10.12871/clicit201421

The Evalita 2014 Dependency Parsing task

Cristina Bosco[1], Felice Dell'Orletta[2], Simonetta Montemagni[2], Manuela Sanguinetti[1], Maria Simi[3]
[1]Dipartimento di Informatica - Università di Torino, Torino (Italy)
[2]Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR, Pisa (Italy)
[3]Dipartimento di Informatica - Università di Pisa, Pisa (Italy)
{bosco,msanguin@di.unito.it},
{felice.dellorletta,simonetta.montemagni@ilc.cnr.it}
simi@di.unipi.it

Abstract

English. The Parsing Task is among the "historical" tasks of Evalita, and in all editions its main objective has been to define and improve state-of-the-art technologies for parsing Italian. The 2014's edition of the shared task features several novelties that have mainly to do with the data set and the subtasks. The paper therefore focuses on these two strictly interrelated aspects and presents an overview of the participants systems and results.

Italiano. *Il "Parsing Task", tra i compiti storici di Evalita, in tutte le edizioni ha avuto lo scopo principale di definire ed estendere lo stato dell'arte per l'analisi sintattica automatica della lingua italiana. Nell'edizione del 2014 della campagna di valutazione esso si caratterizza per alcune significative novità legate in particolare ai dati utilizzati per l'addestramento e alla sua organizzazione interna. L'articolo si focalizza pertanto su questi due aspetti strettamente interrelati e presenta una panoramica dei sistemi che hanno partecipato e dei risultati raggiunti.*

1 Introduction

The Parsing Task is among the "historical" tasks of Evalita, and in all editions its main objective has been to define and improve state-of-the-art technologies for parsing Italian (Bosco and Mazzei, 2013). The 2014's edition of the contest features two main novelties that mainly deal with the internal organization into subtasks and the used data sets.

From Evalita 2007 onwards, different subtasks have been organized focusing on different aspects of syntactic parsing. In Evalita 2007, 2009 and 2011, the tracks were devoted to dependency parsing and constituency parsing respectively, both carried out on the same progressively larger dataset extracted from the Turin University Treebank (TUT[1]), which was released in two formats: the CoNLL–compliant format using the TUT native dependency tagset for dependency parsing, and the Penn Treebank style format of TUT–Penn for constituency parsing. This allowed the comparison of results obtained following the two main existing syntactic representation paradigms as far as Italian is concerned.

In order to investigate the behaviour of parsing systems trained on different treebanks within the same representation paradigm, in 2009 the dependency parsing track was further articulated into two subtasks differing at the level of used treebanks: TUT was used as the development set in the main subtask, and ISST–TANL (originating from the ISST corpus, (Montemagni et al., 2003)) represented the development set for the pilot subtask. Comparison of results helped to shed light on the impact of different training resources, differing in size, corpus composition and adopted annotation schemes, on the performance of parsers.

In Evalita 2014, the parsing task includes two subtasks focusing on dependency parsing only, with a specific view to applicative and multilingual scenarios. The first, henceforth referred to as *Dependency Parsing for Information Extraction* or DPIE, is a basic subtask focusing on standard dependency parsing of Italian texts, with a dual evaluation track aimed at testing both the performance of parsing systems and their suitability to Information Extraction tasks. The second subtask, i.e. *Cross–Language dependency Parsing* or CLaP, is a pilot multilingual task where a source Italian treebank is used to train a parsing model which is then used to parse other (not necessarily typologically related) languages.

[1]http://www.di.unito.it/~tutreeb

Both subtasks are in line with current trends in the area of dependency parsing. In recent years, research is moving from the analysis of grammatical structure to sentence semantics, as testified e.g. by the SemEval 2014 task "Broad-Coverage Semantic Dependency Parsing" aimed at recovering sentence–internal predicate–argument relationships for all content words (Oepen et al., 2014): in DPIE, the evaluation of the suitability of the output of participant systems to information extraction tasks can be seen as a first step in the direction of targeting semantically–oriented representations. From a multilingual perspective, cross–lingual dependency parsing can be seen as a way to overcome the unavailability of training resources in the case of under–resourced languages. CLaP belongs to this line of research, with focus on Italian which is used as source training language.

