Annotation Schema Oriented Validation for Dependency Parsing Evaluation

Cristina Bosco
Università di Torino
Dipartimento di Informatica
E-mail: bosco@di.unito.it

Alberto Lavelli
Fondazione Bruno Kessler, Trento
HLT Research Unit
E-mail: lavelli@fbk.eu

Abstract

Recent studies demonstrate the effects of various factors on the scores of parsing evaluation metrics and show the limits of evaluation centered on single test sets or treebank annotation. The main aim of this work is at contributing to the debate about the evaluation of treebanks and parsers, and, in particular, about the influence on scores of the design of the annotation schema applied in the data. Therefore the paper focusses on a dependency-based treebank whose annotation schema includes relations that can be set at different degrees of specificity, and quantitatively describes how the parser performance is affected when processing a selection of hard to parse constructions taken from a recent evaluation campaign for Italian parsing.

1 Introduction

In most cases parsers are evaluated against gold standard test data and mainly referring to particular resources, see e.g. the recent shared tasks for multilingual parsers [29, 9] and single language parsers (e.g. [17] for German, [4, 5, 6] for Italian, [30] and http://atoll.inria.fr/passage/eval2.en.html for French). Nevertheless, this kind of evaluation has been criticized under various respects, which are strictly related to the nature of treebanks, showing that scores obtained on a single set of data can be significantly limited by a variety of factors among which the following:

- The domains and genres of texts [14].
- The paradigm and metrics used for the evaluation. Starting from [23, 10], PARSEVAL metrics have been criticized for not representing the real quality of parsing, since they neither weight results nor differentiate between linguistically more or less severe errors [31]. By contrast, dependency–based evaluations and metrics are appreciated since they mainly refer to the encoding of predicate argument structures, a crucial factor for several NLP tasks.

- The language, whose characteristics can influence parsing performance; e.g. a long-standing unresolved issue in parsing literature is whether parsing less-configurational languages is harder than parsing English [16], standing the irreproducibility of the results obtained on the Penn Treebank on other languages.

- The frequency in the test data of constructions which are hard to parse, such as coordination or PP-attachment, where the performance of parsers is much lower than the overall score [32].

- The annotation schema on which the evaluation is based, since treebank annotation schemes may have a strong impact on parsing results [31, 16, 24] and cross–framework evaluation is a complex and unresolved issue. Conversions[1], applied for enabling cross-framework comparisons, are difficult [26, 2, 12] and often decrease the reliability of data introducing errors.

The scenario of parsing evaluation is further complicated by the interrelation of these factors. For instance, [8] demonstrated the influence of annotation schemes on some evaluation metrics, and various scholars often considered differences in schemes applied to different languages among the major causes of the different parsing performance for such languages.

New methods have been proposed to increase the reliability of parsing evaluation, e.g. [18, 32, 33]. They are language-oriented and, at least in principle, framework-independent, and have the advantage of annealing the effects of most of the factors that limit the reliability of evaluations based on test sets. Since these methods focus on specific constructions and explicitly take into account the features of the analyzed language, they can provide additional means to assess parser performance on a linguistic level and enable us to develop more informed comparisons of results across different annotation schemes and languages.

In this paper, we present the application of a similar approach to the dependency parsing of Italian. The main aim of this work is at contributing to the debate about the evaluation of parsing results centered on treebanks, to go beyond the simple assessment of results by presenting evidences about the influence on scores of some of the above mentioned factors, i.e. the language, the frequency of hard to parse constructions, and mainly the design of the annotation schema.
Italian has been selected as a case study because the results of the Evalita'09 Parsing Task (henceforth EPT) [6] have shown that performance is now very close to the scores known for English[2] (top systems LAS are 88.73 and 88.67). They were obtained in EPT by systems based on different assumptions, e.g. rule-based, like TULE [22], and statistical parsers, such as DeSR [1] and MaltParser [28, 20][3],

[1] If the evaluation of a parser P is based on a format F, which is different from that of the output of P, a conversion to F is applied to the output of P and/or to the data used for the training of P.

