# Evalita'09 Parsing Task: comparing dependency parsers and treebanks

Cristina Bosco*, Simonetta Montemagni %, Alessandro Mazzei*, Vincenzo Lombardo*, Felice Dell'Orletta%, and Alessandro Lenci+%

* Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149, Torino
% Istituto di Linguistica Computazionale - CNR, Via Moruzzi 1, 56124, Pisa
+ Dipartimento di Linguistica, Università di Pisa, Via Santa Maria 36, 56126, Pisa
{bosco,mazzei,vincenzo}@di.unito.it
{simonetta.montemagni,felice.dellorletta,alessandro.lenci}@ilc.cnr.it

**Abstract.** The aim of Evalita Parsing Task is at defining and extending Italian state of the art parsing by encouraging the application of existing models and approaches. As in the Evalita'07, the Task is organized around two tracks, i.e. Dependency Parsing and Constituency Parsing. As a main novelty with respect to the previous edition, the Dependency Parsing track has been articulated into two subtasks, differing at the level of the used treebanks, thus creating the prerequisites for assessing the impact of different annotation schemes on the parsers performance. In this paper, we describe the Dependency Parsing track by presenting the data sets for development and testing, reporting the test results and providing a first comparative analysis of these results, also with respect to state of the art parsing technologies.

**Keywords:** treebanks, parsers, dependency parsing, evaluation.

## 1 Introduction: motivations of the Parsing Task

The general aim of the Evalita Parsing evaluation campaign is at defining and extending Italian state of the art parsing with reference to existing resources, by encouraging the application of existing models to this language.

In the first edition, in 2007 ([8], [9], [23]), the focus was mainly on the application to the Italian language of various parsing approaches, i.e. rule–based and statistical, and paradigms, i.e. constituency– and dependency–based. The aim was in fact at contributing, with reference to Italian, to the investigation on the causes of the irreproducibility of the parsing results known in literature on languages other than English, attested e.g. by [13] for Czech, [17] for German, [22] for Chinese, [14] for Italian, and on treebanks others than Penn [19]. The same development data, extracted from the Turin University Treebank[1], have been therefore distributed both in dependency (TUT native) and constituency

[1] http://www.di.unito.it/∼tutreeb

format (TUT–Penn), and the Task was articulated in two parallel tracks, respectively based on dependency and constituency paradigm. The results in 2007 for dependency parsing have been evaluated as no far from the state of the art for English, while those for constituency showed a higher distance from it, confirming the hypothesis, acknowledged in the literature, that dependency–based approaches appear to be more adequate for free word order languages like Italian.

The current Evalita edition follows the same approach, i.e. the task is organized around two tracks, i.e. Dependency Parsing and Constituency Parsing. While for constituency parsing the track was organized along the same lines as in 2007 (based on an improved and enlarged release of TUT–Penn), the dependency track has been further articulated into two subtasks differing at the level of the used treebanks, namely:

- a main subtask based on TUT (see section 3.1)
- a pilot subtask based on ISST–TANL[2] treebank (see section 3.2).

This novelty creates the prerequisites for contributing as far as Italian is concerned to a recent line of research focussing on the question of whether and how parsers trained on different syntactic resources differ in their performance (see, among others, [28], [11], [21]). We believe that a comparison of parsing results obtained with respect to treebanks for the same language but differing in size, corpus composition and annotation schemes can help to assess the impact of different training resources, and in particular of different annotation strategies, at the parsing level.

The focus of this paper is on the Dependency track of the Evalita'09 Parsing Task (for the constituency track see [10]). The paper is organized as follows. In the following section, we describe the dependency parsing track and its subtasks. Then, in the third and fourth sections the development data sets and the evaluation measures are illustrated. The last sections are devoted to the presentation of participant's results and to a first comparative analysis across the different subtasks but also with respect to state of the art parsing technologies.

## 2 Definition of the dependency parsing task

As described in the CoNLL competitions ([12], [27]), the parsing task is defined as the activity of assigning a syntactic structure to a given set of PoS tagged sentences (called *test set*), using a fully automatic parser and according to the annotation scheme presented in a large set of sentences (called *training set*[3]). The evaluation for this task is based on a manually annotated or simply revised version of the test set (called *gold standard test set*).

[2] Note that in the papers by the participants to the EVALITA 2009 Parsing Task this resource is often referred to as ISST or ISST–CoNLL or CoNLL–ISST.

[3] In the training of statistical parsing systems, usually the training set is split in two parts, one used for training and the other, referred as *development set*, used for testing during the development. This organisation has been followed in the case of PDS (see section 3.2).

These definitions fully apply to Evalita'09 Dependency Parsing track, both to the Main Dependency Subtask (henceforth, MDS), which uses as training set TUT, and to the Pilot Dependency Subtask (henceforth, PDS), which uses as training set the ISST–TANL resource.

In order to allow for a meaningful and direct comparison between the results achieved in the two subtasks, the test set for MDS and that for PDS have been built by including a common subset of 100 sentences (henceforth referred to as *shared test set*). Even if only MDS was obligatory for all participants, the organizers encouraged the participation to both dependency subtasks; five of the six participants have submitted runs for both MDS and PDS, thereby providing an interesting test bed for investigating the influence of annotation schemes on the trained parsers.