As far as the data set is concerned, in Evalita 2014 the availability of the newly developed *Italian Stanford Dependency Treebank* (ISDT) (Bosco et al., 2013) made it possible to organize a dependency parsing task with three main novelties with respect to previous editions:

1. the annotation scheme, which is compliant to *de facto* standards at the level of both representation format (CoNLL) and adopted tagset (Stanford Dependency scheme, (de Marneffe and Manning, 2008));

2. its being defined with a specific view to supporting Information Extraction tasks, a feature inherited from the Stanford Dependency scheme;

3. the size of the data set, much bigger (around two times larger) than the resources used in previous Evalita campaigns.

The paper is organized as follows. The next section describes the resources that were used and developed for the task. In sections 3 and 4, we will present the subtasks, the participants' systems approaches together with achieved results.

2 A new dataset for the Evalita Parsing Task

Over the last few years, Stanford Dependencies (SD) have progressively gained the status of *de facto* standard for dependency–based treebank annotation (de Marneffe et al., 2006; de Marneffe and Manning, 2008). The *Italian Stanford Dependency Treebank* (ISDT) is the standard-compliant treebank for the Italian language (Bosco et al., 2013; Simi et al., 2014), which was built starting from the *Merged Italian Dependency Treebank* (MIDT) (Bosco et al., 2012), an existing dependency-based Italian treebank resulting in its turn from the harmonization and merging of smaller resources (i.e. TUT and ISST–TANL, already used in previous Evalita campaigns) adopting incompatible annotation schemes. ISDT originates as the result of a joint effort of three research groups based in Pisa (Dipartimento di Informatica – Università di Pisa, and Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR) and in Torino (Dipartimento di Informatica – Università di Torino) aimed at constructing a larger and standard-compliant resource for the Italian language which was expected to create the prerequisites for crucial advancements in Italian NLP.

ISDT has been used in both DPIE and CLaP Evalita 2014 tasks, making it possible to compare parsers for Italian trained on a new, standard-compliant and larger resource, and to assess cross-lingual parsing results using a parser trained on an Italian resource.

The composition of the ISDT resource released for development in both tasks is as follows:

- a data set of around 97,500 tokens, obtained by conversion from TUT, representative of various text genres: legal texts from the Civil code, the Italian Constitution, and European directives; newspaper articles and wikipedia articles;

- a data set of around 81,000 tokens, obtained by conversion from ISST–TANL, including articles from various newspapers.

For what concerns the representation format, ISDT data comply with the standard CoNLL-X format, with UTF-8 encoding, as detailed below:

- sentences are separated by an empty line;

- each token in a sentence is described by ten tab–separated columns;

- columns 1–6 are provided by the organizers and contain: token id, word form, lemma, coarse-grained PoS, fine-grained PoS, and morphology;

- parser results are reported in columns 7 and 8 representing respectively the head token id and the dependency linking the token under description to its head;
- columns 9-10 are not used for the tasks and contain an underscore.

The used annotation scheme follows as close as possible the specifications provided in the SD manual for English (de Marneffe and Manning, 2008), with few variations aimed to account for syntactic peculiarities of the Italian language: the Italian localization of the Stanford Dependency scheme is described in detail in Bosco et al. (2013). The used tagset, which amounts to 41 dependency tags, together with Italian-specific annotation guidelines is reported in the dedicated webpage[2]. For what concerns the rendering of copular verbs, we preferred the standard option of making the copular verb the head of the sentence rather than the so-called Content Head (CH) option, that treats copular verbs as auxiliary modifiers of the adjective or predicative noun complement.

As stated in de Marneffe and Manning (2008), different variants of the typed dependency representation are available in the SD annotation scheme. Among them it is worth reporting here:

- the *basic* variant, corresponding to a regular dependency tree;
- the *collapsed* representation variant, where dependencies involving prepositions, conjunctions as well as information about the antecedent of relative pronouns are collapsed to get direct dependencies between content words. This collapsing is often useful in simplifying patterns in relation extraction applications;
- the *collapsed dependencies with propagation of conjunct dependencies* variant including – besides collapsing of dependencies – also the propagation of the dependencies involving conjuncts.

Note that in the collapsed and propagated variants not all words in a sentence are necessarily connected nor form a tree structure: this means that in these variants a sentence is represented as a set of binary relations (henceforth, we will refer to this representation format as RELS output). This is a semantically oriented representation, typically connecting content words and more suitable for relation extraction and shallow language understanding tasks.