[2] LAS 89,61 [29] is the best result for English dependency parsing, whilst LAS 86.94 [21] is that previously published for Italian in Evalita'07 Parsing Task [4].

[3] See [29] for the results of DeSR and MaltParser in the CoNLL'07 multi-lingual shared task.

evaluated against two different annotation formats, i.e. those of TUT (Turin University Treebank) and ISST-TANL (Italian Syntactic Semantic Treebank [25]).
Our analysis is based on TUT, which allowed for the best results in EPT, and the MaltParser, a statistical parser tested on different languages and treebanks that participated to EPT with results among the best ones. In particular, we will show experiments focussed on a set of Italian hard to parse constructions and three settings of the annotation schema of TUT, which vary with respect to the amount of underlying linguistic information.
The paper is structured as follows: Section 2 gives an overview of the main features of the TUT treebank and its settings. Section 3 describes the methodology and the experiments. Finally, in Section 4 we discuss the results.

2 TUT: data and annotations

TUT[4] is the Italian treebank developed by the Natural Language Processing group of the Department of Computer Science of the University of Turin. The treebank currently includes 2,400 sentences (72,149 annotated tokens in TUT native format) organized in three subcorpora that represent different text genres: newspapers (1,100 sentences), Italian Civil Law Code (1,100 sentences), and 200 sentences from the Italian section of the JRC-Acquis Multilingual Parallel Corpus, a collection of declarations of the European Community shared with the evaluation campaign for parsing French Passage[5].
Even if smaller than other Italian treebanks (i.e. ISST-TANL and the Venice Italian Treebank, VIT, [13]), TUT not only has allowed for best results in EPT, but also makes possible theoretical and applicative comparisons among different formalisms, since TUT is available with annotation formats based on different approaches, e.g. CCG-TUT, a treebank of Combinatory Categorial Grammar derivations [3], and TUT-Penn, a constituency-based treebank [5].

The native annotation scheme of TUT features a pure dependency format centered upon the notion of argument structure, which applies the major principles of Hudson's *word grammar* [15]. This is mirrored, for instance, in the annotation of determiners and prepositions as complementizers of nouns or verbs (see figures below). In fact, since the classes of determiners and prepositions include elements[6] which often are used without complements and can occur alone (like possessive and deictic adjectives or numerals used as pronouns, or prepositions like 'before' and 'after'), all the members of these classes play the same head role when occur with or without nouns or verbs. Moreover, the annotation schema includes null elements to deal with non-projective structures, long distance dependencies, equi

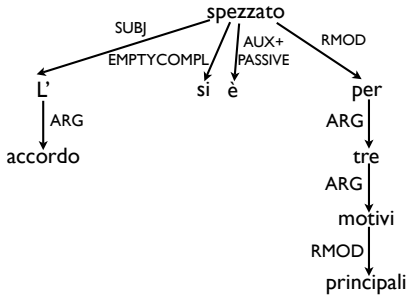
[4] http://www.di.unito.it/∼tutreeb

[5] See http://langtech.jrc.it/JRC-Acquis.html and http://atoll.inria.fr/passage/index.en.html respectively for the JRC-Acquis corpus and Passage.

[6] According to the word grammar, many words qualify as prepositions or determiners which traditional grammar would have classified as adverbs or subordinating conjunctions.

phenomena, pro drop and elliptical structures.

But the most typical feature of the treebank is that it exploits a rich set of grammatical relations designed to represent a variety of linguistic information according to three different perspectives, i.e. morphology, functional syntax and semantics. The main idea is that a single layer, the one describing the relations between words, can represent linguistic knowledge that is proximate to semantics and underlies syntax and morphology, which seems to be unavoidable for efficient processing of human language, i.e. the predicate argument structure of events and states. Therefore, each relation label can in principle include three components, i.e. morpho-syntactic, functional-syntactic and syntactic-semantic, but can be made more or less specialized, including from only one (i.e. the functional-syntactic) to three of them. For instance, the relation used for the annotation of locative prepositional modifiers, i.e. PREP-RMOD-LOC (which includes all the three components), can be reduced to PREP-RMOD (which includes only the first two components) or to RMOD (which includes only the functional-syntactic component).
This works as a means for the annotators to represent different layers of confidence in the annotation, but can also be applied to increase the comparability of TUT with other existing resources, by exploiting the amount of linguistic information more adequate for the comparison, e.g. in terms of number of relations, as happened in EPT. Since in different settings several relations can be merged in a single one (e.g. PREP-RMOD-TIME and PREP-RMOD-LOC are merged in RMOD), each setting includes a different number of relations: the setting based on the single functional-syntactic component (henceforth *1-Comp*) includes 72 relations, the one based on morpho-syntactic and functional-syntactic components (*2-Comp*) 140, and the one based on all the three components (*3-Comp*) 323. In figure 1 the tree (a) for the



