# Croatian Dependency Treebank: Recent Development and Initial Experiments

**Daša Berović*, Željko Agić**, Marko Tadić***

*Department of Linguistics
**Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{dberovic, zagic, marko.tadic}@ffzg.hr

**Abstract**

We present the current state of development of the Croatian Dependency Treebank – with special empahsis on adapting the Prague Dependency Treebank formalism to Croatian language specifics – and illustrate its possible applications in an experiment with dependency parsing using MaltParser. The treebank currently contains approximately 2870 sentences, out of which the 2699 sentences and 66930 tokens were used in this experiment. Three linear-time projective algorithms implemented by the MaltParser system – Nivre eager, Nivre standard and stack projective – running on default settings were used in the experiment. The highest performing system, implementing the Nivre eager algorithm, scored (LAS 71.31 UAS 80.93 LA 83.87) within our experiment setup. The results obtained serve as an illustration of treebank's usefulness in natural language processing research and as a baseline for further research in dependency parsing of Croatian.

**Keywords:** dependency treebank, dependency parsing, Croatian language

## 1. Introduction

The Croatian Dependency Treebank (HOBS further in the text, cf. Tadić 2007) is a dependency treebank built along the principles of Functional Generative Description (FGD) (Sgall et al. 1986), a multistratal model of dependency grammar developed for Czech. In a somewhat simplified version, the FGD formalism was further adapted in the Prague Dependency Treebank (PDT) (Hajič et al. 2000) project and applied for the sentence analysis and annotation on the levels of morphology, syntax – in the form of dependency trees with nodes labelled with syntactic functions – and tectogrammatics. The ongoing construction of HOBS closely followed the guidelines set by the PDT, with their simultaneous adaptation to the specifics of the Croatian language. Currently, HOBS consists of approximately 2870 sentences in the form of dependency trees that were manually annotated with syntactic functions using TrEd (Pajas 2000) as the annotation tool. These sentences, encompassing approximately 70.000 tokens, stem from the CroatiaWeekly 100 kw (CW100) corpus that is a part of the newspaper sub-corpus of the Croatian National Corpus (HNK) (Tadić 2000, 2009). The Croatia Weekly sub-corpus was previously sentence-delimited, tokenized, lemmatized and MSD-annotated by linguists. Thus, each of the analyzed sentences contained the manually assigned information on part-of-speech, morphosyntactic category, lemma, dependency and analytical function for each of the wordforms. Such a course of action, i.e. the selection of the corpus, was taken in order to enable the training procedures of various state-of-the-art dependency parsers (cf. Buchholz and Marsi 2006, Nivre et al. 2007) to choose from a wide selection of different features in experiments with stochastic dependency parsing of Croatian texts. Basic stats for HOBS are given in table 1. Sentences in HOBS are annotated according to the PDT annotation manual for the analytical level of annotation, with respect to differing properties of the Croatian language and consulting the Slovene Dependency Treebank (SDT) project (Džeroski et al. 2007). The utilized analytical functions are thus considered to be compatible with those used in PDT.

| Feature | Experiment | Training | Testing |
|---|---|---|---|
| Sentences | 2699 | 2429.10 | 269.90 |
| Tokens | 66930 | 60237.00 | 6693.00 |
| Lemmas | 8995 | 8524.50 | 2295.60 |
| MSD tags | 798 | 779.60 | 410.10 |
| Functions | 80 | 79.00 | 58.30 |

Table 1. Treebank stats

Section 2 present approaches to adapting the PDT syntactic formalism to the process of manual annotation of Croatian sentences for HOBS with respect to Croatian language specifics. Section 3 presents the results of an initial experiment with dependency parsing of Croatian within the framework of transition-based parsing (cf. Nivre and Nilsson 2006) by using the current version of HOBS for language modelling and validation.