In a similar vein and following closely the SD strategy, in Evalita 2014 different variants of the ISDT resource are exploited. The basic and *collapsed/propagated* representation variants are used in DPIE, whereas CLaP is based on the basic representation variant only. To obtain the *collapsed/propagated* version of ISDT, as well as the participants output, a CoNLL–to–RELS converter was implemented, whose result consists in a set of relations represented as triplets, i.e. name of the relation, governor and dependent. Note that following the SD approach, conjunct propagation is handled only partially by focusing on a limited and safe set of cases.

For CLaP, the Universal version of the basic ISDT variant (henceforth referred to as "uISDT") was used, annotated according to the Universal Stanford Dependencies scheme defined in the framework of *The Universal Dependency Treebank Project* [3]. uISDT was obtained through conversion from ISDT.

3 The Dependency Parsing for Information Extraction subtask

3.1 Task description

DPIE was organized as a classical dependency parsing task, where the performance of different parsers, possibly following different paradigms (statistical, rule-based, hybrid), can be compared on the basis of the same set of test data provided by the organizers.

In order to allow participants to develop and tune their systems, the ISDT resource was split into a training set (165,975 tokens) and a validation set (12,578 tokens). For the purposes of the final evaluation, we developed a new test data set, for a total of 9,442 tokens articulated into three subsets representative of different textual genres:

- a data set of 3,659 tokens extracted from newspaper texts and particularly rich in factual information, a feature making it suitable for evaluating Information Extraction capabilities (henceforth, IE–test)[4];

[2]See: http://medialab.di.unipi.it/wiki/ISDT

[3]https://code.google.com/p/uni-dep-tb/

[4]These texts are part of a benchmark used by Synthema

- a data set of 3,727 tokens from newspaper articles (henceforth, News–test);
- a data set of 2,056 tokens from European directives, annotated as part of the 2012 Shared Task on Dependency Parsing of Legal Texts (Dell'Orletta et al., 2012) (henceforth, SPLeT–test).

The main novelty of this task consists in the methodology adopted for evaluating the output of the participant systems. In addition to the Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS), which represent standard metrics in dependency parsing, we wanted to provide an alternative and semantically-oriented metric to assess the ability of the parsers to produce suitable and accurate output for information extraction applications. Whereas LAS and UAS were computed against the basic SD variant, represented in the CoNLL format, the semantically-oriented evaluation was computed against the *collapsed and propagated* version of the parsers output and was based on a subset of the relation types selected as more relevant, i.e. semantically-loaded.

The dependency relations that were selected for the semantically-oriented evaluation are 18 out of the 41 dependency types, namely: `acomp`, `advcl`, `advmod`, `amod`, `ccomp`, `dobj`, `iobj`, `mark`, `nn`, `nnp`, `npadvmod`, `nsubj`, `nsubjpass`, `prep`, `rcmod`, `tmod`, `vmod`, `xcomp`. Most of them link content words. In this case, used evaluation metrics are: *Precision*, the fraction of correct relations extracted over the total of extracted relations; *Recall*, the fraction of correct relations extracted over the relations to be found (according to the gold standard); and F1, the harmonic mean of the two.

Participants were allowed to use external resources, whenever they deemed it necessary, and to submit multiple runs. In the following section, we describe the main features of the participants' systems, together with achieved results.

3.2 Systems description and results

For DPIE, four participants submitted their results. Here follows an overview of the main features of their parsing systems[5], in order to provide a key to interpret the results achieved.

(`http://www.synthema.it/`) on a common project and kindly offered for the task.

[5]For a detailed description of each participant's system, please refer to the corresponding technical report.

Table 1 summarizes the main features of participants systems, based on three main parameters: 1) whether a single parser or a parser combination has been used; 2) the approach adopted by the parser (statistical, rule-based or hybrid), and 3) whether only the training and development sets provided by the organizers (DPIE only) or rather external resources (Other) have been used.

Participants mostly used publicly available state-of-the-art parsers and used them in different combinations for the task. The parsers that have been used are:

- MALT parser (Nivre et al., 2006): a transition–based dependency parser written in Java, which uses a SVM classifier;
- DeSR parser (Attardi et al., 2009): a transition–based dependency parser written in C++, which can be used with several classifiers including a Multi–Layer Perceptron;
- MATE parser (Bohnet, 2010): the MATE tools, written in Java, include both a graph-based parser and a transition-based parser. The transition-based MATE takes into account complete structures as they become available to re-score the elements of a beam, combining the advantages of transition–based and graph–based approaches. Efficiency is gained through Hash Kernels and exploiting parallelism.
- TurboParser (Martins et al., 2013): a C++ package that implements graph-based dependency parsing exploiting third-order features.
- ZPar (Zang and Nivre, 2011): a transition-based parser that leverages its performance by using considerably richer feature representations with respect to other transition-based parsers. It supports multiple languages and multiple grammar formalisms, but it was especially tuned for Chinese and English.