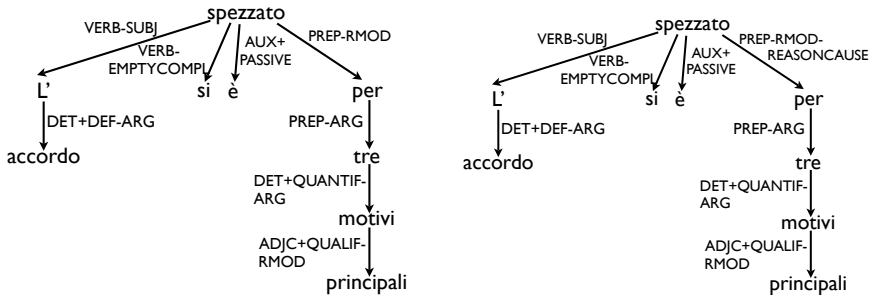
(a)

Figure 1: Sentence ALB–356 in 1–Comp setting, like in EPT.

sentence ALB-356 from TUT corpus, i.e. *"L'accordo si è spezzato per tre motivi principali"* (The agreement has been broken for three main motivations)[7], shows

[7]English translations of the Italian examples are literal and so may appear awkward in English.

the features of the annotation schema. In particular, we see the role of complementizer played by determiners (i.e. the article *"L'"* (The) and the numeral *"tre"* (three)) and prepositions (i.e. *"per* (for)), and the selection of the main verb as head of the structure instead of the auxiliary. If we compare the tree (a) (in figure 1), with the trees (b) and (c) (in figure 2.b and .c), we see also the variation of relations in the three settings for the same sentence. For instance, the relation between *spezzato* (broken) and the prepositional modifier *per tre motivi principali* (for three main motivations), or the argument articles that are ARG in 1-Comp and DET+DEF-ARG (i.e. ARGument of a DEFinite DETerminer) in the other settings. The latter case is an example of relation that does not include semantic information and therefore remains the same in 2- and 3-Comp settings.



(b) (c)

Figure 2: Sentence ALB-356 in: (b) 2-Comp setting; (c) 3-Comp setting.

3 Development of the methodology

The approach we propose is language oriented and construction-based, but it differs e.g. both from those in [18] and in [32]. First, by contrast with [18], we follow a pure dependency approach, i.e. the treebank implements a pure dependency annotation, and our analysis is mainly focused on grammatical relations. Second, the selection of the hard to parse phenomena for our experiments is motivated not only by linguistic and applicative considerations, as in related works, but also driven by the performance of different parsers. Third, the analysis is based on three different annotation schemes which are however extracted from the same treebank rather than derived from different sources. Last but not least, our reference language is Italian, which is considered as relatively free word order like German, but less studied until now than Czech or German.

Assuming that most of the parsing errors are related to some specific relation and construction, like in [18, 32], we begin our analysis by identifying cases that can be considered as hard to parse for Italian. For the results of each of the six

participant parsers on the EPT test set[8] we compute precision and recall[9] for each type of grammatical relations. To further assess the results, we perform the same kind of evaluation on the three relation settings running a 10-fold cross validation on the entire treebank with MaltParser. After identifying the hard to parse relations, we develop a comparative analysis of the behavior of MaltParser in such cases.

