Looking back to the EVALITA Constituency Parsing Task: 2007-2011

Looking Back to the EVALITA Constituency Parsing Task: 2007-2011*

Cristina Bosco, Alessandro Mazzei, and Alberto Lavelli

[1] Dipartimento di Informatica, Università di Torino,
Corso Svizzera 185, I-10149 Torino, Italy
{bosco,mazzei}@di.unito.it
[2] FBK-irst,
via Sommarive 18, I-38123 Povo (TN), Italy
lavelli@fbk.eu

Abstract. The aim of the EVALITA Parsing Task (EPT) is at defining and extending Italian state-of-the-art parsing by encouraging the application of existing models and approaches, comparing paradigms and annotation formats. Therefore, in all the editions, held respectively in 2007, 2009 and 2011, the Task has been organized around two tracks, namely Dependency Parsing and Constituency Parsing, exploiting the same data sets made available by the organizers in two different formats.

This paper describes the Constituency Parsing Task assuming a historical perspective, but mainly focussing on the last edition held in 2011. It presents and compares the resources exploited for development and testing, the participant systems and the results, showing also how the resources and scores improved during the three editions of this contest.

Keywords: Constituency Parsing, Evaluation, Italian.

1 Introduction

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.

As in previous editions, in 2007 [7,10] and 2009 [9,8], in the current edition of the EPT held in 2011 the focus has been 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. Therefore the task has been organized in two tracks, namely Constituency Parsing and Dependency Parsing, giving again the same data for development and testing (respectively annotated in dependency and constituency format) in both tracks. In this way new materials for the development of cross-paradigm analyses about Italian parsing have been made available. The aim of the EPT is in fact at contributing to the literature on parsing results giving information about the behavior of parsing models on Italian, which is a morphologically rich language currently less-resourced with respect e.g. to English or German.

* This work has been partially funded by the PARLI Project (Portale per l'Accesso alle Risorse Linguistiche per l'Italiano MIUR PRIN 2008).

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 46–57, 2012.

In previous EPT editions, the results for dependency parsing have reached performance not far from the state of the art for English, while those for constituency showed a higher distance from it. Instead, in the current edition the major improvement has to be referred to constituency parsing, where scores meaningfully more proximate to the state of the art for English have been achieved. Nevertheless, these results confirm that the scores published for English (of around F 92.1 [14]) using the Penn Treebank (PTB), remain currently irreproducible for Italian.

In this paper, we will especially analyze the constituency parsing track mainly referring to the 2011 edition of the competition, putting it in the context of the more general work on applying statistical parsing techniques to Italian in a comparative perspective with respect to the previous editions of the EPT. In particular, we develop some comparison among the current and previous editions with respect to the data sets, the participant systems and the achieved results.

The paper is organized as follows. In the next section, we summarize the related experiences in statistical constituency parsing in general and with respect to Italian language. In section three, there is a survey of the EPT 2011 with all the information about data sets for training and testing, and the participation results. We conclude with a section devoted to a discussion of the EVALITA experience for constituency parsing.

2 A Bit of History

2.1 Related Work on Constituency Parsing

The starting point of the work on statistical parsing was the release of the PTB [20] in 1992 and the definition of the Parseval metrics [2], which are now considered as the standard de facto evaluation measures in parsing. In the following years, different approaches have been developed focussing at the beginning mainly on PTB and more in general on English language.

More recently, treebanks for languages other than English have been developed, and some limitations of the state-of-the-art approaches emerged. In particular, the initial bias towards specific characteristics of PTB and in general of English. In the perspective of the exploration of a wider variety of languages, it is particularly important the series of workshops on Statistical Parsing of Morphologically Rich Languages [28,26].

2.2 Parsing Italian before EVALITA

The work on statistical parsing of Italian started in 2004 as described in [13]. It was triggered by the availability of the first Italian treebanks, i.e. the Italian Syntactic-Semantic Treebank (ISST, [21]) and the Turin University Treebank (TUT, see Section 3.1). Nevertheless, only the former was exploited in the experiments described in the paper, while the latter was at that time available only in dependency format.

The ISST, developed by the Istituto di Linguistica Computazionale (ILC) and by the University of Pisa, has four levels: morpho-syntactic, two syntactic levels (constituent structure and functional relation), and lexico-semantic. The total size of this treebank is 305,547 word tokens, but only part of the ISST was syntactically annotated at the constituent structure level.