## 3 Data sets: TUT and ISST–TANL

As motivated in the previous sections, the data for the two dependency subtasks originate from two treebanks developed for the Italian language[4], namely TUT and ISST–TANL, which differ significantly at the level of both corpus composition and adopted dependency representations. In this section, the TUT and ISST–TANL treebanks are illustrated with particular emphasis on the main differences between the underlying annotation schemes.

### 3.1 TUT: the Main Dependency Subtask training and test sets

The data proposed in the MDS for the training of parsing systems are from TUT, the treebank for Italian developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin [5]. TUT has been newly released in 2009, after automatic and manual revisions, in an improved version where the annotation is more correct and consistent with respect to the version released for Evalita'07.

Even if smaller than other existing Italian resources, i.e. VIT and ISST–TANL, TUT makes available more annotation formats ([6], [7]) that allowed for a larger variety of training and testing for parsing systems and for meaningful comparisons with theoretical linguistic frameworks. For instance, TUT has been used in Evalita contests in 2007 and 2009 as reference treebank both for dependency and constituency parsing (converted in TUT-Penn format, an application of the Penn Treebank format to the Italian language, see [10]). A more recent project involving TUT concerns instead the development of the CCG-TUT, a treebank of Combinatory Categorial Grammar derivations for Italian [3]. Observe that the current release of TUT benefits of feedbacks about correctness derived both from the conversion processes where TUT is involved and from Evalita'07 contest.

[4] For this language, another treebank has been developed, i.e. the Venice Italian Treebank (VIT) [16].

[5] For the free download of the resource, see `http://www.di.unito.it/~tutreeb`.

The annotation scheme of TUT is centered upon the notion of argument structure and applies the major principles of dependency grammar [20] using a rich set of grammatical relations [6]. Moreover, it includes null elements to deal with non-projective structures, long distance dependencies, equi phenomena, pro drop and elliptical structures, which are quite common in a flexible word order language like Italian. On the one hand, this allows in the most of cases for the representation and the recovery of argument structures associated with verbs and nouns. On the other hand, by using null elements crossing edges and non–projective dependency trees can be avoided.

Nevertheless, in order to increase the comparability with other existing resources and to make possible the application of evaluation measures, the native format of TUT has been automatically converted for the Evalita contests in a format more proximate to the standard. This format differs from the native TUT first because it splits the annotation in the ten standard columns (filling eigth of them), as in CoNLL (and therefore is called CoNLL format), rather than organize them in round and square brackets. Second, because it exploits only part of the rich set of grammatical relations (72 in CoNLL versus 323 in TUT), does not includes pointed indexes[7] and null elements. Nevertheless, TUT in CoNLL format does not include non projective structures.

The training set of the MDS includes 2,400 sentences that correspond to 72,149 annotated tokens in TUT native format, and 66,055 tokens in CoNLL format. The corpus can be separated in three subcorpora, i.e. one from Italian newspapers (1,100 sentences and 30,561 tokens in CoNLL format), one from the Italian Civil Law Code (1,100 sentences and 28,048 tokens), and one from the Italian section of the JRC-Acquis Multilingual Parallel Corpus, a collection of declarations of the European Community[8] (200 sentences and 7,446 tokens). This last small corpus has been recently included in TUT for a collaboration between Evalita Parsing Task and the evaluation campaign for parsing French, Passage[9] that exploits texts from the corresponding French section of the same multilingual corpus.

The test set of the MDS includes 240 sentences (5,287 tokens) and features a balancement alike to that of the training set: 100 sentences (1,782 tokens) from newspapers, 100 sentences (2,293 tokens) from Civil Law Code and 40 sentences (1,212 tokens) from the Passage/JRC-Acquis corpus. The above mentioned shared test set is composed by the 100 sentences of the MDS test set extracted from newspapers; they have been in fact extracted from ISST–TANL test set and newly annotated in TUT for Evalita'09.

[6] See [4], [5].

[7] In TUT native format the representation of amalgamated words uses pointed indexes, e.g. a definite prepositions 'del' occurring as 33th word of a sentence is split in two lines, '33 del (PREP ....' and '33.1 del (ART ....' respectively representing the Preposition and the Article. In CoNLL format, where pointed indexes are not allowed, these two lines became '33 del (PREP ....' and '34 del (ART ....'.

[8] `http://langtech.jrc.it/JRC-Acquis.html`

[9] `http://atoll.inria.fr/passage/index.en.html`

### 3.2 ISST–TANL: the Pilot Dependency Subtask training and test sets

PDS is based on the ISST–TANL dependency annotated corpus, which was jointly developed by the Istituto di Linguistica Computazionale (ILC-CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project[10]. The ISST–TANL dependency annotated corpus originates as a revision of the ISST–CoNLL corpus [25] which was used in the CoNLL 2007 Shared Task on Dependency Parsing [27], and which was built in its turn starting from ISST[11], a multi-layered corpus annotated at the orthographic, morpho–syntactic, syntactic[12] and lexico–semantic levels [24].