## 2. Treebank adaptation

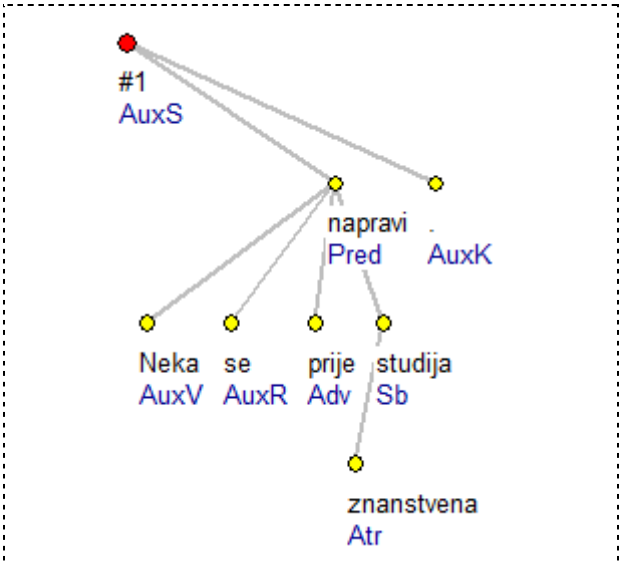
Issues in adapting the PDT formalism to manual annotation of Croatian sentences emerged mainly when annotating predicates, with special emphasis on nominal predicates, somewhat due to the structural differences between the two languages, and somewhat because of approaches to certain issues in the available grammars of Czech and Croatian (cf. Silić and Pranjković 2007). For illustrative purposes, we isolated five different classes of problems with adapting the annotation to specific properties of Croatian with respect to the nominal predicate.

***Problem 1*** In spoken and written Czech negation is connected with the verb itself and imperatives are made with a special suffix. Annotation of particles that compose negation and imperative is thus not provided in the PDT analytical level annotation manual (Hajič et al. 1999, AAL further in the text).

***Solution*** In the annotating system of HOBS the same analytical function (auxiliary verb, AuxV) is assigned to the particle *ne* in the realization of negation and to particles *da* and *neka* in the realization of imperative (figure 1). Analogously, the analytical function AuxV is assigned to negated forms of the auxiliary verb *biti* (en. *to be*), like *nije* or *nisam*. In complex tenses all nodes that are annotated with the analytical function AuxV are directly dependent on the main verb.

***Problem 2*** In PDT, a nominal predicate cannot be expressed with an adverb and a nominal phrase composed of a preposition and a noun. These cases are treated as adverbs and they are annotated with the respective analytical function (Adv).

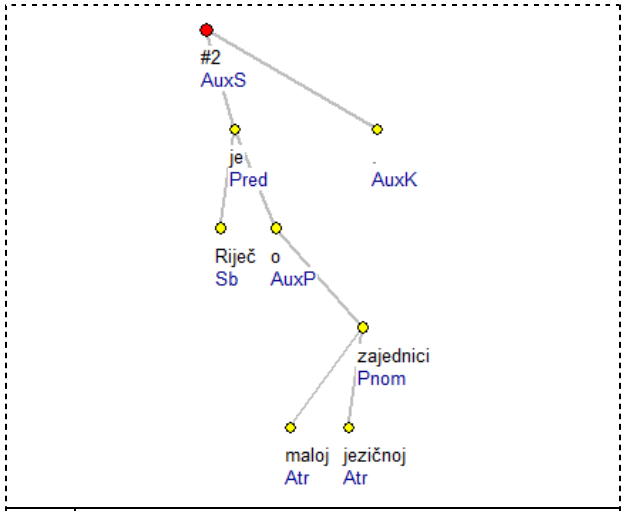
***Solution*** Croatian grammars interpret this case as a part of a nominal predicate, respectively an adjective, so we have annotated them with an analytical function for nominal predicate (Pnom). Furthermore, in Croatian nominal phrases consisting of preposition and noun with an auxiliary verb can also compose a nominal predicate. Accordingly, we propose that in HOBS these cases should be annotated as nominal predicates, unlike in PDT, where they are annotated as adverbs.




| | |
|---|---|
| **hr** | *Neka se prije napravi znanstvena studija.* |
| **en** | *Let's make a scientific study before.* |

Figure 1. Particle *neka* in realization of imperative

Figure 2 shows the annotation of a nominal predicate composed of an auxiliary verb and a nominal phrase consisting of a preposition and a noun. As nominal predicates are also considered as those phrases that are the result of the decomposition of modal verbs to the copula and nominal part – that usually takes the form of an adjective or a nominal phrase consisting of a preposition and a noun (*biti kadar*, *biti u mogućnosti*, en. *to be able to*).

