HamleDT: To Parse or Not to Parse?

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, Jan Hajič

Charles University in Prague, Faculty of Mathematics and Physics,
Malostranské náměstí 25, 118 00, Praha, Czechia
{last-name}@ufal.mff.cuni.cz

Abstract

We propose HamleDT – *HArmonized Multi-LanguagE Dependency Treebank*. HamleDT is a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style. While the license terms prevent us from directly redistributing the corpora, most of them are easily acquirable for research purposes. What we provide instead is the software that normalizes tree structures in the data obtained by the user from their original providers.

Keywords: dependency treebank, annotation scheme, harmonization

1. Introduction

Growing interest in dependency parsing is accompanied (and inspired) by availability of new treebanks for various languages. Shared tasks such as CoNLL 2006 – 2009 have promoted parser evaluation in multilingual setting. However, differences in parsing accuracies on different languages cannot be always attributed to language differences. They are often caused by differing sizes and annotation styles of the treebanks. The impact of data size can be estimated by learning curve experiments but it is difficult to normalize the annotation style. We provide a method (including software that implements the method) to transform the treebanks into one common style. We have studied treebanks of 29 languages and collected a long list of variations. For each phenomenon, we propose one common style and provide a transformation from the original annotation to that style. Besides structure, we also unify the tagsets of both the part-of-speech/morphological tags, and the dependency relation tags.

The common style defined by us serves as a reference point: being able to say "our results are based on HamleDT 1.0 transformations of treebank XY" facilitates comparability of published results. On the other hand, one can use the software to transform all the treebanks into another common style if it suits their needs better. Also, we believe that the unified representation of linguistic content proves advantageous for linguists who want to compare languages based on treebank material.

2. Related Work

Recently there have been a few attempts to address the same problem, namely:

- (Tsarfaty et al., 2011) compare performance of two parsers on different conversions of the Penn Treebank. They do not see the solution in data transformations; instead, they develop evaluation technique that is robust with respect to annotation style.
- (McDonald et al., 2011) experiment with cross-language parser training and they rely on a rather small universal set of part-of-speech tags. They do not transform structure however, and they observe that different annotation schemes across treebanks are responsible for the fact that some language pairs work better together than others.
- Three different dependency parsers developed and tested with respect to two Italian treebanks are compared in (Bosco et al., 2010).
- (Bengoetxea and Gojenola, 2009) evaluate three types of transformations on Basque: projectivization, subordinated sentences and coordination. An important difference between their approach and ours is that their transformations can change tokenization.
- (Nilsson et al., 2006) show that transformations of coordination and verb groups improve parsing.

3. Data

We identified over 30 languages for which treebanks exist and are available for research. Most of the datasets can either be acquired free of charge or they are included in the Linguistic Data Consortium membership fee.

Many treebanks are natively dependency-based but some were originally based on constituents and their conversion included a head-selection procedure. For instance, the Spanish phrase-structure trees were converted to dependencies using a procedure described in (Civit et al., 2006).

HamleDT currently covers the following treebanks and a few others will be added soon (note the ISO 639 codes after the language names—we use them to refer to the languages elsewhere in the paper). Data sizes are summarized in Table 1.

