SLOVAK DEPENDENCY TREEBANK IN UNIVERSAL DEPENDENCIES

Daniel Zeman

Faculty of Mathematics and Physics
Charles University
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract: We describe a conversion of the syntactically annotated part of the Slovak National Corpus into the annotation scheme known as Universal Dependencies. Only a small subset of the data has been converted so far; yet it is the first Slovak treebank that is publicly available for research. We list a number of research projects in which the dataset has been used so far, including the first parsing results.

Keywords: treebank; dependency; universal dependencies; syntax; morphology; tagging; parsing

1 Introduction

Syntactically annotated corpora (treebanks) are important language resources, indispensable for linguistic research and natural language processing alike. Modern treebanks are mostly built on the notion of dependency relations. With the increasing number of languages covered and amount of data available, there is a growing interest in finding one common, linguistically adequate and cross-linguistically applicable annotation style [5, 13]. Universal Dependencies (UD)[1] [6] is an international effort aimed at such an annotation standard; at the same time, UD also releases treebanks annotated according to the UD guidelines, and has arguably become the largest collection of freely available dependency treebanks worldwide.

UD treebanks are released twice a year and every release so far added several languages that had not been part of the previous releases. The group of Slavic languages is represented quite well. [12] gave an early account of Slavic languages in UD 1.1, as well as an overview of other Slavic treebanks outside UD (Table 1). At the time of this writing, UD 2.0 is the most recent release and it comprises 70 treebanks of 50

[1] http://universaldependencies.org/

languages; among them, nearly all[2] Slavic languages are represented with at least a small dataset (Table 2).

In the present article we focus on one of the recent additions to UD, the Slovak Dependency Treebank.

Language	Code	Treebank	Sent	Tok
Bulgarian	[bg]	BulTreeBank	13,221	196K
Church Slavonic	[cu]	PROIEL	7,818	72K
Croatian	[hr]	SETimes.HR	3,736	84K
Czech	[cs]	PDT	87,913	1504K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	SynTagRus	63,000	900K
Slovak	[sl]	SNK	63,238	994K
Slovenian	[sl]	SSJ500K	27,829	500K

Table 1. Dependency treebanks of Slavic languages, as listed by [12] (only some of them were converted to UD at that time).

2 Slovak Dependency Treebank

The data in the Slovak treebank come from the Slovak National Corpus (*Slovenský národný korpus,* SNK)[3] [8]. Over 63,000 sentences (almost one million words) received manual morphological and syntactic annotation, making it one of the three largest treebanks of Slavic languages (after the Czech PDT [1],[4] and with similar size to the Russian SynTagRus [2]). [8] describe the composition of the treebank as 78% fiction, 13% scientific and 9% journalistic text. An important point is that it includes free sources like Wikipedia or folk tales, where intellectual property rights do not complicate access to and distribution of the annotated data. Most sentences were independently annotated by two annotators in order to identify difficult phenomena and reduce annotation errors. The positions where the two annotators disagree would eventually be decided by a third annotator. Unfortunately, this

[2] Serbian and Upper Sorbian are ready to be released in UD 2.1. What remains missing is Lower Sorbian, Bosnian/Montenegrin and Macedonian; and one may also argue for some smaller languages with less clear status such as Kashubian or Rusyn.

[3] http://korpus.juls.savba.sk/

[4] http://ufal.mff.cuni.cz/pdt

final step has not been completed for all the sentences, which also means that the treebank has yet to wait for its full official release.[5]