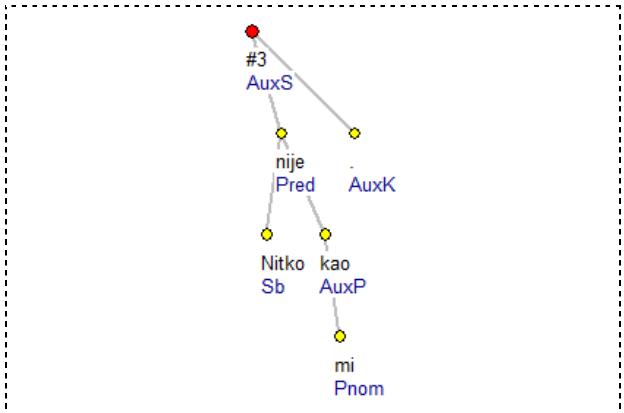



| | |
|---|---|
| **hr** | *Riječ je o maloj jezičnoj zajednici.* |
| **en** | *It is a small language community.* |

Figure 2. Annotation of the nominal predicate composed of an auxiliary verb and a prepositional phrase

***Problem 3*** Silić and Pranjković (2007:290) state that the nominal part of a nominal predicate can be introduced in the sentence by the particle *kao* (en. *like*) – that is not possible in the PDT.

***Solution*** Nominal part of nominal predicate that is introduced by the word *kao* has the same appearance as nominal phrase introduced by the word *poput* or some other preposition, so we decided to treat the word *kao* in nominal predicate as a preposition and annotate it with the corresponding function (figure 3).



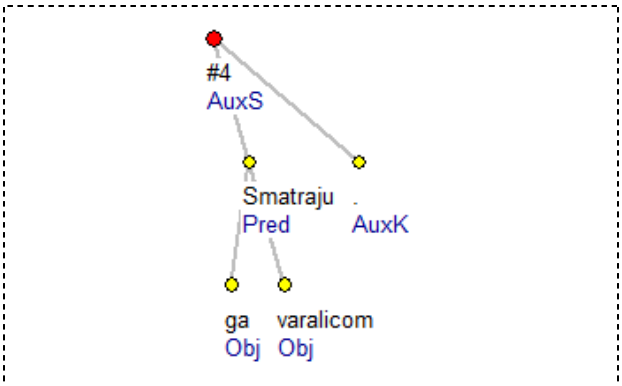

| | |
|---|---|
| **hr** | *Nitko nije kao mi.* |
| **en** | *Nobody is like us.* |

Figure 3. Annotation of the nominal part of nominal predicate that is introduced by the word *kao*

***Problem 4*** In PDT, verbal part of a nominal predicate can be just an auxiliary verb. However, Croatian contains the class of so-called semi-copulative verbs (Silić and Pranjković, 2007:291) that are similar to the auxiliary

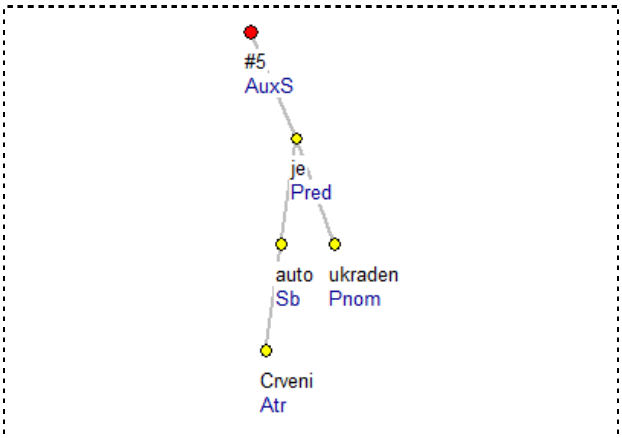
verb *biti*, because they denote that something is attributed to subject or object. Those verbs, just like the verb *biti*, can compose a nominal predicate with a nominal part. In the process of annotation, such verbs should depend on the root of the tree and get the predicate function (Pred), and the nominal part should depend directly on this verb and get assigned as a nominal predicate.

