450 thousand tokens), on which statistical parsers can achieve accuracy scores above 90.0%.

The scores on EPT of statistical parsers are still significantly lower than those achieved by the UniTo_Lesmo parser (LAS: 86.94%, UAS: 90.90%), which is a rule-based parser, whose rules were specifically tuned to TUT.

Finally, it can be interesting to compare the accuracy of dependency parsers on Italian with respect to other Romance languages. In particular, Catalan has been included in the CoNLL 2007 shared parsing task while Portuguese and Spanish were used instead in the CoNLL X shared parsing task. All these three languages, as well as Italian, obtained good performance of the parsers employed in both the competitions. At CoNLL X, Portuguese obtained the second best score[5]; At CoNLL 2007, Italian and Catalan were in the "High Top Score Group" together with English and Chinese (Nivre et al., 2007a). Spanish is the only Romance language that obtained relatively bad performance in the CoNLL X shared task: possible speculations about this data could be motivated may be on the basis of the high

[4] It does exist a further treebank for Italian, i.e. the Venice Italian Treebank (Delmonte, 2008 to appear), but, to our knowledge, there are currently no published results for parsing experiments based on this treebank.

[5] The first by excluding the "easy" task of Japanese parsing (Buchholz and Marsi, 2006)

LAS		UAS		Participant
CoNLL	EPT	CoNLL	EPT	
81.34	77.88	85.54	88.43	UniPi_Attardi
78.67	75.12	82.91	85.81	IIIT_Mannem
80.46	74.85	84.54	85.88	UniStuttIMS_Schiehlen

Table 3: Comparison between EPT and CoNLL-07 LAS and UAS, ordered according to the EPT LAS scores.

average sentence length in the Spanish treebank.

The results obtained for dependency parsing at the EPT can be therefore considered as satisfactory, since they are higher or very close to the state of the art. Moreover, they offered a valuable experimental evidence to previously formulated hypothesis, namely the adequateness of dependency parsing approaches to Italian. This is in line with similar hypotheses formulated for other languages that exhibit free word order.

4.2. Constituency subtask

For the constituency parsing subtask, the results of EPT are less meaningful than those for dependency parsing, because only two systems participated to this subtask and because there is limited empirical evaluation on this kind of parsing applied to Italian language. The comparison will therefore mainly refer to English, which remains the reference language for constituency-based parsing approaches.

In Corazza et al. (2004) the same parser used in EPT (i.e. UniNa_Corazza) was run on an Italian data set composed by about 3,000 sentences from the constituency-based portion of ISST (see the beginning of Section 4.), achieving scores definitely lower than those obtained using as training set a subset of the WSJ of comparable size (see the fourth line of Table 4). The worse results on Italian with respect to English are confirmed in EPT, and the smaller data set in EPT results in lower scores than on ISST (Corazza et al., 2007). However, further experiments after normalizing multiword expressions (cause of the errors reported in Table 2) produced better results, also higher than on ISST (see the second line in Table 4).

When comparing the results obtained on ISST and TUT it is important to underline that the experiments on ISST followed a 10-fold cross-validation protocol, and therefore at each step the training set was larger than for TUT (about 2,700 sentences versus less than 2,000), resulting in a slight bias in favor of the ISST treebank. Despite such bias, the results on TUT (after fixing the misalignment problems) are better than on ISST.

These differences call for some future work on the investigation of the structural differences between the two treebanks.

We expect to obtain better results on the new release of the TUT-Penn currently under development.

4.3. Parsing approaches

The parsing approaches used by the EPT participants included both statistical (5 participants, 4 for dependency and 1 for constituency) and rule-based (3 participants, 2 for dependency and 1 for constituency) parsing.

	Br-R	Br-P	Br-F
EPT official (Italian)	70.81	65.36	67.97
post-EPT (Italian)	71.73	69.88	70.79
ISST (Italian)	68.40	68.58	68.49
WSJ (English)	84.02	83.41	83.71

Table 4: Comparison of performance of UniNa_Corazza on different treebanks. The results on English were obtained using as training set a subset of the WSJ of comparable size (sections 02 & 03).