Language	**Code**	**Treebank**	**Sent**	**Tok**
Belarusian	[be]	UD	393	8K
Bulgarian	[bg]	BulTreeBank	11,138	156K
Church Slavonic	[cu]	PROIEL	6,337	58K
Croatian	[hr]	SETimes.HR	8,889	197K
Czech	[cs]	PDT	87,913	1506K
Czech	[cs]	CAC	24,709	494K
Czech	[cs]	CLTT	1,125	38K
Czech	[cs]	PUD	1,000	19K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	Google	5,030	99K
Russian	[ru]	SynTagRus	61,889	1107K
Russian	[ru]	PUD	1,000	19K
Serbian*	[sr]	SETimes.SR	3,891	87K
Slovak	[sl]	SNK	10,604	106K
Slovenian	[sl]	SSJ200K	8,000	141K
Slovenian	[sl]	SST	3,188	29K
Ukrainian	[uk]	UD	1,706	26K
Upper Sorbian	[hsb]	UD	646	11K

Table 2. Slavic treebanks in UD release 2.0 (plus Serbian, scheduled for UD 2.1). Note that there were two UD 2.0 releases and the counts in this table sum up both: First, training and development data were released in March 2017. The test sets were kept aside for the CoNLL 2017 Shared Task in dependency parsing, and they were released in May 2017 after the shared task.

Morphological annotation in SNK assigns to each word (token) its lemma and a morphological tag that encodes its part of speech and values of relevant morphological features: inflection type, gender, number, case, degree of comparison, agglutination (preposition + pronoun), verbal form, aspect, polarity, voice etc.[6]

[5] The Slovak Language Treebank has been listed in META-SHARE since 2011 (http://metashare.korpus.sk/repository/browse/slovak-treebank/36e46d0a649311e292cd00163e00007874586ecb0acd48909e54babd7c5e7bc2/) but it cannot be downloaded from there.

[6] See http://korpus.juls.savba.sk/attachments/morpho/tagset-www.pdf for documentation of the tagset (in Slovak).

The syntactic annotation follows the annotation guidelines of the "analytical layer" of the Prague Dependency Treebank.[7]

The following steps have been taken to ensure quality of the data. Note that these are filtering steps—on the first sign of a problem, the entire unit (sentence or file) is discarded. In most cases it should be possible to manually fix the problem and retain the sentence; however, the obvious short-term advantage of the filtering approach is that it requires fewer human resources and the problem-free part of the data can be made available sooner.

- Removed files where the morphological annotation was not manual.
- Removed files where the syntactic annotation was done only by one annotator.
- Removed files where the annotators disagree in sentence segmentation (different number of sentences).
- Removed sentences where the annotators disagree in tokenization (different number of tokens).
- Removed empty sentences and sentences with just one token.
- Removed sentences where one or more annotation items (lemmas, morphological tags, dependency relation labels) were empty.

The resulting corpus consisted of 40,350 sentences and 671,968 tokens. Every sentence in this data set had two complete dependency trees from two annotators. In general, the contrasted annotations were not expected to differ on the word level and in morphological annotation because the annotators were focusing on syntax (while morphology was inherited from pre-existing annotation of SNK). However, it seems that they occasionally modified the lower layers: there were 747 mismatches in word forms, 2 in lemmas and 3 in morphological tags. Again, all affected sentences were removed.

As for the syntactic annotation, the two annotators agreed on 80.34% dependencies (both the parent node index and the dependency label). If we disregard the dependency labels and only look at the parent node assignment, the agreement rate rises to 87.31%. It means that in 6.97% of all tokens (35% of dependency errors) the sole disagreement is in the label (termed *analytical function* or *afun* in Prague-style treebanks).

[7] See http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html for documentation.