In [13], two state-of-the-art parsers, namely the Stanford parser [16,17] and the Bikel parser [1], were in fact compared on the basis of a portion of the ISST which contains about 3,000 sentences (89,941 tokens) from the financial domain, and some experiment is performed on a subset of the WSJ of size comparable with ISST.

The Bikel's parser is an implementation of Collins' probabilistic parser [11] and can be viewed in the framework of the lexicalized grammar approaches traditionally considered for Probabilistic Lexicalized Context-Free Grammars (PLCFGs). Each parse tree is represented as the sequence of decisions corresponding to the head-centered, top-down derivation of the tree. Probabilities for each decision are conditioned on the lexical head.

Table 1. Results of Bikel parser and of different configurations of the Stanford parser on WSJ (training: sections 02 & 03; test: section 23) and on ISST (with 10-fold cross-validation)

	LR	**LP**	F_1	**Parser**
WSJ	83.41	84.02	83.71	Bikel
	77.89	77.04	77.46	Stanford - noPA
	78.69	75.89	77.27	Stanford - PA
	78.26	76.00	77.12	Stanford - noPA tagPA
	79.70	75.76	77.68	Stanford - PA tagPA
	78.42	76.52	77.46	Stanford - noPA tagPA h=2
	79.56	75.97	77.73	Stanford - PA tagPA h=2
ISST	68.58	68.40	68.49	Bikel
	59.88	60.00	59.94	Stanford - noPA
	60.78	59.36	60.06	Stanford - PA
	67.08	64.72	65.88	Stanford - noPA tagPA
	66.42	62.15	64.21	Stanford - PA tagPA
	66.96	64.88	65.80	Stanford - noPA tagPA h=2
	66.31	62.19	64.18	Stanford - PA tagPA h=2

The Stanford lexicalized probabilistic parser implements a factored model, which considers separately the Probabilistic Context Free Grammar (PCFG) phrase structure model and the lexical dependency model. The preferences corresponding to these two different models are then combined by efficient exact inference, using an A* algorithm. The Stanford parser allows different configurations of the model, by specializing non-terminal labels on the basis of the parent tag (parent annotation or PA) and of the sisters (hMarkov=2). Also the Part of Speech (PoS) tags can be specialized on the basis of the parent tag (tagPA).

The results reported in Table 1 for Bikel and Stanford parsers show a substantial difference in performance with the state-of-the-art results on English.

2.3 EVALITA 2007 & 2009

As far as Italian is concerned, the experiences in statistical constituency parsing done in the context of the EVALITA evaluation campaign represent the next step after those described in [13]. In this section, we describe the 2007 and 2009 editions of the EPT (see section 3.1 for the current edition). We focus, in particular, on the size and features of the datasets, a short description of the participant systems and the achieved results.

EVALITA 2007. For the EPT 2007, the training set was composed by 2,000 sentences that correspond to about 53,700 tokens[1]. It included two equally sized subcorpora, one from the Italian Civil Law Code (i.e. CODCIV) and one from Italian newspapers (i.e. NEWSPAPER), both made available in the TUT–Penn format (see section 3.1 for more details about the format). The test set was composed by 200 sentences (4,962 tokens) and is balanced with respect to genres as the training set.

The teams which participated were two, namely that of Anna Corazza, Alberto Lavelli and Giorgio Satta, and that of Emanuele Pianta.

The team composed by Corazza, Lavelli, and Satta [12] participated with an adaptation to Italian of Collins' probabilistic parser (as implemented by Dan Bikel) achieving the best result for this task. Pianta [24] instead participated with a left corner parser for Italian, based on explicit rules manually coded in a unification formalism.

The results reported in Table 2 refer respectively to the evaluation on the entire test set (all test set), on the parts of the test set respectively extracted from the Civil Law Code (CODCIV) and newspapers (NEWSPAPER). Even for the best scoring system, i.e. that of Corazza, Lavelli, and Satta, the results were very far from those known for English at that time. As for the subcorpora, we can see that the best results refer to the subcorpus including legal text.

EVALITA 2009. For the EPT 2009, the training set has been increased with 200 new sentences to include 2,200 sentences that correspond to about 58,600 tokens. As in 2007, the corpus is organized in two subcorpora, i.e. one from Italian newspaper (NEWSPAPER) and one from the Italian Civil Law Code (CODCIV), made available in the TUT–Penn format (see section 3.1 for more details about the format). The test set included 200 sentences (4,074 tokens) and is balanced as the development set, one half from newspapers and the other half from the Civil Law Code.