- Arabic (ar): Prague Arabic Dependency Treebank 1.0 / CoNLL 2007 (Smrž et al., 2008)[1]
- Basque (eu): Basque Dependency Treebank (larger version than CoNLL 2007 generously provided by IXA Group) (Aduriz et al., 2003)
- Bengali (bn): *see Hindi*
- Bulgarian (bg): BulTreeBank (Simov and Osenova, 2005)[2]
- Catalan (ca) and Spanish (es): AnCora (Taulé et al., 2008)
- Czech (cs): Prague Dependency Treebank 2.0 / CoNLL 2009 (Hajič et al., 2006)[3]
- Danish (da): Danish Dependency Treebank / CoNLL 2006 (Kromann et al., 2004), now part of the Copenhagen Dependency Treebank[4]
- Dutch (nl): Alpino Treebank / CoNLL 2006 (van der Beek et al., 2002)[5]
- English (en): Penn TreeBank 2 / CoNLL 2009 (Surdeanu et al., 2008)[6]
- Estonian (et): Eesti keele puudepank / Arborest (Bick et al., 2004)[7]
- Finnish (fi): Turku Dependency Treebank (Haverinen et al., 2010)[8]
- German (de): Tiger Treebank / CoNLL 2009 (Brants et al., 2002)[9]
- Greek (modern) (el): Greek Dependency Treebank (Prokopidis et al., 2005)
- Greek (ancient) (grc) and Latin (la): Ancient Greek and Latin Dependency Treebanks (Bamman and Crane, 2011)[10]
- Hindi (hi), Bengali (bn) and Telugu (te): Hyderabad Dependency Treebank / ICON 2010 (Husain et al., 2010)
- Hungarian (hu): Szeged Treebank (Csendes et al., 2005)[11]
- Italian (it): Italian Syntactic-Semantic Treebank / CoNLL 2007 (Montemagni et al., 2003)[12]
- Japanese (ja): Verbmobil (Kawata and Bartels, 2000)[13]
- Latin (la): *see Greek (ancient)*
- Persian (fa): Persian Dependency Treebank (Rasooli et al., 2011)[14]
- Portuguese (pt): Floresta sintá(c)tica (Afonso et al., 2002)[15]
- Romanian (ro): Romanian Dependency Treebank (Călăcean, 2008)[16]
- Russian (ru): Syntagrus (Boguslavsky et al., 2000)
- Slovene (sl): Slovene Dependency Treebank / CoNLL 2006 (Džeroski et al., 2006)[17]
- Spanish (es): *see Catalan*
- Swedish (sv): Talbanken05 (Nilsson et al., 2005)[18]
- Tamil (ta): TamilTB (Ramasamy and Žabokrtský, 2012)[19]
- Telugu (te): *see Hindi*
- Turkish (tr): METU-Sabanci Turkish Treebank (Atalay et al., 2003)[20]

3.1. Train/test division

Many treebanks (especially those from CoNLL) define a train/test data split. This is important for comparability of experiments with automated tagging and parsing. We thus decided to define test subsets for the remaining treebanks, too. On doing so, we tried to keep the test size similar to the majority of CoNLL 2006/2007 test sets, i.e. roughly 5000 tokens.

4. Harmonization

Our effort aims at identifying all syntactic constructions for which there is at least one treebank where the annotation systematically differs from other treebanks. In a typical case, such constructions can be identified automatically using existing syntactic and morphological tags, i.e. with little or no lexical knowledge. Thanks to this fact we were able to design algorithms to normalize the annotations to one style.

[1] http://padt-online.blogspot.com/2007/01/conll-shared-task-2007.html
[2] http://www.bultreebank.org/indexBTB.html
[3] http://ufal.mff.cuni.cz/pdt2.0/
[4] http://code.google.com/p/copenhagen-dependency-treebank/
[5] http://odur.let.rug.nl/~vannoord/trees/
[6] http://www.cis.upenn.edu/~treebank/
[7] http://www.cs.ut.ee/~kaili/Korpus/puud/
[8] http://bionlp.utu.fi/fintreebank.html
[9] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
[10] http://nlp.perseus.tufts.edu/syntax/treebank/greek.html, http://nlp.perseus.tufts.edu/syntax/treebank/latin.html
[11] http://www.inf.u-szeged.hu/projectdirs/hlt/index_en.html
[12] http://medialab.di.unipi.it/isst/
[13] http://www.sfs.uni-tuebingen.de/en/tuebajs.shtml
[14] http://dadegan.ir/en/persiandependencytreebank
[15] http://www.linguateca.pt/floresta/info_floresta_English.html
[16] http://www.phobos.ro/roric/texts/xml/
[17] http://nl.ijs.si/sdt/
[18] http://www.msi.vxu.se/users/nivre/research/Talbanken05.html
[19] http://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1/
[20] http://www.ii.metu.edu.tr/content/treebank