***Solution*** (Silić and Pranjković 2007) provide the list of the semi-copulative verbs, but it is not finite and unambiguous. Besides, this semi-copulative predicate is mentioned just in their grammar, but not in others. Considering that we decided to annotate these cases following the PDT manual. Figure 4 shows the sentence in which semi-copulative verb *smatraju* (en. *consider*) is annotated as ordinary verbal predicate and the noun *varalicom* (en. *fraud*) – that according to (Silić and Pranjković 2007) should be annotated as a nominal part of nominal predicate – is annotated as an object.




| | |
|---|---|
| **hr** | *Smatraju ga varalicom.* |
| **en** | *They consider him a fraud.* |

Figure 4. Annotation of the nominal predicate composed of semi-copulative verb and noun

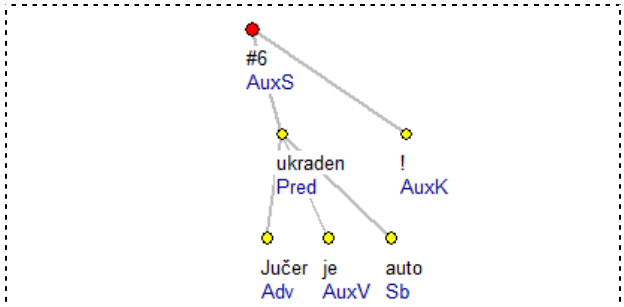



| | |
|---|---|
| **hr** | *Crveni auto je ukraden.* |
| **en** | *The red car was stolen.* |

Figure 5. Annotation of the nominal predicate in the sentence without an adverb

***Problem 5*** Distinction of the nominal predicate and passive verb forms appears as another issue with the annotation of nominal predicates. According to the PDT annotation manual (Hajič et al., 1999:34) the only way to deciding whether these are a nominal predicate composed of an adjective or a passive form realized by a perfect participle is the intuition of the annotator based on sentence context. The annotator should assess whether the focus of the sentence is on the action which is realized or on assigning attributes to the subject of the sentence.

***Solution*** Distinction of nominal predicate and passive forms in HOBS can be made according to the realization of the adverb in the sentence. If the adverb is not realized in the sentence, we conclude that the focus of the sentence is on the subject, so it is a nominal predicate. If there is an adverb that specifies the action of the sentence, we conclude that it is a realization of passive form by a perfect participle. Figure 5 shows the sentence in which there is no adverb, and the adjective *ukraden* (en. *stolen*) specifies the subject phrase *crveni auto* (en. *red car*) – according to that we annotated phrase je *ukraden* as a nominal predicate in which *ukraden* is a nominal part of the nominal predicate. In figure 6, there is an adverb *jučer* (en. *yesterday*), so the phrase je *ukraden* is annotated as a passive verb form in which *ukraden* is annotated as a verbal predicate and je is annotated as an auxiliary verb.




| | |
|---|---|
| **hr** | *Jučer je ukraden auto.* |
| **en** | *Yesterday a car was stolen.* |

Figure 6. Annotation of the passive verb form in the sentence with an adverb

## 3. Parsing

Our illustrational experiment with parsing was basically envisioned as a tenfold cross-validated run of several MaltParser (Nivre et al. 2006) parsing algorithms on HOBS. Thus, the task required pre-processing of the treebank, choosing the parsing algorithms and evaluation metrics and tools.