Finally, let us consider “complete matches” – sentences whose dependency trees from the two annotators were identical in all aspects of annotation. These sentences constitute the most trustworthy core of the corpus, as it is unlikely that two annotators independently make the same error. Only completely matching sentences were selected for the first UD release of the treebank: reliability of the annotation got the top priority. Of course, there are also some downsides to this decision. The first of them is linked to filtering in general: the resulting corpus does not contain whole documents, making any discourse-level studies impossible. The second drawback is perhaps even more serious: the treebank contains a high proportion of short sentences because the more words in the sentence, the higher is the probability of an annotation error. Before removing sentences with annotation mismatches, the average sentence length in the treebank was 16.7 tokens. When only complete matches remained, the average length dropped to 10.0 tokens. (The longest completely matched sentence contained 54 tokens.) Such a corpus is unbalanced and some more complex grammatical structures may be seriously underrepresented in it. It is thus highly desirable to extend the corpus and add more sentences in the future. However, the filtered portion is arguably much better than nothing, and can be used to train statistical parsers for Slovak; with 10,604 sentences and 106,043 tokens it is still a medium-sized treebank, surpassing by an order of magnitude treebanks that are available for some other languages.

Given that there was no official download site for the Slovak treebank, the filtered part was first released, with the permission from the Ľudovít Štúr Institute of Linguistics, in the LINDAT/CLARIN digital library[8] [3]. This release retained the original Prague-style annotation before conversion to the UD standard.

3 Conversion to Universal Dependencies

The conversion of the annotation to the scheme defined in Universal Dependencies consists of two partially independent steps: 1. converting the morphological tags to universal POS tags and features, and 2. converting the dependency relations.

The Interset Perl library[9] [11] was used to convert the values of morphological categories to UD features; since the internal representation of Interset is defined as a kind of Interlingua for

[8] http://hdl.handle.net/11234/1-1822

[9] http://ufal.mff.cuni.cz/interset

morphosyntactic tagsets, and because the features in UD are based directly on Interset, the conversion was rather straightforward. Most of the categories encoded in SNK tags were directly mappable to UD feature values, with the exception of *paradigma* (inflection class) for which there is no direct counterpart in UD.

The situation is less straightforward with part-of-speech categories (the first character of SNK tags). UD guidelines define 17 universal part-of-speech tags (UPOS) that are rather coarse-grained but assumed to be sufficient for any natural language. If more fine-grained distinctions are needed, they should be encoded by additional features (UPOS tags are separate from features in UD).

SNK	Description	UPOS	Features
S	noun	NOUN, PROPN	
A	adjective	ADJ	
P	pronoun	PRON, DET	
N	numeral	NUM	
V	verb	VERB, AUX	
G	participle	ADJ	VerbForm=Part
D	adverb	ADV	
E	preposition	ADP	
O	conjunction	CCONJ, SCONJ	
T	particle	PART	
J	interjection	INTJ	
R	reflexive pronoun	PRON	Reflex=Yes
Y	conditional morpheme	AUX	Mood=Cnd
W	abbreviation	X	Abbr=Yes
Z	punctuation	PUNCT	
Q	unidentifiable	X	Hyph=Yes
#	non-word element	X	
%	citation in foreign language	X	Foreign=Yes
0	digit	NUM	NumForm=Digit

Table 3. Correspondence between SNK POS tags and UPOS (universal part-of-speech tags).