The teams which participated were again two, that of Alberto Lavelli (FBK-irst) and Anna Corazza (Università "Federico II" di Napoli) and that of Federico Sangati (University of Amsterdam), i.e. Lavelli et al. [19] and Sangati [25]. The parser from FBK-irst and Università "Federico II" di Napoli adopts a probabilistic context-free grammars model, while that from the University of Amsterdam adopts the DOP model.

[1] Only words and punctuation marks are considered as tokens.

Table 2. EPT 2007: results on the entire test set and on the two subcorpora (CODCIV and NEWSPAPER)

	LR	**LP**	F_1	**Participant**
all test set	70.81	65.36	67.97	Corazza, Lavelli, Satta
	38.92	45.49	41.94	Pianta
CODCIV	74.31	70.11	72.15	Corazza, Lavelli, Satta
	41.55	49.92	45.35	Pianta
NEWSPAPER	67.31	60.60	63.78	Corazza, Lavelli, Satta
	36.28	41.06	38.52	Pianta

Table 3. EPT 2009: results on the entire test set and on the two subcorpora (NEWSPAPER and CODCIV)

	LR	**LP**	F_1	**Participant**
all test set	80.02	77.48	78.73	Lavelli, Corazza
	78.53	73.24	75.79	Sangati
CODCIV	83.15	78.33	80.66	Lavelli, Corazza
	80.47	73.69	76.93	Sangati
NEWSPAPER	76.08	76.34	76.21	Lavelli, Corazza
	76.08	72.65	74.33	Sangati

Lavelli and Corazza exploited for this edition of the EPT the Berkeley parser[2] [22] which outperformed the Bikel's parser, i.e. the best scored system applied by the same team in 2007. The Berkeley parser is based on a hierarchical coarse-to-fine parsing, where a sequence of grammars is considered, each being the refinement, namely a partial splitting, of the preceding one. Its performance is at the state of the art for English on the PTB and it outperforms other parsers in languages different from English, namely German and Chinese [22]. Indeed, a good compromise between efficiency and accuracy is obtained by a node splitting procedure, where splits which do not help accuracy are immediately pruned. Training is based on a discriminative framework, as discussed in [23]. Aiming at maximizing F_1, it has been applied a parser version without reranking according to likelihood.

Sangati parser is an adaptation of the Data Oriented Parsing (DOP) model [3]. This is a generative statistical model that computes parsing probabilities on the basis of tree

[2] `http://nlp.cs.berkeley.edu/Main.html#Parsing`

fragments. Parsing trees in the training set are decomposed into sub-trees, i.e. fragments, by assuming a tree-substitution combination operator. The frequencies of these fragments are the basis to compute scores in the parsing phase. Indeed, each potential parsing tree for sentences in the test set is scored by using the probabilities of the fragments in the parsing tree. In contrast to standard DOP model, Sangati decided to use only those fragments which are occurring at least two times in the training data.

In Table 3, the results of the evaluation both on the entire test set and its subcorpora are presented. We can observe that the best results have been achieved by Lavelli, but according to the p-value the difference between the first and second score for recall cannot be considered as significant[3]. As in 2007 the best scores refer to the legal text genre.

3 EVALITA 2011

For the last edition of the EVALITA evaluation campaign the Parsing Task included the same data sets both for dependency and constituency, but the data for training were improved with respect to quality and quantity with respect to the past.

3.1 EVALITA 2011 Dataset

In the EPT 2011, the data proposed for the training and development of parsing systems are, as in previous editions, from TUT, the treebank for Italian developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin[4]. TUT has been newly released for the last time in 2011, after automatic and manual revisions, in an improved version where both the consistency of the annotation and the size of the treebank are improved with respect to the previous releases. In particular, for what concerns size, TUT is currently similar to the other Italian resources, i.e. Venice Italian Treebank [27] and ISST–TANL [21] (see also the subsection 2.2). Moreover, TUT makes available different annotation formats [5] that allow for a larger variety of training and testing for parsing systems and for meaningful comparisons with theoretical linguistic frameworks, i.e. the native TUT, the TUT-Penn, and the CCG-TUT which is an application to Italian of the Combinatory Categorial Grammar [4].