Language	Primary data source	Pri. tree type	Used data source	Sents.	Tokens	Train / test [% snt]	Avg. snt. length	Nproj. deps. [%]
Arabic (ar)	PADT	dep	CoNLL 2007	3043	116793	96 / 4	38.38	0.37
Basque (eu)	BDT	dep	primary	11226	151604	90 / 10	13.50	1.27
Bengali (bn)	HyDT	dep	ICON 2010	1129	7252	87 / 13	6.42	1.08
Bulgarian (bg)	BTB	phr	CoNLL 2006	13221	196151	97 / 3	14.84	0.38
Catalan (ca)	AnCora	phr	CoNLL 2009	14924	443317	88 / 12	29.70	0.00
Czech (cs)	PDT	dep	CoNLL 2007	25650	437020	99 / 1	17.04	1.91
Danish (da)	DDT	dep	CoNLL 2006	5512	100238	94 / 6	18.19	0.99
Dutch (nl)	Alpino	phr	CoNLL 2006	13735	200654	97 / 3	14.61	5.41
English (en)	PTB	phr	CoNLL 2009	40613	991535	97 / 3	24.41	0.39
Estonian (et)	EP	phr	primary	1315	9491	90 / 10	7.22	0.07
Finnish (fi)	Turku	dep	primary	4307	58576	90 / 10	13.60	0.51
German (de)	Tiger	phr	CoNLL 2009	38020	680710	95 / 5	17.90	2.33
Greek (el)	GDT	dep	CoNLL 2007	2902	70223	93 / 7	24.20	1.17
Greek (grc)	AGDT	dep	primary	21160	308882	98 / 2	14.60	19.58
Hindi (hi)	HyDT	dep	ICON 2010	3515	77068	85 / 15	21.93	1.12
Hungarian (hu)	Szeged	phr	CoNLL 2007	6424	139143	94 / 6	21.66	2.90
Italian (it)	ISST	dep	CoNLL 2007	3359	76295	93 / 7	22.71	0.46
Japanese (ja)	TüBa	dep	CoNLL 2006	17753	157172	96 / 4	8.85	1.10
Latin (la)	LDT	dep	primary	3473	53143	91 / 9	15.30	7.61
Persian (fa)	PDT	dep	primary	12455	189572	97 / 3	15.22	1.77
Portuguese (pt)	Floresta	phr	CoNLL 2006	9359	212545	97 / 3	22.71	1.31
Romanian (ro)	RDT	dep	primary	4042	36150	93 / 7	8.94	0.00
Russian (ru)	Syntagrus	dep	primary	34895	497465	99 / 1	14.26	0.83
Slovene (sl)	SDT	dep	CoNLL 2006	1936	35140	79 / 21	18.15	1.92
Spanish (es)	AnCora	phr	CoNLL 2009	15984	477810	90 / 10	29.89	0.00
Swedish (sv)	Talbanken	phr	CoNLL 2006	11431	197123	97 / 3	17.24	0.98
Tamil (ta)	TamilTB	dep	primary	600	9581	80 / 20	15.97	0.16
Telugu (te)	HyDT	dep	ICON 2010	1450	5722	90 / 10	3.95	0.23
Turkish (tr)	METU	dep	CoNLL 2007	5935	69695	95 / 5	11.74	5.33

Table 1: Overview of data resources processed by the date of publication of this paper. The average sentence length is the number of tokens divided by the number of sentences (note that some treebanks, e.g. Bengali and Telugu, work with sentence chunks as if they were tokens; others, e.g. Arabic and Persian, use extended tokenization that splits certain words). The last column gives the percentage of nodes attached nonprojectively.

Our default normalized form is mostly derived from the annotation style of the Prague Dependency Treebank. It is a matter of convenience for large part: This is the scheme the authors feel most at home, and many of the included treebanks already use a style similar to PDT. We do not want to assert that the PDT style is objectively better than the other styles. (Note however that in case of coordination, the PDT style provides more expressive power than the other options.)

The normalization procedure involves both structural transformation and dependency relation relabeling. While we strive to design the structural transformations as reversible as possible, we do not attempt to save all the information stored in the labels. The DEPREL tagsets are *very* different across the treebanks, ranging from simple statements such as "this is a noun phrase modifying something" over standard *subject, object* etc. relations, to deep-level functions of Pāṇinian grammar such as *karma* and *karta*. It does not seem possible to unify these tagsets without manual relabeling of the whole treebanks.

We use a lossy scheme that maps the DEPREL tags on the moderately-sized tagset of PDT analytical functions (more or less the same as the DEPREL tags in CoNLL Czech data).

Occasionally the original structure and dependency labels are not enough to determine the normalized output, as we also need to consider the part-of-speech, the word form or even the values of morphological features. Since the POS/morphological tagsets also vary greatly across treebanks, we use the Interset approach described by (Zeman, 2008) to access all the morphological information. As a by-product, the normalized treebanks provide Interset-unified morphology, too.