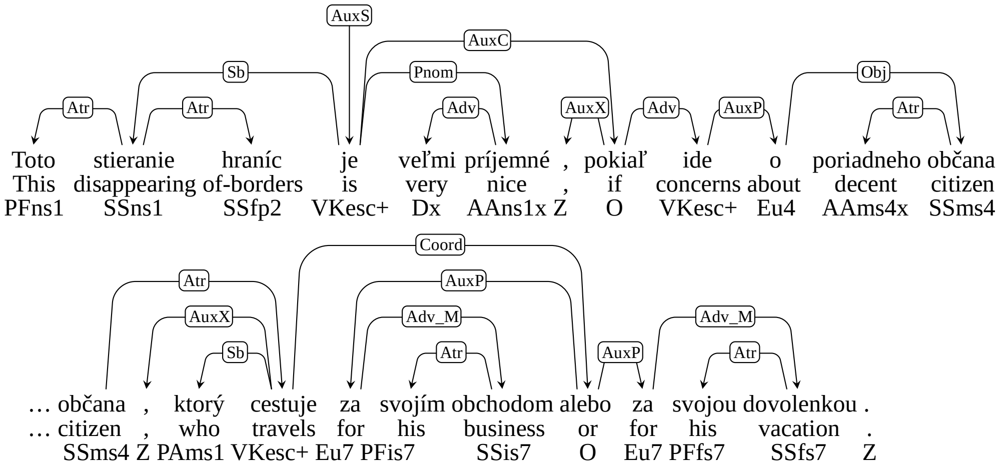
AuxS
AuxC
Sb
Pnom
Atr
Atr
Adv
AuxX
Adv
AuxP
Obj
Atr
Toto stieranie hraníc je veľmi príjemné , pokiaľ ide o poriadneho občana
This disappearing of-borders is very nice , if concerns about decent citizen
PFns1 SSns1 SSfp2 VKesc+ Dx AAns1x Z O VKesc+ Eu4 AAms4x SSms4
Coord
Atr
AuxP
AuxX
Adv_M
Sb
Atr
AuxP
Adv_M
Atr
… občana , ktorý cestuje za svojím obchodom alebo za svojou dovolenkou .
… citizen , who travels for his business or for his vacation .
SSms4 Z PAms1 VKesc+ Eu7 PFis7 SSis7 O Eu7 PFfs7 SSfs7 Z

Fig. 1. Example of an original Prague-style dependency tree.

Table 3 shows the correspondences between SNK POS tags and UPOS. Certain ambiguities are relatively easy to solve—for example, common and proper nouns are distinguished by subsequent characters in the SNK tag. Other ambiguities cannot be resolved by looking at tags alone, and they are addressed outside Interset, taking also the word and its lemma into account. Thus all *pro-adjectives* (pronouns inflecting and behaving like adjectives) are listed and re-tagged DET in UD (see [12] for a discussion of pronouns vs. determiners in Slavic languages). The feature PronType (pronominal type) is set for all pronouns, determiners and pronominal adverbs. Ordinal and multiplicative numerals are distinguished from cardinals by the feature NumType and by changing their tag to ADJ or ADV (the NUM tag is reserved for definite cardinal numbers). Similarly, a word list is used to distinguish coordinating and subordinating conjunctions.

Another change involves polarity of verbs. In SNK, the negative forms with the prefix *ne-* are treated as derivational morphology: they are not encoded in the morphological tags and negative verbs have different lemmas than their affirmative counterparts (e.g. *obviniť* “to accuse” – *neobviniť* “not to accuse”). The Slovak UD data, on the other hand, use the affirmative lemma for both forms and set the Polarity feature to either “Pos” or “Neg”. This is in line with the UD guidelines and improves parallelism to the Czech treebanks in UD. Note that negative verbs can be recognized using simple regular expressions, but one must watch for a few exceptions where an affirmative verb starts with *ne-* *(nechať, nechávať, nenávidieť)*.