3.2 Development Set

The data format adopted for constituency parsing is the TUT–Penn, which is an application of the PTB format to the Italian language [8]. In this format, the kind and structure of the constituents are the same as in PTB for English, but the inventory of functional tags is enriched with some relations needed to represent e.g. the subject in

[3] Note that the difference between two results is taken to be significant if $p < 0.05$ (see `http://www.cis.upenn.edu/~dbikel/software.html#comparator`)

[4] For the free download of the resource, which is covered by a Creative Commons licence, see `http://www.di.unito.it/~tutreeb`

post-verbal position. Moreover, in order to describe the rich inflectional system of Italian language, the TUT–Penn format adopts a different and richer set of Part of Speech tags with respect to the PTB.

The training data consist in 3,452 sentences corresponding to 94,722 tokens and belong to five different text genres organized in the following subcorpora:

- NEWS and VEDCH, two collections of sentences from Italian newspapers (700 + 400 sentences and 31,299 tokens)
- CODCIV, a collection of sentences from the Italian Civil Law Code (1,100 sentences and 28,045 tokens)
- EUDIR, a collection of declarations of the European Community from the Italian section of the JRC-Acquis Multilingual Parallel Corpus[5] (201 sentences and 7,455 tokens)
- Wikipedia, a collection of sentences from the Italian section of Wikipedia (459 sentences and 14,746 tokens)
- COSTITA, the full collection of sentences of the *Costituzione Italiana* (682 sentences and 13,177 tokens)

3.3 Test Set

The test set is composed by 300 sentences (i.e. 7,325 tokens) balanced around as in the development set: 150 sentences from Civil Law Code, 75 sentences from newspapers and 75 sentences from Wikipedia, which is a new text genre for the constituency track with respect to previous editions.

3.4 Experimental Results

We had only one participant to the constituency track, i.e. Lavelli [18], whose parser adopts the same probabilistic context-free grammar model exploited by the same author in EPT 2009, namely the Berkeley parser.

The evaluation of the participation results for the constituency track is presented in Table 4. It can be observed that the best results have again been achieved on the data

Table 4. EPT 2011: results on the entire test set and on the three subcorpora (CODCIV, NEWSPAPER and WIKIPEDIA)

	Size (sentences)	**LR**	**LP**	F_1
all test set	300	83.42	83.96	83.69
CODCIV	150	87.41	87.14	87.27
NEWSPAPER	75	78.22	76.72	77.46
WIKIPEDIA	75	77.49	79.30	78.38

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

extracted from the Civil Law Code, and the scores for the data from Wikipedia, i.e. the new text genre for this task, are very close to those from newspapers.

Note that, as in the previous editions, the results on the test set were evaluated taking into account punctuation.

4 Discussion

Observing the EPT experience during all the six years, without doubt we can see a trend of significant improvement of the scores for constituency parsing, but this can be ascribed to several factors.

First, this can be motivated by the selection of applied algorithms, which have been also made progressively more adequate and tuned for the reference language and for the data sets. Nevertheless, because of the relatively scarce participation to the constituency parsing contests (which unfortunately never consisted in more than two teams), we have quite limited evidence e.g. about the adequacy to Italian of constituency parsing approaches.

Second, the improvement of results is also determined by the availability of data sets improved with respect to both quality and size. Concerning in particular the size of the data sets, the data available for training today is almost the double of the amount available in 2007 (as shown in Table 5). And there is a corresponding improvement of the performance (in terms of best F_1 score) of about 20%.

To investigate the influence of the treebank size on performance, we carried on further experiments.

In the first experiment we have exploited a subset of the WSJ treebank of a size comparable with that of TUT, i.e. the sections 02 and 03 (consisting of 2,416 sentences). The results in Table 6 show that the performance of the parser on the two treebanks is very similar.

Moreover, we have performed a set of experiments to draw the learning curve and assess the influence of the training set size on the performance. We randomized the training set and selected three subsets containing 50%, 75% and 90% of the sentences of the training set, see (Table 7). For all the experiments, the performance was evaluated on the original test set. We tested the statistical significance of the difference in performance between the results obtained using the entire training set and those exploiting only 90% of it. The test shows that the difference is not significant. This result needs to be further investigated through other experiments, but it suggests that the treebank has currently reached an adequate size.