The ISST–CoNLL corpus is a subset of the balanced ISST partition of about 80,000 tokens (for a total of 4,162 sentences) exemplifying general language usage and consisting of a selection of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). This ISST subset was semi–automatically converted into the CoNLL format and made available for the CoNLL 2007 evaluation campaign. In particular, ISST–CoNLL was built by combining information from the morpho–syntactic and syntactic dependency annotation levels of ISST through a semi–automatic conversion process in charge of a) combining information coming from different annotation levels, and – most importantly – b) converting the ISST dependency annotation scheme into the CoNLL 2007 tabular format. Concerning b), conversion had to cope with the fact that in ISST dependency relations were expressed in terms of binary relations holding between two lexical heads belonging to major lexical classes only (i.e. non-auxiliary verbs, nouns, adjectives and adverbs): in fact, in ISST information about grammatical words (e.g. determiners, prepositions, auxiliaries) was encoded in terms of features associated with the participants to the relation. This implies that during the conversion process the dependency relations involving grammatical words had to be reconstructed from the ISST original annotation and the already existing dependency relations had to be revised accordingly. Other conversion issues which

[10] The TANL project (2007–2009), whose extended title is "Analisi di Testi per il Semantic Web e il Question Answering" (Text Analysis for the Semantic Web and Question Answering), is a project funded by Fondazione Cassa di Risparmio di Pisa coordinated by Giuseppe Attardi and involving the Informatics and Linguistics departments of Pisa University, and the Istituto di Linguistica Computazionale of CNR. It aims at developing linguistic technologies for the Italian language in order to build a Question Answering system based on semantic information: more details at `http://medialab.di.unipi.it/wiki/SemaWiki`.

[11] ISST is a multi–layered annotated corpus of Italian which represents one of the main outcomes of an Italian national project, SI–TAL, funded by the Italian Ministry of Science and Research and coordinated by Antonio Zampolli for the design and development of an integrated suite of tools and resources for Italian Natural Language Processing. ISST was developed between 1999 and 2001.

[12] In ISST syntactic annotation is distributed over two different levels, the constituent structure level and the dependency annotation level. In this context we focussed on the dependency annotation layer only.

had to be addressed are concerned with: multi-headed tokens, which caused the dependency structure not to be a tree; empty tokens, representing omitted subjects due to the pro–drop property of Italian; identification of the sentence root; insertion of dependencies involving punctuation (for more details see [25]).

The ISST–TANL dependency annotated corpus is a revised version of the ISST–CoNLL corpus, where revisions – all performed manually – were mainly concerned with a revised dependency Tag Set and annotation criteria.

As far as PDS is concerned, the evaluation has been based on three data sets:

1. Training Corpus, containing data annotated using the TANL tagset to be used for development and training of the pilot subtask participating systems (2,868, sentences for a total of 66,528 tokens);
2. Development Corpus, a smaller corpus to be used for development (241 sentences corresponding to 4,745 tokens);
3. Test Set, containing blind test data for the evaluation (260 sentences and 5,011 tokens).

Note that the PDS Test Set includes 100 sentences (extracted from newspapers) which are shared with the MDS Test Set, i.e. the above mentioned (see section 3.1) *shared test set*.

### 3.3 TUT vs ISST–TANL resources

A comparison of the TUT and ISST–TANL dependency annotated corpora should take into account both the corpus composition and the adopted annotation schemes. In fact, in principle both factors can influence the final parsing performance.

As described in sections 3.1 and 3.2, the composition of the TUT and ISST–TANL corpora differs significantly. Whereas the TUT corpus is articulated into different sections (namely, Newspapers, Italian Civil Code, declarations of the European Community) which were selected as representative of different types of language use, the ISST–TANL corpus was fully extracted from the "balanced partition" of the ISST corpus containing a selection of newspapers and periodicals articles testifying general language usage. As an evidence about the difference among text genres, we can observe e.g. the average sentence lenght which varies in the MDS test set from 17.82 (for the shared section), to 22.93 (for Civile Code), to 30.30 (for the declarations of the European Community).

For what concerns annotation, although both schemes belong to the dependency paradigm, they show significant differences which may be of some help in explaining different performance results achieved by the same systems in MDS and PDS. In order to give the reader the flavour of how and to what extent the two annotations differ, in tables 1 and 2 respectively we report the TUT (in CoNLL format) and ISST–TANL annotations for the same sentence extracted from the shared test: *La coppia, residente a Milano anche se di origini siciliane, stava trascorrendo un periodo di vacanza*, lit. 'The couple, living in Milan even if of origins sicilian, was having a period of holiday', 'The couple, living in Milan although of Sicilian origin, was having a period of holiday'.