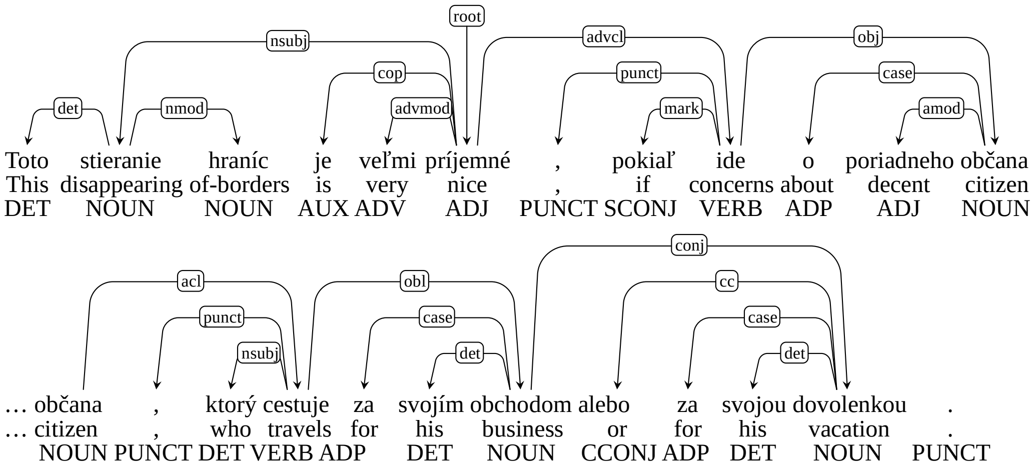
root
nsubj
advcl
obj
cop
punct
case
det
nmod
advmod
mark
amod
Toto stieranie hraníc je veľmi príjemné , pokiaľ ide o poriadneho občana
This disappearing of-borders is very nice , if concerns about decent citizen
DET NOUN NOUN AUX ADV ADJ PUNCT SCONJ VERB ADP ADJ NOUN
conj
acl
obl
cc
punct
case
case
nsubj
det
det
… občana , ktorý cestuje za svojím obchodom alebo za svojou dovolenkou .
… citizen , who travels for his business or for his vacation .
NOUN PUNCT DET VERB ADP DET NOUN CCONJ ADP DET NOUN PUNCT

Fig. 2. The tree from Figure 1 converted to Universal Dependencies.

Further part-of-speech adjustments occur during the transformation of the syntactic structure. For instance, the verb *byť* “to be” usually functions as an auxiliary verb or a copula. If it is found in one of these functions, its tag is changed from VERB to AUX.

Conversion of syntactic annotation is illustrated in Figures 1 and 2. Besides simple relabeling of dependency relations (see also Table 4), it involves several structural transformations:

- Copula verb heads the non-verbal predicate in the Prague style while the non-verbal predicate is the head in UD: *je príjemné* “is nice”.
- In Prague, preposition is plugged as a connector between its noun and the parent of the prepositional phrase. In UD, prepositions are leaves attached to their nouns: *o občana* “about citizen”, *za obchodom* “for business”.
- In Prague, subordinating conjunction is plugged as a connector between the predicate of the subordinate clause and its parent. In UD, subordinating conjunctions are leaves attached to the predicates: *pokiaľ ide* “if it concerns”.
- In Prague, coordination is headed by a conjunction or punctuation symbol; the child nodes are marked as either members of coordination (“_M” attached to afun) or modifiers shared by the members (no suffix). In UD, coordination is headed by the first conjunct (member) and the subsequent conjuncts are attached to it. Shared modifiers cannot be

distinguished from private modifiers of the first conjunct.

SNK	Description	UD
Adv	adverbial modifier	obl, advmod, advcl
Apos	apposition	appos*, punct*
Atr	attribute	amod, det, nummod, nmod, flat, acl
Atv	verbal attribute	acl
AtvV	verbal attribute	xcomp
AuxC	subordinating conjunction	mark*
AuxG	non-comma punctuation	punct
AuxK	sentence-final punctuation	punct*
AuxO	semantically redundant	discourse
AuxP	preposition	case*
AuxR	reflexive passive	expl:pass
AuxT	inherently reflexive verbs	expl:pv
AuxV	auxiliary verb	aux, aux:pass
AuxX	comma	punct
AuxY	extra conjunction	cc, mark
AuxZ	emphasizer	advmod:emph
Coord	coordination head	cc*, conj*, punct*
ExD	ex-dependent (ellipsis)	vocative, advcl, orphan*, dep
Obj	object	obj, iobj, ccomp, xcomp
Pnom	nominal predicate	cop*
Pred	main predicate	root, parataxis
Sb	subject	nsubj, nsubj:pass, csubj, csubj:pass

Table 4. Correspondence between SNK (Prague style) and UD dependency relations. The correspondences marked * are indirect: a structural transformation is necessary when the source relation occurs; the target relation may appear in the resulting structure but it will hold between a different pair of nodes.