We provide below a short description of the parsing solutions adopted by each participant.

Attardi et al. (University of Pisa) The final runs submitted by this team used a combination of four parsers: MATE in the standard graph-based configuration; DeSR, with the Multilayer Perceptron algorithm; a new version of the DeSR parser, introducing graph completion; TurboParser.

Participant	#Parser/s used	Approach	Development
Attardi et al.	Combination	Statistical	DPIE only
Lavelli	Combination	Statistical	DPIE only
Mazzei	Combination	Statistical	DPIE only
Grella	Single	Hybrid	Other

Table 1: Systems overview based on number of parsers, approach and resources used.

Parser combination was based on the technique described in Attardi, Dell'Orletta (2009). Submitted runs differ at the level of the conversion applied to the corpus, performed in pre- and a post-processing steps, consisting in local restructuring of the parse-trees.

Lavelli (FBK-irst) This participant used the following parsers: ZPar; the graph-based MATE parser combined with the output of TurboParser (full model) using stacking; Ensemble (Surdenau and Manning, 2010), a parser that implements a linear interpolation of several linear-time parsing models. For the submission, the output of the following 5 parsers have been combined: graph-based MATE parser, transition-based MATE parser, TurboParser (full model), MaltParser (Nivre's arc-eager, PP-head, left-to-right), and MaltParser (Nivre's arc-eager, PP-head, right-to-left).

Mazzei (University of Torino) The final runs submitted by this participant resulted from the combination of the following parsers: MATE; DeSR parser with the Multi-Layer Perceptron algorithm; MALT parser.
Parser combination was based on the technique described in (Mazzei and Bosco, 2012), which applies a majority vote algorithm.

Grella (Parsit, Torino) This participant used a proprietary transition-based parser (ParsIt) based on a Multi-Layer Perceptron algorithm. The parser includes PoS tagging and lemmatization, using a dictionary of word forms with associated PoS, lemmas and morphology, and a subcategorization lexicon for verbs, nouns, adjectives and adverbs. In addition, the parser exploits a vectorial semantic space obtained by parsing large quantities of text with a basic parser. The parser was trained on a set of around 7,000 manually-annotated sentences, different from the ones provided for the task, and the output was converted into the ISDT scheme with a rule-based converter. The development resources were used in order to develop and test the converter from the output parser format into the ISDT representation format.

Tables 2 and 3 report the results for each run submitted by each participant system for the first evaluation track. In Table 2, the overall performance of parsers is reported in terms of achieved LAS/UAS scores, without considering punctuation. Since achieved results were very close for most of the runs, we checked whether the difference in performance was statistically significant by using the test proposed by Dan Bikel[6]. We considered that two runs differ significantly in performance when the computed p value is below 0.05. This was done by taking the highest LAS score and assessing whether the difference with subsequent values was significant or not; the highest score among the remaining ones whose difference was significant was taken as the top of the second cluster. This was repeated until the end of the list of runs. In Table 2, we thus clustered together the LAS of the runs whose difference was not significant according to the Bikel's test: the top results include all runs submitted by Attardi et al. and one of the runs by Lavelli.

Table 3 reports the performance results for each subset of the test corpus, covering different textual genres. It can be noticed that the best results are achieved with newspaper texts, corresponding to the IE and News test sets: in all runs submitted by participants higher results are obtained with the IE-test, whereas with the News-test LAS/UAS scores are slighly lower. As expected, for all participants the worse results refer to the test set represented by legal texts (SPLeT).

The results of the alternative and semantically-oriented evaluation, computed against the *collapsed* and *propagated* version of the systems out-

[6]The Randomized Parsing Comparator, whose script is now available at: `http://pauillac.inria.fr/~seddah/compare.pl`

Participant	LAS	UAS
Attardi run1	87.89	90.16
Attardi run3	87.84	90.15
Attardi run2	87.83	90.06
Lavelli run3	87.53	89.90
Lavelli run2	87.37	89.94
Mazzei run1	87.21	89.29
Mazzei run2	87.05	89.48
Lavelli run1	86.79	89.14
Grella	84.72	90.03

Table 2: DPIE subtask: participants' results, according to LAS and UAS scores. Results are clustered on the basis of the statistical significance test.