The treebank was stored in the native TrEd feature structure (FS) format. Using TrEd, we converted the treebank into the Czech sentence tree structure (CSTS) format and then easily translated this format into the CoNLL format by simple regular expressions. Further, we implemented a script for CoNLL token validation and filtered out sentences with invalid tokens. The results of this filtering are given in table 1. Token encoding issues invalidated 171 sentences and thus left a total of 66.930 tokens that were initially available for the experiment. The before-mentioned token encoding issues were mainly

caused by missing escape sequences for decimal numbers within FS-formatted sentences and are currently being corrected. The sentence pool was shuffled and ten pairs of (training set, testing set) samples were selected for the cross-validation. For each of the pairs, the training set consisted of 90% of the treebank sentences and 10% remaining sentences for the testing set. Basic stats for these pairs are also provided in table 1.

Out of various available features of the MaltParser parser generator system, we chose only three algorithms for the experiment. The three are both limited to the set of projective sentences and run in linear time – the Nivre eager, Nivre standard and stack projective algorithm. All the other available algorithms were excluded from this experiment because of simplicity, time constraints and the preliminary nature of these tests. Default settings for all algorithms were selected, i.e. no feature modifications have been made for fine-tuning the algorithms to specific properties of Croatian. Each of the algorithms, or parsers, was first trained on each of the ten training sets, creating 30 different parsing models. The models were then used by MaltParser in parsing mode to parse the respective testing sets. Evaluation was done by using MaltEval (Nilsson and Nivre 2008).

| Metric | Eager | Standard | Stack proj. |
|---|---|---|---|
| LAS | 71.31±0.64 | 68.09±0.81 | 70.60±0.65 |
| UAS | 80.93±0.57 | 81.33±0.75 | 81.51±0.62 |
| LA | 83.87±0.44 | 77.75±0.68 | 82.38±0.53 |

Table 2. Parsing accuracy

| Stage | Eager | Standard | Stack proj. |
|---|---|---|---|
| Training | 56.43±0.77 | 61.29±2.32 | 62.47±1.97 |
| Testing | 10.43±0.21 | 10.33±0.27 | 11.32±0.22 |

Table 3. Execution time (in minutes)

Evaluating the overall accuracy scores of the three systems, given in table 2, and its top-performing system implementing the Nivre eager algorithm, it is apparent – although the scores are somewhat as expected – that room for improvements exists and that improvements are, in fact, required if data-driven dependency parsers derived from HOBS are to be used for further treebank enrichment and information retrieval/extraction tasks. Overall, the Nivre eager algorithm is the top-performer, outperformed only in label attachment by the stack projective algorithm. Comparing these results with the ones obtained for similar languages within the CoNLL 2006 and CoNLL 2007 shared tasks (Buchholz and Marsi 2006, Nivre et al. 2007), these scores for parsing Croatian texts would be grouped with the languages similar in morphosyntactic properties and treebank sizes. It is important to note that the results obtained for Slovene in the 2006 shared task are comparable.

Being that training dependency parsing models is known to be a relatively time-consuming task, we also measured the training and parsing times – they are given in table 2. Both training and testing for the three algorithms was done by using three IBM x3400 servers with Intel Xeon E5405 2 GHz CPU and 2 GB RAM. The training process lasted for approximately an hour for each of the language models, and the parsing for the test samples lasted approximately ten minutes.

## 4. Conclusions and future work

In the paper we presented the current state of the Croatian Dependency Treebank and results of an initial experiment with data-driven transition-based dependency parsing of Croatian by using the Croatian Dependency Treebank and the Malt-Parser parser generator system.

Future research plans are expectedly extensive. The treebank requires both enlargement and enhancement and extensive efforts are currently underway with respect to these goals. Regarding dependency parsing of Croatian by using HOBS, we plan to undergo various research directions in order to increase overall parsing accuracy. Firstly, we shall investigate the performance of other state-of-the-art data-driven dependency parsers such as DeSR (Attardi et al. 2007), MST (McDonald et al. 2006) and IDP (Titov and Henderson 2007). Secondly, fine-tuning of all the available parameters for these and the MaltParser should be investigated with respect to the specific properties of Croatian. Experiment with combining parsers and different parsing settings along the lines of experiments with the Index Thomisticus treebank (Passarotti and Del'Orletta 2010) should also be conducted. Specifically, we would like to look into the possibilities of hybridization of the before-mentioned state-of-the-art data-driven parsers by linking them with language specific resources such as valency lexicons (e.g. CROVALLEX, Mikelić Preradović et al. 2009). These research paths will be accompanied by a more elaborate investigation into all the different, i.e. treebank-encoded properties of Croatian language influencing the various aspects of dependency parsing accuracy.