**Table 1.** TUT annotation in CoNLL format of sentence from the shared test set.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | La | IL | ART | ART | DEF—F—SING | 14 | SUBJ | _ _ |
| 2 | coppia | COPPIA | NOUN | NOUN | COMMON—F—SING | 1 | ARG | _ _ |
| 3 | , | # | PUNCT | PUNCT | _ | 2 | OPEN+ PARENTHETICAL | _ _ |
| 4 | residente | RISIEDERE | VERB | VERB | MAIN—PARTICIPLE—PAST—INTRANS—SING—ALLVAL | 2 | RMOD+ RELCL+REDUC | _ _ |
| 5 | a | A | PREP | PREP | MONO | 4 | INDCOMPL | _ _ |
| 6 | Milano | MILANO | NOUN | NOUN | PROPER—F—SING—CITY | 5 | ARG | _ _ |
| 7 | anche | ANCHE | ADV | ADV | CONCESS | 8 | RMOD | _ _ |
| 8 | se | SE | CONJ | CONJ | SUBORD—COND | 4 | RMOD | _ _ |
| 9 | di | DI | PREP | PREP | MONO | 8 | ARG | _ _ |
| 10 | origini | ORIGINE | NOUN | NOUN | COMMON—F—PL | 9 | ARG | _ _ |
| 11 | siciliane | SICILIANO | ADJ | ADJ | QUALIF—F—PL | 10 | RMOD | _ _ |
| 12 | , | # | PUNCT | PUNCT | _ | 2 | CLOSE+ PARENTHETICAL | _ _ |
| 13 | stava | STARE | VERB | VERB | AUX—IND—IMPERF—INTRANS—3—SING | 14 | AUX+ PROGRESSIVE | _ _ |
| 14 | trascorrendo | TRASCORRERE | VERB | VERB | MAIN—GERUND—PRES—TRANS—SING | 0 | TOP | _ _ |
| 15 | un | UN | ART | ART | INDEF—M—SING | 14 | OBJ | _ _ |
| 16 | periodo | PERIODO | NOUN | NOUN | COMMON—M—SING | 15 | ARG | _ _ |
| 17 | di | DI | PREP | PREP | MONO | 16 | RMOD | _ _ |
| 18 | vacanza | VACANZA | NOUN | NOUN | COMMON—F—SING | 17 | ARG | _ _ |
| 19 | . | #˙ | PUNCT | PUNCT | _ | | | |
| 14 | END | _ | _ | | | | | |

**Table 2.** ISST–TANL annotation

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | La | lo | R | RD | num=s—gen=f | 2 | det | _ _ |
| 2 | coppia | coppia | S | S | num=s—gen=f | 13 | subj | _ _ |
| 3 | , | , | F | FF | _ | 4 | punc | _ _ |
| 4 | residente | residente | A | A | num=s—gen=n | 2 | mod | _ _ |
| 5 | a | a | E | E | _ | 4 | comp_loc | _ _ |
| 6 | Milano | milano | S | SP | _ | 5 | prep | _ _ |
| 7 | anche_se | anche_se | C | CS | _ | 4 | con | _ _ |
| 8 | di | di | E | E | _ | 4 | conj | _ _ |
| 9 | origini | origine | S | S | num=p—gen=f | 8 | prep | _ _ |
| 10 | siciliane | siciliano | A | A | num=p—gen=f | 9 | mod | _ _ |
| 11 | , | , | F | FF | _ | 4 | punc | _ _ |
| 12 | stava | stare | V | VA | num=s—per=3—mod=i—ten=i | 13 | modal | _ _ |
| 13 | trascorrendo | trascorrere | V | V | mod=g | 0 | ROOT | _ _ |
| 14 | un | un | R | RI | num=s—gen=m | 15 | det | _ _ |
| 15 | periodo | periodo | S | S | num=s—gen=m | 13 | obj | _ _ |
| 16 | di | di | E | E | _ | 15 | comp | _ _ |
| 17 | vacanza | vacanza | S | S | num=s—gen=f | 16 | prep | _ _ |
| 18 | . | . | F | FS | _ | 13 | punc | _ _ |

By comparing tables 1 and 2, it can be noticed that differences lie at the level of both morpho–syntactic tagging and dependency annotation. If we focus on dependency annotation, we can observe that a first dimension of variation is concerned with the inventory of assumed dependency types. Consider as an example the relation holding between the words *coppia* 'couple' and *residente* 'living': in TUT *residente* is interpreted as the head of a relative clause whereas in ISST–TANL it is treated as a modifier. However, even when – at first glance – the two schemes show common dependency types, they can diverge at the level of their associated meaning. This is the case, for instance, of the "obj" relation which in the TUT annotation scheme refers to the direct argument (either in the nominal or clausal form) occurring at least and most once and expressing the subcategorized object, and in ISST–TANL is meant to denote the relation holding between a verbal head and its non–clausal direct object (other dependency types are foreseen to mark clausal complements).

Other important dimensions of variation are concerned with other aspects. With respect to head selection, following the Word Grammar framework [20] TUT always assigns heads on the basis of syntactic criteria, i.e. in all constructions involving one function word and one content word (e.g. determiner–noun, preposition–noun, complementizer–verb) the head role is always played by the function word. By contrast, in ISST–TANL head selection follows from a combination of syntactic and semantic criteria: i.e. whereas in the determiner–noun and auxiliary–verb constructions the head role is assigned to the semantic head (the noun and the verb respectively), in preposition–noun and complementizer–verb constructions the head role is played by the element which is subcategorized by the governing head, namely the preposition and the complementizer. In the annotation example in tables 1 and 2 such a difference emerges clearly. With respect to projectivity constraint: whereas in TUT the projectivity constraint is assumed[13], ISST–TANL corpus recognizes the need for non–projective representations due to the free word order property of the Italian language. Other important differences between TUT and ISST–TANL are concerned with the treatment of coordination and punctuation, i.e. phenomena which are particularly problematic to deal with in the dependency framework. For instance, in the example TUT recognizes a parenthetical structure between the two occurring commas and marks it with specific dependency types; ISST–TANL follows a different strategy to deal with it, i.e. the two paired commas are both connected to the head of the delimited phrase. Last but not least, distinct tokenization and sentence splitting criteria are assumed in the two resources with repercussions at different levels; e.g. whereas TUT annotated sentences conform to the single root constraint, in ISST–TANL there may be multiple–rooted sentences.