	IE	News	SPLeT
Attardi run1	88.64	87.77	86.77
Attardi run3	88.29	88.25	86.33
Attardi run2	88.55	88.09	86.01
Lavelli run3	88.71	87.68	85.21
Lavelli run2	88,8	87,29	84,99
Mazzei run1	88,2	87,64	84,71
Mazzei run2	88,2	86,94	85,21
Lavelli run1	87,72	87,39	84,1
Grella	86,96	84,54	81,08

Table 3: Systems results in terms of LAS on different textual genres.

put, are reported in Table 4, where Precision, Recall and F1 score for the set of selected relations are reported for each participant's run. In this case we did not perform any test of statistical significance. By comparing the results reported in tables 2 and 4, it is interesting to note differences at the level of the ranking of achieved results: besides the 3 runs by Attardi et al. which are top-ranked in both cases although with a different internal ordering, two runs by Mazzei (run2) and Lavelli (run1) respectively from the second cluster in table 2 show higher precision and recall than e.g. run3 by Lavelli which was among the top-ranked ones. The reasons underlying this state of affairs should be further investigated. It is however interesting to report that traditional parser evaluation with attachment scores (LAS/UAS) may not be always helpful for researchers who want to find the most suitable parser for their IE application, as suggested among others by Volokh and Neumann (2012).

We also performed a dependency–based evaluation, in order to identify low scored relations shared by all parsers. It turned out that `iobj` (indirect object), `nn` (noun compound modifier), `npadvmod` (noun phrase as adverbial modifier), `tmod` (temporal modifier) are hard to parse relations for all parsers, although at a different extent: their average F1 score computed on the best run of each participant ranges between 46,70 (`npadvmod`) and 56,25 (`tmod`). This suggests that either we do not have enough information for dealing with semantically–oriented distinctions (as in the case of `iobj`, `npadvmod` and `tmod`), or more simply the dimension of the training corpus is not sufficient to reliably deal with them (see the `nn` relation whose frequency of occurrence in Italian is much lower than in English).

Participant	Precision	Recall	F1
Attardi run1	81.89	90.45	85.95
Attardi run3	81.54	90.37	85.73
Attardi run2	81.57	89.51	85.36
Mazzei run2	80.47	89.98	84.96
Lavelli run1	80.30	88.93	84.39
Mazzei run1	80.88	87.97	84.28
Lavelli run2	79.13	87.97	83.31
Grella	80.15	85.89	82.92
Lavelli run3	78.28	88.09	82.90

Table 4: DPIE subtask: participants' results, according to Precision, Recall and F1 score of selected relations, computed against the *collapsed* and *propagated* variant of the output.

4 The Cross-Language dependency Parsing subtask

CLaP is a cross-lingual transfer parsing task, organized along the lines of the experiments described in McDonald et al. (2013). In this task, participants were asked to use their parsers trained on the Universal variant of ISDT (uISDT) on test sets of other languages, annotated according to the Universal Dependency Treebank Project guidelines. The languages involved in the task are all the

languages distributed from the Universal Dependency Treebank Project with the exclusion of Italian, i.e.: Brazilian-Portuguese, English, Finnish, French, German, Indonesian, Japanese, Korean, Spanish and Swedish.

Participant systems were provided with:

- a development set consisting of uISDT, the universal version of ISDT used for training in DPIE and obtained through automatic conversion, and validation sets of about 7,500 tokens for each of the eleven languages of the Universal Dependency Treebank;
- a number of test sets (one for each language to be dealt with) for evaluation, with gold PoS and morphology and without dependency information; these data sets consist of about 7,500 tokens for each of the eleven languages of the Universal Dependency Treebank. Test sets were built by randomly extracting sentences from SD treebanks available at `https://code.google.com/p/uni-dep-tb/`. For languages which opted for the Content Head (CH) option in the treatment of copulas, sentences with copular constructions were discarded.

The use of external resources (e.g. dictionaries, lexicons, machine translation outputs, etc.) in addition to the corpus provided for training was allowed. Participants in this task were also allowed to focus on a subset of languages only.

4.1 System description and results

Just one participant, Mazzei, submitted the system results for this task. He focused on four languages only: Brazilian-Portuguese, French, German and Spanish.

Differently from the approach previously adopted, for CLaP Mazzei used a single parser, the MALT parser. The adopted strategy is articulated in three steps as follows: 1) each analyzed test set was word-for-word translated into Italian using Google Translate; 2) the best feature configuration was selected for each language using MaltOptimizer (Ballesteros, 2012) on the translated development sets; 3) for each language the parsing models were obtained by combining the Italian training set with the translated development set.