## 5. Acknowledgements

Special thanks to our colleagues Tena Gnjatović and Ida Raffaelli from the Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, for substantial contributions to the process of manual annotation of sentences for HOBS.

The results presented here were partially obtained from research within projects ACCURAT (FP7, grant 248347), CESAR (ICT-PSP, grant 271022) funded by EC, and and partially from projects 130-1300646-0645 and 130-1300646-1776 funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

## 6. References

Agić Ž, Šojat K, Tadić M. (2010). An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank. Proceedings of the 32nd International Conference on Information Technology Interfaces, Zagreb, SRCE University Computer Centre, University of Zagreb, 2010. pp. 55-60.

Attardi G, Dell'Orletta F, Simi M, Chanev A, Ciaramita M. (2007). Multilingual Dependency Parsing and Domain Adaptation using DeSR. Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, Prague.

Buchholz S, Marsi E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL), New York, NY, pp. 149-164.

Džeroski S, Erjavec T, Ledinek N, Pajas P, Žabokrtský Z, Žele A. (2006). Towards a Slovene Dependency Treebank. Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06, 24-26 May 2006. Genoa.

Hajič J. (1996). Formal Representation of Language Structures. TELRI Newsletter, 3, pp. 12-19. See URL http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch06.html.

Hajič J et al. (1999). Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory. Praha: Karlová Univerzita. See URL http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch06.html.

Hajič J, Böhmová A, Hajičová E, Vidová Hladká B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. Treebanks: Building and Using Parsed Corpora, Amsterdam, Kluwer, 2000.

McDonald R, Lerman K, Pereira F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL).

Mikelić Preradović N, Boras D, Kišiček S. (2009). CROVALLEX: Croatian Verb Valence Lexicon. Proceedings of the 31st International Conference on Information Technology Interfaces, pp. 533-538. See URL http://cal.ffzg.hr/crovallex/index.html.

Nilsson J, Nivre J. (2008). MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech-Paris, ELRA, 2008.

Nivre J, Hall J, Nilsson J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), May 24-26, 2006, Genoa, Italy, pp. 2216-2219.

Nivre J, Hall J, Kübler S, McDonald R, Nilsson J, Riedel S, Yuret D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007, Prague, Czech Republic, pp. 915-932.

Pajas P. (2000). Tree Editor TrEd, Prague Dependency Treebank, Charles University, Prague. See URL http://ufal.mff.cuni.cz/ pajas/tred.

Passarotti M, Dell'Orletta F. (2010). Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin. Proceedings of the Seventh conference on International Language Resources and Evaluation, ELRA, 2010.

Sgall P, Hajičová E, Panevová J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht, D. Reidel Publishing Company.

Silić J, Pranjković I. (2007). Gramatika hrvatskoga jezika. Zagreb: Školska knjiga.

Tadić M. (2002). Building the Croatian National Corpus. Proceedings of the 3rd International Conference on Language Resources and Evaluation, ELRA.

Tadić M. (2007). Building the Croatian Dependency Treebank: the initial stages. Suvremena lingvistika, 63, pp. 85-92.

Tadić M. (2009). New version of the Croatian National Corpus. After Half a Century of Slavonic Natural Language Processing. Brno, Masaryk University, 2009, pp. 199-205.

Titov I, Henderson J. (2007). Fast and Robust Multilingual Dependency Parsing with a Generative Latent Variable Model. CoNLL 2007 Shared Task, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07), Prague, Czech Republic, 2007.