## 4 Evaluation measures

The standard methodology for the evaluation of dependency parsers is to apply them to a test set and compare their output to the gold standard test set, i.e. the test set annotated according to the treebank used for the development of the parsers. Among the most widely used evaluation metrics, we have selected for the evaluation of dependency parsing official results in MDS and PDS those used in the CoNLL parsing shared task, i.e. LAS (Labeled Attachment Score) that is the percentage of tokens with correct head and dependency type. Morever, in accord with literature, we report too the UAS (Unlabeled Attachment Score) measure, i.e. the percentage of tokens with correct head [12, 27]. Note that the use of a single accuracy metric is possible in dependency parsing thanks to the single-head property of dependency trees. This hypothesis unify the measures of precision and recall and makes parsing resemble a tagging task, where every word is to be tagged with its correct head and dependency type [29].

[13] In TUT native format, for the annotation of non-projective structures the annotation of null elements allows for the recovery of corresponding projective structures. Since CoNLL does not admit the use of null elements, they are deleted in TUT in CoNLL format, but the projectivity is maintained.

The evaluation for the MDS and for the PDS (as well as the development of data for the contest) have been separately performed by the groups responsible of the two subtasks, namely the Natural Language group of the Department of Computer Science of Turin University for the former and ILC–CNR and Pisa University for the latter.

With respect to the definition of the task, in order to account for the large variety of parsing systems and the evaluation of all the submitted results, we have considered acceptable a number of discrepancies between the gold standard test set for MDS and PDS (annotated according to TUT or ISST–TANL, as described below) and the participant output. Among these discrepancies we mention, in particular, the application of partly different PoS tagsets with respect to that annotated in the MDS and PDS test sets distributed to the participants and to the gold standard test sets.

## 5 Results

The participants[14] to the Evalita'09 Parsing Task were six, all participated to MDS, five presented results also for PDS, and two also for the constituency parsing track (see [10] for the results of the Evalita'09 constituency track). Among the participants of the dependency track, two participated also to the Evalita'07 Parsing Task [8], [9].

Two participant systems, namely UniTo_Lesmo_DPAR and CELI_Dini_DPAR, are rule–based parsers. UniTo_Lesmo_DPAR system is a wide coverage parser, which has been applied to various domains and which has been the starting point for the development of TUT. The CELI_Dini_DPAR uses the Xerox Incremental Parser (XIP, [1]) with a hand–written grammar which was developed through a cycle of implementation, verification and debugging exploiting TUT as a gold standard.

The other three participating systems belong to the class of statistical parsers, following different models, inference and learning methods. In particular, the FBKirst_Lavelli_DPAR and UniPi_Attardi_DPAR systems are both transition–based dependency parsers. The former uses MaltParser [26] with a non–deterministic transition system for mapping sentences to dependency trees (Covington's non–projective system [15]) and a SVM classifier to predict the next transition for every possible system configuration. UniPi_Attardi_DPAR uses DeSR, a Shift/Reduce deterministic transition–based parser that by using special rules is able to handle non–projective dependencies in linear time complexity; in particular, in Evalita'09 the system is tested using three different configurations – namely a left to right DeSR, right to left DeSR, and a stacked Reverse Revision system – whose final output is eventually combined using a linear time method

[14] The name of each system that participated to the contest is composed according to the following pattern: institution_author_DPAR, where DPAR stands for Dependency Parsing (to distinguish them from possible data of constituency parsing by the same participant).

[2]. As machine learning algorithms, DeSR uses SVM and Multilayer Perceptron. The third system is by Søgaard and Rishøj; it uses vine parsing algorithm [18] and a two stage approach to create labeled dependency trees. They use a first–order MIRA–informed Covington algorithm [15] as their baseline parser and POS–specific hard constraints on dependency length.

The evaluation of the participation results for the dependency track is presented separately for each subtask, respectively in tables 3 and 5 for MDS, and 4 and 6 for PDS. Assuming LAS as the main evaluation measure, we can observe that the best results for the MDS (see table 3) have been achieved by both the UniTo_Les-mo_DPAR and the UniPi_Attardi_DPAR, since the difference between their scores cannot be considered as statistically significant according to the p-value[15]. The average LAS and UAS of the participants is respectively 82.88 and 87.96.

**Table 3.** Dependency parsing MDS: evaluation on all the test set (240 sentences).

| **LAS** | **UAS** | **p-value** | **Participant** |
|---|---|---|---|
| 88.73 | 92.28 | 0.472 | UniTo_Lesmo_DPAR |
| 88.67 | 92.72 | 0.0001 | UniPi_Attardi_DPAR |
| 86.5 | 90.96 | 0.005 | FBKirst_Lavelli_DPAR |
| 84.98 | 89.07 | 0.0001 | UniAmsterdam_Sangati_DPAR |
| 80.42 | 89.05 | 0.0001 | UniCopenhagen_Soegaard_DPAR |
| 68 | 77.95 | | CELI_Dini_DPAR |

In the PDS the best results have been achieved instead by the UniPi_Attardi_DPAR (see table 4), i.e. LAS 83.38 and UAS 87.71. For this subtask, the average LAS is 74.73 and the average UAS is 80.65.