3.1 Selecting phenomena and features

Observing the average score of the six parsers which participated in EPT we can identify the following hard to parse constructions:

- the predicative complement of the object, i.e. PREDCOMPL+OBJ (which occurs 141 times in the full treebank, i.e. 0.19%). For instance, in *"Il parlamentare si è detto **favorevole** ad una maggiore apertura delle frontiere ai rifugiati politici."*(The parliamentarian itself has said **in favour** of a major opening of frontiers to the political refugees.)

- the indirect object, i.e. INDOBJ (which occurs 325 times, i.e. 0.45%). For instance, in *"Noi non permetteremo **a nessuno** di imbrogliarci."* (We will not allow **to anybody** to cheat us.)

- various relations involved in coordinative structures that represent comparisons (e.g. COORDANTEC+COMPAR and COORD+COMPAR (which occurs 64 times, i.e. 0,08%), like in *"Usa un test **meno** raffinato **di quello tradizionale**."* ([He] exploits a test **less** refined **than the traditional one**.)).

- various relations for the annotation of punctuation, in particular SEPARATOR, OPEN+PARENTHETICAL (which occurs 1,116 times, i.e. 1.5%) and CLOSE +PARENTHETICAL (which occurs 1097 times, i.e. 1.5%)). For instance, SEPARATOR (which occurs 1,952 times, i.e. 2.7%) is used in cases where commas play the role of disambiguating marks and an ambiguity could result if the marks were not there [19], e.g. in *"Quando il meccanismo si inceppa, è il disastro."* (When the mechanism hinds itself, is a disaster). OPEN+/CLOSE+PARENTHETICAL are instead used for the annotation of paired punctuation that marks the parenthetical in *"Pochi quotidiani **,** solo quelli inglesi**,** saranno oggi in vendita."* (Few newspapers**,** only those English**,** will be today on sale.).

Since not all the grammatical relations of 1-Comp occur in the test set, the above list cannot be in principle considered as representative of how hard to parse is the treebank (and the Italian language). A 10-fold cross validation performed on the whole TUT with the 1-Comp setting shows that other low-scored relations exist, but since they appear with a very low frequency we did not include them in our

[8]The EPT test set included 240 sentences (5,287 tokens) balanced alike to those of the treebank used for training: 100 sentences (1,782 tokens) from newspapers, 100 (2,293 tokens) from Civil Law Code and 40 (1,212 tokens) from the Passage/JRC-Acquis corpus.

[9]The evaluation has been performed by using the MaltEval tools [27].

experiments. This shows however that the test set, even if it shows the same balancement of TUT, does not represent at best the treebank in terms of relations and constructions. Moreover, a comparison with ISST-TANL, based on the EPT results and developed in [6] and [7], shows that similar relations, in particular coordination and punctuation, are low-scored also in this other resource, notwithstanding the different underlying annotation schema where, e.g. it is the determiner which depends on the noun. Nevertheless this comparison is of limited interested, since in ISST-TANL the annotation of punctuation is far less fine-grained than in TUT.

3.2 Comparing the test set and the whole treebank

The comparisons of this section exploit the relation settings of TUT, and are oriented to the assessment of the influence of the annotation schema design on parsing results. They show that the evaluation has to be weighted observing at least the distribution and kind of hard to parse constructions and the degree of difficulty of hard to parse constructions, which can vary in the test set and in the whole treebank.

First of all, we test the hypothesis that the test set is an aggregate over a highly skewed distribution of relations and constructions, where the frequency of hard to parse phenomena can be different from that of the whole treebank. The application of MaltParser on all the treebank with the 1-Comp setting, like in the EPT test set, exploiting a 10-fold cross validation strategy shows that this hypothesis is correct, since the performance significantly varies when the parser is applied to the EPT test set rather than to all the treebank, i.e. from LAS 86.5 and UAS 90.96, in the test set [20], to LAS 83.24 e UAS 87.69 in all TUT[10]. This suggests that the distribution of hard to parse phenomena is not the same in both cases.