Table 5 reports the results in terms of LAS, UAS and also LA (Label Accuracy Score). Unlike DPIE, the punctuation is included in the evaluation metrics.

	LAS	**UAS**	**LA**
Brazilian-Portuguese	71.70	76.48	84.50
French	71.53	77.30	84.41
German	66.51	73.86	79.14
Spanish	72.39	77.83	83.30

Table 5: CLaP results in terms of LAS, UAS, LA on the test sets.

The reported results confirm that using training data from different languages can improve accuracy of a parsing system on a given language: this can be particularly useful for improving the accuracy of parsing less–resourced languages. As expected, the accuracy achieved on the German test set is the lowest: typologically speaking, within the set of languages taken into account German is the most distant language from Italian. These results can be considered in the framework of the work proposed by Zhao et al. (2009), in which the authors translated word-for-word the training set in the target language: interestingly, Mazzei followed the opposite approach and achieved promising results.

5 Acknowledgements

Roberta Montefusco implemented the scripts for producing the collapsed and propagated version of ISDT and for the evaluation of systems in this variant. Google and Synthema contributed part of the resources that were distributed to participants.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proceedings of the 2nd Workshop of Evalita 2009*, Springer–Verlag, Berlin Heidelberg.

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse Revision and Linear Tree Combination for Dependency Parsing. In *Proceedings of Human Language Technologie* (NAACL 2009), ACL, pp. 261–264.

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An Optimization Tool for MaltParser. In *Proceedings of the System Demonstration Session of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2012). ACL, pp. 58–62.

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING 2010). ACL, pp. 89–97.

Cristina Bosco and Alessandro Mazzei. 2013. The EVALITA Dependency Parsing Task: from 2007 to 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone and Emanuele Pianta (eds.) *Evaluation of Natural Language and Speech Tools for Italian*, Springer–Verlag, Berlin Heidelberg, pp. 1–12.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2012. Harmonization and merging of two Italian dependency treebanks. In *Proceedings of the LREC Workshop on Language Resource Merging*, ELRA, pp. 23–30.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th ACL Linguistic Annotation Workshop and Interoperability with Discourse*, ACL, pp. 61–69

Felice DellOrletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, Giulia Venturi. 2012. The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts. In *Proceedings of the 4th Workshop on Semantic Processing of Legal Texts (SPLeT 2012)*, held in conjunction with LREC 2012, Istanbul, Turkey, 27th May, pp. 42-51.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), ELRA, pp. 449–454.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependencies representation. In *Coling2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, CrossParser 08*, ACL, pp. 1–8 .

Marie-Catherine de Marneffe and Christopher D. Manning 2008. *Stanford Typed Dependencies manual* (Revised for the Stanford Parser v. 3.3 in December 2013). `http://nlp.stanford.edu/software/dependencies_manual.pdf`.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non–projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), ACL, pp. 617–622.

Alessandro Mazzei, and Cristina Bosco. 2012. Simple Parser Combination. In *Proceedings of Semantic Processing of Legal Texts (SPLeT-2012)*, ELRA, pp. 57–61.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of 51st annual meeting of the Association for Computational Linguistics* (ACL'13), ACL, pp. 92-97.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Alessandro Lenci, Ornella Corazzari, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Roberto Basili, Remo Raffaelli, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Fabio Pianesi, Nadia Mana and Rodolfo Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Anne Abeillè (ed.) *Building and Using syntactically annotated corpora*, Kluwer, Dordrecht, pp. 189–210.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC '06), ELRA, pp. 2216–2219.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, Yi Zhang. 2014. SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Republic of Ireland, pp. 63–72.

Maria Simi, Cristina Bosco and Simonetta Montemagni. 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC'14), ELRA, pp. 83–90.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble Models for Dependency Parsing: Cheap and Good. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ACL, pp. 649–652.

Alexander Volokh and Günter Neumann. 2012. Task–oriented dependency parsing evaluation methodology. In Proceedings of the *IEEE 13th International Conference on Information Reuse & Integration, IRI*, Las Vegas, NV, USA, August 8-10, 2012, pp. 132–137.

Yue Zhang and Joakim Nivre. 2011. Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL*, ACL, pp. 188-193.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of ACL/IJCNLP*, ACL, pp. 55-63.