Here is a selection of phenomena that we observed and, to various degrees for various languages, included in our normalization scenario. (Language codes in brackets give examples of treebanks where the particular approach is employed.)

4.1. Coordination

Capturing coordination in a dependency framework has been repeatedly described as difficult for both treebank designers and parsers. Our analysis revealed four families of approaches that further vary in attachment of punctuation, shared modifiers etc.: Prague (all con-

juncts headed by the conjunction) [ar, bn, cs, el, eu, grc, hi, la, nl, sl, ta, te], Mel'čukian (the first/last conjunct is the head, others organized in a chain) [de, en, ja, ru, sv, tr], Stanford (the first/last conjunct is the head, others attached directly to it) [bg, ca, es, fi, it, pt] and Tesnièrian (no common head, all conjuncts attached directly to the node modified by the coordination structure) [hu]. Furthermore, the Prague style provides for nested coordinations, as in *"apples and oranges or pears and lemons"*. It also distinguishes between shared modifiers, as the subject in *"Mary came and cried"*, from private modifiers of the conjuncts, as in *"John came and Mary cried"*. As this distinction is missing in non-Prague-style treebanks, we cannot recover it reliably. We apply a few heuristics but in most cases the modifiers of the head conjunct will remain private modifiers after normalization.

Danish employs a mixture of the Stanford and Mel'čukian styles, where the last conjunct is attached indirectly via the conjunction. The Romanian and Russian treebanks omit punctuation tokens (these do not have corresponding nodes in the tree); in the case of Romanian, this means that coordinations of more than two conjuncts get split.

4.2. Prepositions

Prepositions (or postpositions) can either govern their noun phrase [cs, sl, en, …] or they can be attached to the head of the NP [hi]. When they govern the NP, other modifiers of the main noun are usually attached to the noun but they can also be attached to the preposition [de]. The label of the relation of the PP to its parent can be found at the prepositional head [de, en, nl], or the preposition, despite serving as head, gets an auxiliary label (such as `AuxP` in PDT) and the real label is found at the NP head [cs, sl, ar, el, la, grc].

We propose the [cs] approach here.

4.3. Subordinated Clauses

Roots (predicates) of relative clauses are usually attached to the noun they modify (example: in *"the man who came yesterday"*, *"came"* would be attached to *"man"* and *"who"* would be attached to *"came"* as its subject). Some clauses use a subordinating conjunction (complementizer; e.g. *"that, dass, que, che"* if not used as a relative pronoun/determiner, example: *"the man said that he came yesterday"*). The conjunction can either be attached to the predicate of the embedded clause [es, ca, pt, de, ro] or it can lie between the clause and the main predicate it modifies [cs, en, hi, it, ru, sl]. In the latter case the label of the relation of the clause to its parent can be assigned to the conjunction [en, it, hi] or to the clausal predicate [cs, sl]. The comma before the conjunction is attached either to the conjunction or to the predicate of the clause. The Romanian treebank is segmented to clauses instead of sentences, so every clause has its own tree and inter-clausal relations are not annotated.

We propose the [cs] approach here.

4.4. Verb Groups

Various sorts of verbal groups include analytical verb forms (such as auxiliary + participle), modal verbs with infinitives and similar constructions. Dependency relations, both internal (between group elements) and external (leading to parent on one side and verb modifiers on the other side), may be defined according to various criteria: content verb vs. auxiliary, finite form vs. infinitive, subject-verb agreement (typically holds for finite verbs and participles but not for infinitives). Participles often govern auxiliaries [es, ca, it, ro], elsewhere the finite verb is the head [pt, de, nl, en, sv] or both approaches are possible based on semantic criteria [cs]. In [hi, ta], the content verb (which could be a participle or a bare verb stem) is the head and auxiliaries (finite or participles) are attached to it. The head typically bears the label describing the relation of the group to its parent. As for child nodes, subject and negative particle (if any) are often attached to the head, especially if it is the finite element [de, en] while the arguments (objects) are attached to the content element whose valency slot they fill (often participle or infinitive). Sometimes even the subject [nl] or the negative particle [pt] can be attached to the non-head content element. Various infinitive-marking particles (English *"to"*, Swedish *"att"*, Bulgarian *"da"*) can be treated similarly to subordinating conjunctions, can govern the infinitive [en, bg] or be attached to it. In [pt], prepositions used between main verb and the infinitive (*"estão a usufruir"*) are attached to the infinitive. In [bg], all modifiers of the verb including the subject are attached to *"da"* instead of the verb below. We intend to unify verbal groups under one common approach but the current version of HamleDT does not do so yet. This part is more language-dependent than the others and further analysis is needed.