The rule-based parsers are UniTo_Lesmo and UniRoma2_Zanzotto for dependency parsing and FBKirst_Pianta for constituency parsing.

The two most accurate statistical dependency parsers (UniPi_Attardi and IIIT_Mannem) use variants of a shift-reduce parsing algorithm. A machine learning classifier is trained on the corpus to predict the proper parsing action. In the submitted runs, both systems used variants of the Perceptron Algorithm: UniPi_Attardi used a second order perceptron while IIIT_Mannem used MIRA. An unofficial run of UniPi_Attardi using an SVM classifier achieved a further slight improvement, showing that the choice of learning algorithm can be critical. In UniPi_Attardi, non-projectivity is handled by specific reduce actions: however the number of non-projective relation is quite small in the TUT corpus to have significant impact on the overall accuracy: there are only 19 non-projective occurrences among the over 5000 tokens in the test set.

The parser UniStuttIMS_Schiehlen follows the approach by McDonald in (McDonald et al., 2005), selecting the Maximum Spanning Tree with the highest score, among all possible projective dependency trees. The weights for the scoring function are learned through iterations on the training data. The parser uses the same weights but different features also to assign dependency labels to relations. For EPT the second order features of McDonald algorithm were switched off.

In the dependency subtask, statistical parsers have achieved notable results although the development set is smaller than in CoNLL 2007. As mentioned earlier, the LAS accuracy achieved by the best parser is superior to those of Maltparser and IDP, two of the best single parsers participating in the CoNLL 2007 shared task.

Among the rule-based dependency parsers, differences in tuning may explain the large difference in accuracy between the parser UniTo_Lesmo, which achieved the best overall accuracy for this task, and the UniRoma2_Zanzotto. The statistical constituency parser (UniNA_Corazza) is an adaptation of the Collins' parser (Collins, 1999), as implemented by Dan Bikel[6] (Bikel, 2004). Adaptation of Collins' parser to the TUT included the identification of rules for finding lexical heads, and the selection of a lower threshold for unknown words (as the amount of available data is much lower than for WSJ). No language-dependent adaptations (such as the tree transformations introduced by Collins for the PennTreeBank) were introduced.

The rule-based constituency parser (FBKirst_Pianta) is a left corner parser based on explicit rules manually coded in a unification formalism. The grammar is inspired to the Lexical Functional Grammar linguistic theory and encodes various kinds of linguistic information in parallel: constituency, grammatical functions and semantics. As the linguistic coverage of the grammar is still quite limited, the parser produces complete parse trees for a small percentage of sentences. A number of strategies to recover from parsing failures were applied and evaluated.

5. Conclusions and future work

The paper describes the EPT, the first contest among parsing systems for Italian. The availability of a dependedency and a constituency-based annotation for the same Italian corpus in this contest has allowed for a comparison between parsing performances on different representation paradigms. The paper contrasted these results with previous parsing experiences referring in particular to those for Italian, for free word order and Romance languages.

By showing a higher distance from the state of the art for constituency than for dependency parsing for Italian, the results of the EPT confirm the hypothesis, known in literature, that dependency structures are more adequate for the representation of this language. Under this respect, even if different standard measures and different number of participants to the dependency and constituency subtasks make a direct comparison difficult, EPT contributed to the investigation on parsing system portability.

Furthermore, an important consequence of the Evalita activity has been to strengthen the interactions among groups working on Italian parsing and treebanks. As an immediate effect, thanks to the cooperation between organizers and participants, EPT resulted in an increased quality for the reference treebank, i.e. TUT, which has been newly released in December 2007 in the dependency format, and will be soon newly released in the constituency format too. Hopefully, the EPT will lead to a common future effort towards an in-depth comparison of annotation schemes and to the development of larger integrated resources in order to give more adequate answers to the questions left open by Evalita about the features of treebank annotations, such as the cardinality of the relation sets and the granularity of PoS tagging implemented by different corpora.