With respect to the subcorpora that represent the text genres of TUT within the test set of the MDS, we can observe that the best results have been achieved on the set of sentences extracted from the civil law code (see table 5), i.e. 92.63 for LAS (by the UniPi_Attardi_DPAR) and 95.51 for UAS (by the UniAmsterdam_Sangati_DPAR). The worst results have been instead obtained on the shared test set, where the best results are LAS 84.68 and UAS 89.73 (by the UniTo_Lesmo_DPAR).

For what concerns PDS, although the ISST–TANL corpus is not further organised into subcorpora representative of different text genres we partitioned the test corpus into two different subsections, namely the shared test set (100

[15] The difference between two results is taken to be significant if $p < 0.05$ (see `http://depparse.uvt.nl/depparse-wiki/AllScores` and `http://nextens.uvt.nl/~conll/software.html#eval`).

**Table 4.** Dependency parsing PDS: evaluation on all the test set (260 sentences).

| LAS | UAS | p-value | Participant |
|---|---|---|---|
| 83.38 | 87.71 | 0.0001 | UniPi_Attardi_DPAR |
| 80.54 | 84.85 | 0.0012 | FBKirst_Lavelli_DPAR |
| 78.51 | 85.81 | 0.0001 | UniCopenhagen_Soegaard_DPAR |
| 73.44 | 80.80 | 0.0001 | UniTo_Lesmo_DPAR |
| 57.81 | 64.10 | | CELI_Dini_DPAR |

**Table 5.** Dependency parsing MDS: evaluation on the shared test set (100 sentences from newspaper), the civil law (100 sentences) and passage (40 sentences).

| shared | | civillaw | | passage | | Participant |
|---|---|---|---|---|---|---|
| **LAS** | **UAS** | **LAS** | **UAS** | **LAS** | **UAS** | |
| 82.60 | 89.17 | 92.63 | 95.38 | 90.10 | 92.90 | UniPi_Attardi_DPAR |
| 84.68 | 89.73 | 91.54 | 94.64 | 89.36 | 91.58 | UniTo_Lesmo_DPAR |
| 79.91 | 87.15 | 90.23 | 93.33 | 89.11 | 91.75 | FBKirst_Lavelli_DPAR |
| 76.66 | 87.99 | 89.93 | 95.51 | 87.87 | 93.89 | UniAmsterdam_Sangati_DPAR |
| 72.84 | 81.93 | 86.04 | 90.27 | 80.94 | 85.31 | UniCopenhagen_Soegaard_DPAR |
| 63.86 | 70.15 | 70.74 | 74.97 | 68.89 | 73.35 | CELI_Dini_DPAR |

**Table 6.** Dependency parsing PDS: evaluation on the shared test set (100 sentences from newspapers), and the remaining test corpus (160 sentences).

| shared | | rest | | Participant |
|---|---|---|---|---|
| **LAS** | **UAS** | **LAS** | **UAS** | |
| 84.67 | 88.99 | 82.70 | 87.04 | UniPi_Attardi_DPAR |
| 81.12 | 85.02 | 80.24 | 84.76 | FBKirst_Lavelli_DPAR |
| 78.61 | 85.26 | 78.45 | 86.10 | UniCopenhagen_Soegaard_DPAR |
| 75.12 | 82.58 | 72.56 | 79.88 | UniTo_Lesmo_DPAR |
| 60.78 | 67.07 | 56.27 | 62.55 | CELI_Dini_DPAR |

sentences) and the remaining test sentences (160). Due to the fact that the shared test set was built by enforcing a constraint relative to the sentence length which could not exceed 40 tokens, it is worth considering the results obtained by the parsing systems on the two corpora subsections. In fact, the average sentence length in the two subsections, both representing the same text type, is significantly different, namely 17.16 tokens in the shared test set against 20.59 in the remaining sentences. Within the same text type, the sentence length can be taken to be indicative – at least to some extent – of the linguistic complexity of the corpus. By comparing the results obtained in the two ISST–TANL corpus subsections (see table 6), in both cases the best LAS and UAS scores are obtained by the UniPi_Attardi_DPAR. It should be noted, however, that slightly higher LAS and UAS scores have been obtained by all systems with respect to the shared test set, thus confirming our hypothesis that the shared test set is less complex to parse than the rest.

It is worth now to compare the results obtained in the two subtasks. Let us start by comparing the overall results achieved in MDS and PDS as reported in tables 3 and 4. It can be noticed that the best results refer to the MDS and the difference from the best LAS in MDS to the best LAS in PDS is about 5.35, while 4.57 is the difference for the best UAS scores in MDS and PDS; the difference between the average scores in MDS and PDS is 8.15 for LAS and 7.31 for UAS.