4.5. Determiner Heads

The Danish treebank is probably the most extraordinary one. Nouns often depend on determiners, numerals etc.: the opposite of what the rest of the world is doing.

We propose to attach articles (determiners) to their nouns, and numerals to the counted nouns.[21]

4.6. Punctuation

Paired punctuation (quotation marks, brackets, parenthesizing commas) is typically attached to the head of the segment between the marks. Occasionally it is attached one level higher, to the parent of the enclosed segment, which may break projectivity [pt]. Non-coordinating unpaired punctuation symbols are usually attached to a neighboring symbol or its parent. In [it], left paired marks are attached to the next token and all the others to the previous token.

[21] Note however that numeral heads are not restricted to [da]. Czech has a complex set of rules about numerals, which may result under some circumstances in the numeral serving as head.

Sentence-final punctuation is attached to the artificial root node [cs, ar, sl, grc, ta], to the main predicate [bg, ca, da, de, en, es, et, fi, hu, pt, sv], to the predicate of the last clause [hi], to the previous token [eu, it, ja, nl]. In [la] there is no final punctuation. In [bn, te] it is rare but when present, it can govern a few previous tokens! In [tr], it is attached to the artificial root node but instead of being sibling of the main predicate, the punctuation governs the predicate.
We propose to attach the sentence-final punctuation to the artificial root node; paired punctuation to the root of the subtree inside; for the other punctuation occurrences, further analysis is needed.

4.7. Tokenization and Sentence Segmentation

The only aspect we do not intend to change is tokenization. Our harmonized trees always have the same number of nodes as the original annotation, despite some variability in approaches we observe in the treebanks. Some treebanks collapse multi-word expressions into single nodes [ca, da, es, eu, fa, hu, it, nl, pt]. In [hu], collapsing is restricted to personal names. In [fa], it is used for analytical verb forms. The word form of the node is composed of all the participating words, joined by the underscore character or even by a space [fa].
In [bn, te], dependencies are annotated between chunks, instead of words. So one node may comprise a whole verbal group with all auxiliaries, or a noun with its postposition(s).
On the other hand, there are treebanks [ar, fa] where orthographic words can be split into syntactically autonomous parts. An Arabic example: وبالفالوجة = *wabiālfālūjah* = *wa*/CONJ + *bi*/PREP + *AlfAlwjp*/NOUN_PROP = "and in al-Falujah".
In [ro, ru], punctuation tokens are ignored and do not get a node in the tree structure.
Occasionally [bn, hi, te] we see an inserted NULL node, mostly for participants deleted on surface, as in this Hindi example: दीवाली के दिन जुआ खेलें मगर NULL घर में या होटल में. = *dīvālī ke dina juā khelaṁ magara NULL ghara meṁ yā hoṭala meṁ.* = "On Diwali they gamble but [they do so] at home or hotel." (The NULL node stands for the deleted phrase *they do so.*)
Similarly to tokenization, we also take sentence segmentation as fixed, despite some less usual solutions: in [ar], sentence-level units are paragraphs rather than sentences (which explains the high average segment length in Table 1). In contrast, [ro] annotates every clause as a separate tree.

5. Obtaining HamleDT

While the license terms of some of the treebanks prevent us from directly redistributing them (in the original or normalized form), most of them are easily acquirable for research purposes. What we provide instead is the software that normalizes tree structures in the data obtained by the user from their original providers. News will be announced at our web site:

http://ufal.mff.cuni.cz/hamledt

All the normalizations are implemented in Treex (formerly TectoMT) (Žabokrtský et al., 2008), a modular open-source framework for structured language processing, written in Perl.[22] In addition to the normalization scripts for each language/treebank, Treex contains also other transformations, so for example, coordinations in any treebank can be converted from Prague style to Stanford style.
Finally, there is the tree editor TrEd[23] that can open Treex files and display original and normalized trees side-by-side on multiple platforms.