Acknowledgments

A. Lavelli has been partially supported by the X-Media project (http://www.x-media-project.org) sponsored by the European Commission (IST-FP6-026978). G. Satta has been partially supported by MIUR under project PRIN No. 2003091149_005.

[6] *http://www.cis.upenn.edu/~dbikel/#stat-parser*

6. References

G. Attardi and M. Simi. 2007. DeSR at the evalita dependency parsing task. *Intelligenza artificiale*, IV(2).

Daniel M. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4).

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English. In *Proceedings of the Speech and Natural Language Workshop*, Pacific Grove, CA.

C. Bosco and V. Lombardo. 2004. Dependency and relational structure in treebank annotation. In *Proceedings of the COLING'04 workshop on Recent Advances in Dependency Grammar*, Geneve, Switzerland.

C. Bosco and V. Lombardo. 2006. Comparing linguistic information in treebank annotations. In *Proceedings of LREC '06.*

C. Bosco, A. Mazzei, and V. Lombardo. 2007. Evalita parsing task: an analysis of the first parsing system contest for Italian. *Intelligenza artificiale*, IV(2).

C. Bosco. 2007. Multiple-step treebank conversion: from dependency to Penn format. In *Proceedings of the ACL'07 workshop on Linguistic Annotation (LAW).*

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL-X.*

A. Chanev. 2005. Portability of dependency parsing algorithms. An application for Italian. In *Proceedings of TLT-2005.*

E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park.

M. Collins, J. Hajic, L. Ramshaw, and C. Tillmann. 1999. A statistical parser of Czech. In *Proceedings of ACL'99.*

M. Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.

A. Corazza, A. Lavelli, G. Satta, and R. Zanoli. 2004. Analyzing an Italian treebank with state-of-the-art statistical parser. In *Proceedings of TLT-2004*.

A. Corazza, A. Lavelli, and G. Satta. 2007. Phrase-based statistical parsing. *Intelligenza artificiale*, IV(2).

R. Delmonte. 2008 - to appear. *Strutture sintattiche dall'analisi computazionale di corpora di italiano.* Franco Angeli, Milano.

A. Dubey and F. Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of ACL'03*.

D. Gildea. 2001. Corpus variation and parser performance. In *Proceedings of EMNLP'01.*

J. Hall, J. Nilsson, J. Nivre, G. Eryigit, B. Megyesi, M. Nilsson, and M. Saers. 2007. Single malt or blended? a study in multilingual parser optimization. In *Proceedings of CoNLL-2007 Shared Task. EMNLP-CoNLL.*

R. Hudson. 1984. *Word Grammar*. Basil Blackwell, Oxford and New York.

S. Kübler and J. Prokič. 2006. Why is German dependency parsing more reliable than constituent parsing? In *Proceedings of TLT-2006*.

S. Kübler. 2005. How do treebank annotation schemes influence parsing results? or how not to compare apples and oranges. In *Proceedings of RANLP*.

R. Levy and C. Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of ACL'03*.

W. Maier. 2006. Annotation schemes and their influence on parsing results. In *Proceedings of ACL'06 Student Research Workshop*.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé, editor, *Building and Using syntactically annotated corpora*. Kluwer, Dordrecht.

T. Nakagawa. 2007. Multilingual dependency parsing using Gibbs sampling. In *Proceedings of CoNLL- 2007 Shared Task*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007a. The CoNLL-2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*, Prague.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007b. MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2).

A. Ratnaparki. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of EMNLP-97*.

I. Titov and J. Henderson. 2007. Fast and robust multilingual dependency parsing with a generative latent variable model. In *Proceedings of CoNLL-2007 Shared Task. EMNLP-CoNLL.*

F. Xia. 2001. *Automatic grammar generation from two different perspectives*. Ph.D. thesis, University of Pennsylvania.