In order to test the hypothesis that the degree of difficulty of the same hard to parse constructions can vary in the test set with respect to the treebank, we first analyze the performance of MaltParser on all TUT with the 3 settings, and, second, we analyze the variation of precision and recall for each hard to parse case according to the three settings. As table 1 shows, the performance in terms of UAS is

	1-Comp	**2-Comp**	**3-Comp**
LAS	83.24	82.56	78.77
UAS	87.69	87.60	87.20

Table 1: MaltParser scores in 10-fold cross validation over the whole treebank.

not significantly influenced by the different settings, since the difference concerns the relation labels rather than the tree structures. Instead, LAS decreases when the number of relations is enlarged in settings that should be more informative, going from 72 (1-Comp), to 140 (2-Comp), to 323 relations (3-Comp). The larger amount of relations occurring a small number of times in 2- and 3-Comp (with

[10]This is only partially explained by the sentence length, which is lower than 40 words only in the test set, and by the smaller size of the training set for the 10-fold cross validation.

respect to 1-Comp) increases the sparseness of relations and negatively influences the performance. Also the stability across all settings of the performance only on more frequent relations, further supports this conclusion.

Now we focus on single hard to parse relations in order to show the variation of parser performance in the three settings. Tables 2, 3 and 4 show that the parser behavior varies in different way for different relations and sometimes following a different trend with respect to the results on all the treebank. For instance, for

	EPT	**1-Comp**	**2-Comp**	**3-Comp**
prec	50.00	89.66	83.33	86.21
rec	25.00	54.17	52.08	52.08

Table 2: MaltParser scores for COORD+COMPAR with different settings.

COORD+COMPAR (table 2) the best performance is in 1-Comp and the worst in the EPT test set. For PREDCOMPL+OBJ (table 3), instead, the best performance

	EPT	**1-Comp**	**2-Comp**	**3-Comp**
prec	50	57.81	60.00	61.16
rec	40	52.48	53.19	52.48

Table 3: MaltParser scores for (VERB-)PREDCOMPL+OBJ with different settings.

is in 3-Comp and the worst in the EPT test set. Therefore, in this case there is a contrast with the general trend shown in table 1, since the results are significantly better when the relation labels include the morphological component.

	EPT	**1-Comp**	**2-Comp**	**3-Comp**
prec	68.97	57.00	55.96	48.26
rec	58.82	52.35	50.49	63.19

Table 4: MaltParser scores for (VERB-)INDOBJ with different settings.

For what concerns instead punctuation, we observe that it is not always considered when performing evaluation. As we have seen before, in our evaluation punctuation is instead taken into account, but the related relations are among the low-scored ones. For instance, SEPARATOR (see section 3.1) is in the set of the 9 most frequent relations[11] (in 1-Comp setting in both all the treebank and the test set) and occurs around 2,000 times in the full treebank, but it is the one scoring the lower

[11]The ten most frequent relations in all the 1-Comp treebank (with respect to 72,149 annotated tokens) are ARG (30.3%), RMOD (19.2%), OBJ (4.5%), SUBJ (3.9%), END (3.3%), TOP (3.2%), COORD2ND+BASE (3.1%), COORD+BASE (3.1%), SEPARATOR (2.7%), INDCOMPL (1.9%).

in precision and recall of this set for all the parsers participating to EPT. Therefore, in the perspective of a comparison with other evaluations and resources, it would be useful to see how our results vary when punctuation is excluded, as in table 5. The UAS and LAS scores of MaltParser are in all TUT settings 3.5 points higher

	1-Comp	**2-Comp**	**3-Comp**
LAS Punct	83.24	82.56	78.77
LAS noPunct	86.78	86.02	81.88
UAS Punct	87.69	87.60	87.20
UAS noPunct	91.10	91.01	90.70

Table 5: MaltParser scores on 1-, 2- and 3-Comp TUT with and without punctuation, in 10-fold cross validation.

when the punctuation is not taken into account. As for ISST-TANL, the experiments show that the difference in performance when considering or not considering punctuation is between 1.76 and 2.50 according to different parser parameters. This lower difference can be explained by the different annotation of punctuation, less fine-grained in ISST-TANL where a single relation PUNC is used. This means that some improvement in parsing can be obtained by more adequate processing of punctuation, as said e.g. in [11], and/or by more adequate annotation of it. In fact punctuation is often relevant from a linguistic point of view as a marker of clause or phrase boundaries, thus if a parser does not predict it correctly, it can lead to incorrect parses and lower scores when evaluated against a resource that annotates punctuation.