Table 5. Constituency parsing: evaluation in the three editions and training data set size

year	**training tokens**	**best LR**	**best LP**	**best F_1**
2007	53,656	70.81	65.36	67.97
2009	58,609	80.02	77.48	78.73
2011	94,722	83.42	83.96	83.69

Table 6. Performance of the Berkeley parser using as training set a subset of the WSJ consisting of sections 02 and 03 and as test set section 21

	LR	**LP**	F_1
WSJ	83.45	82.17	82.80

Table 7. Evaluation of the improvement of scores versus increase of data size for training

portion of the training set	**best LR**	**best LP**	**best** F_1
50%	81.11	80.73	80.92
75%	81.56	81.34	81.45
90%	83.73	83.66	83.70

It is moreover interesting to note the variation of the performance with respect to text genres, which is around the same in all the EPT editions and is confirmed also in the dependency parsing track [6]. The language of the Civil Law Code shows in all the analyzed EVALITA experiences scores higher than for the other text genres (e.g. newspaper). For instance, if we observe the F_1 of the best scoring systems, we see a variation of 8.37 points between legal and newspaper text genre in 2007, 4.45 in 2009 and 11.05 in 2011.

This variation has to be motivated by carefully taking into account not only the features of each text genre, but also the annotation applied to the data.

As far as the features of the legal language are concerned, it should be observed that the Civil Law Code corpus is featured by a little bit higher frequency (around 2%) of null elements, punctuation marks and non-projective structures with respect to the newspaper corpus. The average sentence length is around the same for both these corpora, but the distribution of lengths strongly vary in legal and newspaper texts: sentences shorter than 7 words represent more than 12.3% in the Civil Law Code versus 4.4% in Newspaper. In spite of this, three quarters of the legal corpus is composed by sentences longer than 10 words, while around 43% of the sentences of Newspaper corpus are featured by this same length.

Nevertheless, we underline that there are important differences among the texts belonging to the legal domain itself. For instance, the experiments reported in [15] and in [29] demonstrate that the legal texts annotated according to the (dependency-based) ISST–TANL scheme and extracted from European Commission, Italian State and Piedmont Region laws are harder to parse with respect to the texts extracted from newspapers and then probably with respect also to those of the Civil Law Code. A more detailed analysis of the features of the legal language, a comparison between different kinds of legal language and the investigation of the influence of the applied annotation and representation paradigm are beyond the scope of this paper, but can the object of future work.

Finally, even if in 2011 they show a substantial improvement, the results for constituency parsing remain significantly lower than those achieved by applying dependency-based approaches (see [6]).

The limited amount of data provided by the EPT editions, together with the scarce availability of published experiments about the application of constituency parsing to Italian, make difficult to formulate reliable hypotheses about this language. Nevertheless, several evidences can be extracted from the experiments performed on the languages belonging, like Italian, to the family of morphologically rich languages [28,26]. In this kind of languages morphological differences of word forms express information concerning the arrangement of words into syntactic units or cues to syntactic relations. This leads to a very large number of possible word forms, but also to free constituent order, discontinuity and pro–drop. On the one hand, where words are featured by a larger variety of inflected forms, they can more often freely change their position with respect to languages which rely on rigid phrase structure, like English and Chinese. On the other hand, rich morphological information in the Verbal head of clauses can predispose to omission of overt subjects, i.e. pro–drop. A wide literature shows that the most morphologically rich languages share scores of standard metrics for statistical parsing significantly lower than English, and the dependency paradigm has been demonstrated as more suitable for such kind of languages with respect to the constituency one.

5 Conclusions

The EVALITA Parsing Tasks held during the last six years have been devoted to the definition and extension of the state-of-the-art for Italian parsing. Taking into account all the events of this evaluation campaign and mainly focussing on the last one held in 2011, the paper especially describes the evolution of the constituency parsing for a language which can be considered under various respects as belonging to the family of morphologically rich languages. It describes therefore the data sets used both in the training and evaluation, the applied parsing systems and the results they achieved on the basis of these data in all the editions of the contest. Finally, a brief discussion about the results is presented.