Things change quite significantly if the comparison is carried out with respect to the shared test set (see first two columns of tables 5 and 6): in this case, the best scores are obtained by the UniTo_Lesmo_DPAR for the MDS (LAS: 84.68; UAS: 89.73) and by the UniPi_Attardi_DPAR for the PDS (LAS: 84.67; UAS: 88.99). Interestingly enough, no significant difference can be noticed between the best LAS and UAS scores obtained in the two subtasks; concerning the difference between the average results obtained in MDS and PDS, it ranges from 0.718 for LAS to 1.73 for UAS.

If we focus on the performance of individual parsing systems which participated to both subtasks, it is interesting to note that a significant difference can be observed in the performance in the two subtasks with respect to the shared test set. There are three parsing systems, namely the stochastic ones (UniPi_Attardi, FBKirst_Lavelli and UniCopenhagen_Søgaard), showing higher LAS scores in PDS, whereas the reverse holds for the rule–based parsers (i.e. UniTo_Lesmo_DPAR and CELI_Dini_DPAR) which achieve best results in the MDS.

## 6 Discussion

The results obtained in the dependency parsing contest are very promising and positively compare with other experiences in this area, also with the state of the art for English dependency parsing (LAS $89,61\%$) as well as for Japanese (LAS $91,65\%$) [27].

Due to the fact that MDS is based on the same (revised) treebank of Evalita'07, and PDS on the same (revised) resource used for Italian in the multi–lingual

track of CoNLL 2007 Shared Task on Dependency Parsing [27], the more obvious comparison that we can develop is with these experiences.

With respect to the previous edition of the dependency parsing task in Evalita'07, there is an impressive improvement of the results and not only for the best scores. For instance, the best result in Evalita'07 was LAS 86.94 and UAS 90.90 (of the UniTo_Lesmo_DPAR), while today is LAS 88.73 (both by the UniTo_Lesmo_DPAR in the MDS) and UAS 92.72 (by the UniPi_Attardi_DPAR in MDS); the average LAS is passed from 72.48 (in Evalita'07) to 82.88 (in the MDS), and the average UAS from 83.09 (in Evalita'07) to 87.96 (in the MDS).

In the comparison with the multi–lingual track of CoNLL 2007 Shared Task on Dependency Parsing, results are in line with the state–of–the–art dependency parsing. Results in CoNLL–2007 as far as Italian is concerned and PDS EVALITA–2009 show the same range of variation: for LAS, scores range from 84.40 to 59.75 in CoNLL–2007 and from 83.38 to 57.81 in PSD; for UAS, from 87.91 to 65.52 in CoNLL–2007 and from 87.71 to 64.10 in PSD.

The possibility of testing the participant systems on two subtasks involving different annotated corpora allowed us to gain two important lessons. First of all, the results confirm that statistical approaches show more flexibility in adapting themselves to new texts and domains. The top rule–based parser in MDS (UniTo_Lesmo_DPAR) scores no significantly better than the best stochastic parser (UniPi_Attardi_DPAR). Conversely, while the latter parser is still the best system in the PDS, with a reduction of only 3 points in the LAS score, the UniTo_Lesmo_DPAR achieves only the 73.44 of LAS, with more than 15 points of reduction with respect to its performance in the MDS; note that the distance becomes shorter in both cases if we focus on the shared test set only. The major robustness and adaptability of stochastic approaches is no surprise, but it is worth stressing that this widespread claim finds another neat confirmation in the Evalita'09 dependency parsing task. On the other hand, rule based systems have confirmed to be adequate to analyze linguistic texts where the annotation adhere to a very regular analysis, as the case of some parenthetical structures in TUT, where in contrast statistical systems are not able to recognize the simple (context-free) rule underlying the constructions.

A second important lesson can be gained by the system results on the shared test set by the stochastic systems. As we noted above, all these systems score significantly worse in the MDS shared test set, rather than in the PDS one. The differences here can be found at two different levels, namely the annotation scheme and the training corpora. Concerning the former, a plausible still provisional hypothesis we can put forward is that the TUT annotation scheme is responsible – at least to some extent – for this different performance. This fact raises the crucial issue (indeed one that is too often underestimated) that annotation schemes are not all equal, when they are used to create data for the training of statistical parsers. It could perhaps be the case that some syntactic distinctions encoded in one annotation scheme can not be easily learned by the parser, or simply that the they are too sparse in the training data, which therefore should be enlarged in a significant way. Whatever the specific reason of the

different performances of the systems in the shared test set, the results suggest the need for some deeper reflections on parsing annotation schemes, showing that the improvement of parsing technology should proceed hand in hand with the development of more suitable representations for annotating syntactic data. Another possible explanation is concerned with the features of the training corpora in the two subtasks: it could be the case that the TUT training corpus does not provide enough evidence to tackle the linguistic constructions occurring in the shared test set. This raises another important issue, concerning the composition of the training corpora, which according to the Evalita'09 results can play a significant role in the parsers performance. As in the previous case, further analysis is needed to understand better its role.

**Acknowledgments.** We would like to thank Livio Robaldo and Alessia Bianchini for their support in the annotation of the TUT test set, and Eva Maria Vecchi for her support with the ISST–TANL corpus. We also would like to thank Simone Marchi for his help with the Evalita'09 Pilot Dependency Task web site.