As for the comparison with other languages, we have seen that part of the hard to parse phenomena for Italian are included also in the test suites proposed for German, e.g. forms of coordination. But, since the lists presented in [18] and in [32] are mainly linguistically motivated and not quantitatively determined, we cannot go beyond this observation and further extend the comparison here.

For what concerns single phenomena, following the idea that parsing can be made more or less hard by the availability of different amount of linguistic information, we have seen that different effects can be caused by the use of more or less informative grammatical relations. The results demonstrate, in particular, that the evaluation based on the test set is limited with respect to the distribution and kind of hard to parse constructions, which in the test set and in the treebank can be different, and the degree of difficulty of hard to parse constructions, which in the test set and in the treebank can be not the same.

4 Conclusions and future work

Most parser evaluations are based on single resources, but the design and features of the treebank used for testing can strongly influence the results.

This paper presents issues for the development and application to Italian parsing of a methodology for the validation of the evaluation of parsing results. Starting from the results of an evaluation campaign for Italian parsing, i.e. EPT, it provides evidence about the skewedness of the test set of this contest. The experiments presented confirm the hypothesis that evaluations based on test sets and single resources present several shortcomings. They demonstrate, in particular, that the validity of an evaluation based on a test set and a single resource is limited with respect to the distribution and kind of hard to parse constructions, which in the test set and in the treebank can be different, and with respect to the degree of difficulty of hard to parse constructions, which in the test set and in the treebank can vary.
A variety of directions for future research is raised by the present work that go beyond the simple assessment of the results to give suggestions for both treebank design and the development of more informed evaluation methodologies. Among them, in particular, a deeper analysis of the presented data and results by trying new experiments based also on parsers that apply different approaches, e.g. TULE; the comparison with other existing resources and annotation schemes, and last but not least the comparison with other languages.

References

[1] Attardi, G., Dell'Orletta, F., Simi, M. and Turian, J. (2009) Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of Evalita'09*, Reggio Emilia.

[2] Bick, E. (2006) Turning a dependency treebank into a PSG-style constituent treebank. In *Proceedings of LREC'06*, pp. 1961–1964, Genova.

[3] Bos, J., Bosco, C. and Mazzei, A. (2009) Converting a Dependency Treebank to a Categorial Grammar Treebank for Italian. In *Proceedings of TLT-8*, pp. 27–38, Milano.

[4] Bosco, C., Mazzei, A. and Lombardo, V. (2007) Evalita Parsing Task: an analysis of the first parsing system contest for Italian. In *Intelligenza Artificiale*, Vol. 2, pp. 30–33.

[5] Bosco, C., Mazzei, A. and Lombardo, V. (2009) Evalita Parsing Task 2009: constituency parsing and a Penn format for Italian. In *Proceedings of Evalita'09*, Reggio Emilia.

[6] Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F. and Lenci, A. (2009) Evalita'09 Parsing Task: comparing dependency parsers and treebanks. In *Proceedings of Evalita'09*, Reggio Emilia.

[7] Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A., Lesmo, L., Attardi, G., Simi, M., Lavelli, A., Hall, J., Nilsson,

J. and Nivre, J. (2010) Comparing the Influence of Different Treebank Annotations on Dependency Parsing. In *Proceedings of LREC'10*, pp. 1794–1801, Malta.

[8] Boyd, A. and Meurers, D. (2008) Revisiting the impact of different annotation schemes on PCFG parsing: a grammatical dependency evaluation. In *Proceedings of ACL'08: HLT Workshop on parsing German*, pp. 24-32, Columbus Ohio.

[9] Buchholz, S. and Marsi, E. (2007) CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL-X*, pp. 149-164, New York.

[10] Carroll, J., Briscoe, T. and Sanfilippo, A. (1998) Parser evaluation: a survey and a new proposal. In *Proceedings of LREC'98*, pp. 447-454, Granada.