References

1. Bikel, D.M.: Intricacies of Collins' parsing model. Computational Linguistics 30(4), 479–511 (2004)
2. Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., Strzalkowski, T.: A procedure for quantitatively comparing the syntactic coverage of English. In: Proceedings of the Speech and Natural Language Workshop, Pacific Grove, CA, pp. 306–311 (1991)
3. Bod, R.: A computational model of language performance: Data oriented parsing. In: Proceedings of the 14th International Conference on Computational linguistics (CoLing 1992), Nantes, France, pp. 855–859 (1992)
4. Bos, J., Bosco, C., Mazzei, A.: Converting a dependency treebank to a Categorial Grammar treebank for Italian. In: Proceedings of the 8th Workshop on Treebanks and Linguistic Theories (TLT 2008), Milan, Italy, pp. 27–38 (2009)

5. Bosco, C.: Multiple-step treebank conversion: from dependency to Penn format. In: Proceedings of the Linguistic Annotation Workshop (LAW) 2007, Prague, pp. 164–167 (2007)
6. Bosco, C., Mazzei, A.: The EVALITA Dependency Parsing Task: From 2007 to 2011. In: Magnini, B., Cutugno, F., Falcone, M., Pianta, E. (eds.) EVALITA 2012. LNCS(LNAI), vol. 7689, pp. 1–12. Springer, Heidelberg (2012)
7. Bosco, C., Mazzei, A., Lombardo, V.: Evalita Parsing Task: an analysis of the first parsing system contest for Italian. Intelligenza Artificiale 2(IV), 30–33 (2007)
8. Bosco, C., Mazzei, A., Lombardo, V.: Evalita 2009 Parsing Task: constituency parsers and the Penn format for Italian. In: Proceedings of Evalita 2009, Reggio Emilia, Italy (2009)
9. Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A.: Evalita 2009 Parsing Task: comparing dependency parsers and treebanks. In: Proceedings of Evalita 2009, Reggio Emilia, Italy (2009)
10. 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 the Sixth International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco, pp. 2066–2073 (2008)
11. Collins, M.: Three generative, lexicalized models for statistical parsing. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, pp. 16–23 (1997)
12. Corazza, A., Lavelli, A., Satta, G.: Phrase-based statistical parsing. Intelligenza Artificiale 2(IV), 38–39 (2007)
13. Corazza, A., Lavelli, A., Satta, G., Zanoli, R.: Analyzing an Italian treebank with state-of-the-art statistical parsers. In: Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004), Tübingen, Germany, pp. 39–50 (2004)
14. McClosky, D., Charniak, E., Johnson, M.: When is self-training effective for parsing? In: Proceedings of the 22th International Conference on Computational linguistics (CoLing 2008), pp. 561–568 (2008)
15. Dell'Orletta, F., Marchi, S., Montemagni, S., Venturi, G., Agnoloni, T., Francesconi, E.: Domain adaptation for dependency parsing at Evalita 2011. In: Working Notes of EVALITA 2011 (2012)
16. Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: Advances in Neural Information Processing Systems 15 (NIPS 2002), Vancouver, Canada (2002)
17. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp. 423–430 (2003)
18. Lavelli, A.: The Berkeley parser at the EVALITA 2011 Constituency Parsing Task. In: Working Notes of EVALITA 2011 (2011)
19. Lavelli, A., Corazza, A.: The Berkeley Parser at the EVALITA 2009 constituency parsing task. In: Proceedings of the EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian (2009)
20. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313–330 (1993)
21. 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 the Italian Syntactic-Semantic Treebank. In: Abeillé, A. (ed.) Building and Using Syntactically Annotated Corpora, pp. 189–210. Kluwer, Dordrecht (2003)

22. Petrov, S., Klein, D.: Improved inference for unlexicalized parsing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York, pp. 404–411 (2007)
23. Petrov, S., Klein, D.: Discriminative log-linear grammars with latent variables. In: Advances in Neural Information Processing Systems 20 (NIPS 20), Vancouver, Canada, pp. 1153–1160 (2008)
24. Pianta, E.: Recovering from failure with the GraFo left corner parser. Intelligenza Artificiale 2(IV), 34–35 (2007)
25. Sangati, F.: A simple DOP model for constituency parsing of Italian sentences. In: Proceedings of the EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian (October 2009)
26. Seddah, D., Tsarfaty, R., Foster, J. (eds.): Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages. Association for Computational Linguistics, Dublin, Ireland (October 2011)
27. Tonelli, S., Delmonte, R., Bristot, A.: Enriching the Venice Italian Treebank with dependency and grammatical relations. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 1920–1924 (2008)
28. Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., Tounsi, L.: Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, Los Angeles, CA, USA, pp. 1–12 (2010)
29. Venturi, G.: Design and development of TEMIS: a syntactically and semantically annotated corpus of italian legislative texts. In: Proceedings of the Workshop on Semantic Processing of Legal Texts (SPLeT 2012), pp. 1–12 (2012)