## References

1. At-Mokhtar, S., Chanod, J. P., Roux, C.: Robustness below shallowness: Incremental deep parsing. Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data, pp. 121–144 (2002)
2. Attardi, G., Dell'Orletta, F.: Reverse Revision and Linear Tree Combination for Dependency Parsing. In: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies short papers (NAACL HLT) 2009 conference, pp. 261–264. Boulder, Colorado (2009)
3. Bos, J., Bosco, C., Mazzei, A.: Converting a Dependency-Based Treebank to a Categorial Grammar Treebank for Italian. In: Proceedings of the 8th Workshop on Treebanks and Linguistic Theories. Milan, in press (2009)
4. Bosco, C., Lombardo, V., Vassallo, D., Lesmo, L.: Building a Treebank for Italian: a Data-driven Annotation Schema. In: Proceedings of LREC'00, pp. 99–106. Athens, Greece (2000)
5. Bosco, C.: A grammatical relation system for treebank annotation. Unpublished PhD thesis disussed at the University of Turin (2004)
6. Bosco, C., Lombardo, V.: Comparing linguistic information in treebank annotations. In: Proceedings of LREC'06, pp. 1770–1775. Genova (2006)
7. Bosco, C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Workshop on Linguistic Annotation at the ACL'07, pp. 164–167 (2007)
8. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale, vol. 12, pp. 30–33 (2007)
9. Bosco, C., Mazzei, A., Lombardo, V., Attardi, G., Corazza, A., Lavelli, A., Lesmo, L., Satta, G., Simi, M.: Comparing Italian parsers on a common treebank: the Evalita experience. In: Proceedings of LREC'08, pp. 2066–2073. Marrakesh, Morocco (2008)
10. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task 2009: constituency parsing and a Penn format for Italian. In: Proceedings of EVALITA 2009 (2009)

11. Boyd, A., Meurers, D.: Revisiting the impact of different annotation schemes on PCFG parsing: a grammatical dependency evaluation. In: Proceedings of the ACL Workshop on Parsing German - PaGe '08, Association for Computational Linguistics, pp- 24–32. Morristown, NJ, USA (2008)
12. Buchholz, S., Marsi, E.: CoNLL-X Shared Task on Multilingual Dependency Parsing. In: Proceedings of the CoNLL-X, pp. 149–164 (2006)
13. Collins, M., Hajic, J., Ramshaw, L., Tillmann, C.: A Statistical Parser of Czech. In: Proceedings of ACL'99, pp. 505-512 (1999)
14. Corazza, A., Lavelli, A., Satta, G., Zanoli, R.: Analyzing an Italian treebank with state-of-the-art statistical parser. In: Proceedings of TLT-2004, pp. 39–50 (2004)
15. Covington., M. A.: A fundamental algorithm for dependency parsing. In: Proceedings of 39th Annual ACM Southeast Conference, pp. 95–102 (2001)
16. Delmonte, R.: Strutture sintattiche dall'analisi computazionale di corpora di italiano. Franco Angeli, Milano (2008)
17. Dubey, A., Keller, F.: Probabilistic parsing for German using sister-head dependencies. In: Proceedings of ACL'03, pp. 96–103 (2003)
18. Eisner J., Smith N.A.: Parsing with soft and hard constraints on dependency length. In: Proceedings of the International Workshop on Parsing Technologies (IWPT), pp. 30–41. Vancouver (2005)
19. Gildea, D.: Corpus variation and parser performance. In: Proceedings of EMNLP'01, pp. 167–172 (2001)
20. Hudson, R.: Word grammar. Basil Blackwel, Oxford and New York (1984)
21. Kübler, S., Rehbein, I., van Genabith, J.: TePaCoC - a corpus for testing parser performance on complex German grammatical constructions. In: Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories. Groningen, The Netherlands (2009)
22. Levy, R., Manning, C.: Is it harder to parse Chinese, or the Chinese treebank? In: Proceedings of ACL'03, pp. 439–446 (2003)
23. Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., Mazzei, A., Lombardo, V., Bertagna, F., Calzolari, N., Toral, A., Bartalesi Lenzi, V., Sprugnoli, R., Speranza, M.: Evaluation of Natural Language Tools for Italian: EVALITA 2007. In: Proceedings of LREC'08, pp. 2536–2543. Marrakesh, Morocco (2008)
24. Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte R.: Building and using parsed corpora. In Abeill A. (ed.), Building and using Parsed Corpora, Language and Speech Series, pp. 189–210. Kluwer, Dordrecht (2003)
25. Montemagni, S., Simi, M.: The Italian dependency annotated corpus developed for the CoNLL–2007 Shared Task. Technical Report, January 2007, available at `http://www.ilc.cnr.it/tressi_prg/ISST@CoNNL2007/ISST/ISST@CoNNL2007.pdf`
26. Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., Marsi E.: MaltParser: a language-independent system for data-driven dependency parsing. Natural Language Engineering 13(2), pp. 95–135 (2007)
27. Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the EMNLP-CoNLL, pp. 915–932 (2007)
28. Nivre J., Nilsson J., Hall J.: Generalizing Tree Transformations for Inductive Dependency Parsing. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 968–975 (2007)
29. Kübler S., McDonald R., Nivre J.: Dependency parsing. Morgan and Claypool Publishers (2009)