[11] Cheung, J. C.K. and Penn, G. (2009) Topological field parsing of German. In *Proceedings of ACL-IJCNLP'09*, pp. 64-72, Singapore.

[12] Clark, S. and Curran, J. R. (2007) Formalism-independent parser evaluation with CCG and DepBank. In *Proceedings of ACL'07*, pp. 248-255, Prague.

[13] Delmonte, R. (2008) *Strutture sintattiche dall'analisi computazionale di corpora di italiano.* Milano: Franco Angeli.

[14] Gildea, D. (2001) Corpus variation and parser performance. In *Proceedings of EMNLP'01*, pp. 167-202, Pittsburg.

[15] Hudson, R. (1984) *Word grammar*, Oxford and New York: Basil Blackwell.

[16] Kübler, S., Hinrichs, H. and Maier, W. (2006) Is it really that difficult to parse German?. In *Proceedings of EMNLP'06*, pp. 111-119, Sydney.

[17] Kübler, S. (2008) The PaGe 2008 shared task on parsing German. In *Proceedings of ACL Workshop on parsing German*, pp. 55-63, Columbus Ohio.

[18] Kübler, S., Rehbein, I. and van Genabith, J. (2009) TePaCoC a corpus for testing parser performance on complex German grammatical constructions. In *Proceedings of TLT-7*, pp. 15–28, Groningen: The Netherlands.

[19] Jones, B. E. M. (1994) Exploring the role of punctuation in parsing natural text. In *Proceedings of COLING'94*, pp. 421-425, Kyoto.

[20] Lavelli, A., Hall, J., Nilsson, J. and Nivre, J. (2009) MaltParser at the Evalita 2009 Dependency parsing task. In *Proceedings of Evalita'09*, Reggio Emilia.

[21] Lesmo, L. (2007) The rule-based parser of the NLP group of the University of Torino. In *Intelligenza Artificiale*, Vol. 2, pp. 46–47.

[22] Lesmo, L. (2009) The Turin University Parser at Evalita 2009. In *Proceedings of Evalita'09*, Reggio Emilia.

[23] Lin, D. (1995) A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI'95*, pp. 1420-1427, Montreal.

[24] Maier, W. (2006) Annotation schemes and their influence on parsing results. In *Proceedings of COLING-ACL'06 Student Research Workshop*, pp. 19-24, Sydney.

[25] Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Pirrelli, V., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F. and Delmonte, R. (2003) Building the Italian Syntactic-Semantic treebank. In *Building and using Parsed Corpora* A. Abeillè (ed.), pp. 189–210, Dordrecht: Kluwer.

[26] Musillo, G., Sima'an, K. (2002) Towards comparing parsers from different linguistic frameworks. An information theoretic approach. In *Proceedings of Workshop Beyond PARSEVAL - Towards improved evaluation measures for parsing systems at the LREC'02*, pp. 44–51, Las Palmas Canary Islands.

[27] Nilsson, J., Nivre, J. (2008) MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. In *Proceedings of LREC'08*, pp. 161–166, Marrakech.

[28] Nivre, J., Hall, J. and Nilsson, J. (2006) MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of LREC'06*, pp. 2216–2219, Genova.

[29] Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S. and Yuret, D. (2007) The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of EMNLP-CoNLL'07*, pp. 915–932, Prague.

[30] Paroubek, P., Vilnat, A., Loiseau, S., Hamon, O., Francopoulo, G., and Villemonte de la Clergerie, E. (2008) Large scale production of syntactic annotations to move forward. In *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 36–43, Manchester UK.

[31] Rehbein, I. and van Genabith, J. (2007) Treebank annotation schemes and parser evaluation for German. In *Proceedings of EMNLP-CoNLL'07*, pp. 630–639, Prague.

[32] Rimell, L. and Clark, S. and Steedman, M. (2009) Unbounded dependency recovery for parser evaluation. In *Proceedings of EMNLP'09*, pp. 813–821, Singapore.

[33] Tam, W. L., Sato, Y., Miyao, Y. and Tsujii, J. (2008) Parser evaluation across frameworks without format conversion. In *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 29–35, Manchester UK.