6. Conclusion

We proposed a method for automatic normalization of annotation styles of many publicly available treebanks for various languages. The method applies transformation rules conditioned on the original structural annotation, dependency labels and morphosyntactic tags. We also unify the tag sets for parts of speech, morphosyntactic features and dependency relation labels. We take care to make the structural transformations and the morphosyntactic tagset unification as reversible as possible (we do not attempt the same with dependency relations).
We provide an implementation of the transformations in the open-source framework Treex. It can also be used for transforming the data to any other annotation style, besides the one we propose. A subset of the treebanks whose license terms permit redistribution is available directly from us. For the rest, the user has to acquire the original data first, then to apply our transformation tool.

7. Acknowledgements

The authors wish to express their gratitude to all the creators and providers of the respective corpora.
The work on this project was supported by the Czech Science Foundation grant no. P406/11/1499, and GAUK 116310. The work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

8. References

Itzair Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Aitzpea Garmendia, and Maite Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories.*

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. "Floresta sintá(c)tica": a treebank for Portuguese. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1968–1703.

Nart B. Atalay, Kemal Oflazer, Bilge Say, and Informatics Inst. 2003. The annotation process in the

[22] http://ufal.mff.cuni.cz/treex/
[23] http://ufal.mff.cuni.cz/tred/

Turkish treebank. In *In Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora (LINC*.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In Caroline Sporleder, Antal Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98. Springer Berlin Heidelberg.

Kepa Bengoetxea and Koldo Gojenola. 2009. Exploring treebank transformations in dependency parsing. In *Proceedings of the International Conference RANLP-2009*, pages 33–38, Borovets, Bulgaria, September. Association for Computational Linguistics.

Eckhard Bick, Heli Uibo, and Kaili Müürisep. 2004. Arborest – a VISL-style treebank derived from an Estonian constraint grammar corpus. In *Proceedings of Treebanks and Linguistic Theories*.

Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadezhda Frid. 2000. Dependency treebank for Russian: Concept, tools, types of information. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 987–991. Association for Computational Linguistics Morristown, NJ, USA.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Alessandro Lenci, Leonardo Lesmo, Giuseppe Attardi, Maria Simi, Alberto Lavelli, Johan Hall, Jens Nilsson, and Joakim Nivre. 2010. Comparing the influence of different treebank annotations on dependency parsing.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Montserrat Civit, Maria Antònia Martí, and Núria Bufí. 2006. Cat3LB and Cast3LB: From constituents to dependencies. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, pages 141–152. Springer.

Mihaela Călăcean. 2008. Data-driven dependency parsing for Romanian. Master's thesis, Uppsala University, August.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *TSD*, volume 3658 of *Lecture Notes in Computer Science*, pages 123–131. Springer.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdeněk Žabokrtský, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proceedings of the Fifth International Language Resources and Evaluation Conference, LREC 2006*, pages 1388–1391, Genova, Italy. European Language Resources Association (ELRA).

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking Finnish. In Markus Dickinson, Kaili Müürisep, and Marco Passarotti, editors, *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.

Samar Husain, Prashanth Mannem, Bharat Ambati, and Phani Gadde. 2010. The ICON-2010 tools contest on Indian language dependency parsing. In *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing*, Kharagpur, India.

Yasuhiro Kawata and Julia Bartels. 2000. Stylebook for the Japanese treebank in Verbmobil. In *Report 240*, Tübingen, Germany, September 29.

Matthias T. Kromann, Line Mikkelsen, and Stine Kern Lynge. 2004. Danish dependency treebank.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, *Building and using Parsed Corpora*, Language and Speech series, pages 189–210, Dordrecht. Kluwer.

Jens Nilsson, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proceedings of the NODALIDA Special Session on Treebanks*.

Jens Nilsson, Joakim Nivre, and Johan Hall. 2006. Graph transformations in data-driven dependency parsing. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 257–264. Association for Computational Linguistics.

Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, and Stelios Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 149–160.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague dependency style treebank for Tamil. In *Proceedings of LREC 2012*, İstanbul, Turkey.

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. 2011. A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 227–231, Poznań, Poland.

Kiril Simov and Petya Osenova. 2005. Extending the annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 173–184, Barcelona, December.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague Arabic dependency treebank: A word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco. European Language Resources Association.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *LREC*. European Language Resources Association.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 385–396, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. 2002. Chapter 5. the Alpino dependency treebank. In *Algorithms for Linguistic Processing NWO PIONIER Progress Report*, Groningen, The Netherlands.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA. Association for Computational Linguistics.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008*, pages 28–30, Marrakech, Morocco, May. European Language Resources Association (ELRA).