The conversion procedure is not trivial because sometimes the rules outlined above interact. Notice how coordination is combined with prepositional phrases in our example—in the Prague style, the real type of the relation between *cestuje* and *za obchodom*, “Adv”, is revealed two levels lower than in the UD tree. Fortunately, there was already software for conversion between the Prague style and UD. The publicly available Treex package[10] [7] in the configuration described in [13] (with some extensions) was reused to convert the Slovak treebank.

The Slovak UD treebank first appeared in the Universal Dependencies release 1.4 in November 2016. In order to facilitate reproducibility of machine learning experiments, the dataset was split to training, development and test section, respectively (Table 5).

Section	Sentences	Tokens
Training	8,483	80,575
Development	1,060	12,440
Test	1,061	13,028

Table 5. The official split of the Slovak UD treebank into training, development and test data.

The second edition that included Slovak, UD 2.0 in March 2017, followed the updated version of the UD guidelines, v2. (All examples in the present article also relate to the v2 guidelines.) The data split was the same as in UD 1.4 but the test sets were released separately after the CoNLL 2017 shared task in dependency parsing.

4 Usage and Related Work

The mere fact that the treebank is available under a free license is very important for the Slovak language in the field of natural language processing. Being a part of a large collection like Universal Dependencies is a bonus that significantly increases visibility of the corpus. According to the statistics published by LINDAT/CLARIN, there have been 79 unique downloads of the Prague-style release of the Slovak Dependency Treebank, 2592 unique downloads of UD 1.4 and 1985 unique downloads of UD 2.0 (as of July 19, 2017).

The treebank can be searched on-line in the PML-TQ search engine maintained by the Charles University in Prague[11] and in the SETS engine at the University of Turku.[12]

Soon after its first release, the treebank was picked (together with Czech, Slovenian, Croatian, Danish, Swedish and Norwegian) by the organizers of the VarDial 2017 shared task in parsing closely related

[10] http://ufal.mff.cuni.cz/treex

[11] https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/ud20_sk/

[12] http://bionlp-www.utu.fi/dep_search/

languages [10]. A much larger shared task was organized as part of the CoNLL 2017 conference[13] [14]. The topic was end-to-end parsing from raw text, via automatic tokenization, sentence segmentation, lemmatization and morphological tagging to universal dependencies. The task set the new state of the art in dependency parsing for 45 languages, including Slovak. Baseline models were produced by the UDPipe system[14] [9]; this parser is open-source and available together with the pre-trained language models. Twenty of the systems competing in the shared task managed to surpass the baseline result; some of them are freely available, too.

The best results for Slovak were achieved by the team from Stanford: 83.86% content-word labeled attachment score (CLAS), **86.04%** labeled attachment score (LAS) and 89.58% unlabeled attachment score (UAS). The parser was processing raw text (that is, it could not access gold-standard sentence segmentation, tokenization and morphology). All models were only trained on the training portion of the Slovak UD treebank. However, since much larger tagged data are available in the Slovak National Corpus, there is room for a significant boost of the tagging accuracy, which in turn may improve parsing results (but note that some parsers do not need morphology on input).

In connection to the shared task, large web corpora have been collected from CommonCrawl and Wikipedia for all the languages. The data have been automatically segmented, lemmatized, tagged and parsed by UDPipe, so there is now also a parsebank of Slovak comprising over 59 million sentences (811 million words) [4]. The first 2 million words have been indexed and made searchable through the SETS engine in Turku.

5 Conclusion and Outlook

We have presented the first public release of the Slovak Dependency Treebank and its automatic conversion to Universal Dependencies using rule-based heuristics and correspondence tables. We have shown that the release practically immediately put Slovak in several interesting NLP research projects where multilingual approaches are studied.

The current version contains only sentences with 100% inter-annotator agreement. This temporary measure ensures quality of the syntactic annotation but it also means that the released dataset is relatively small

[13] http://universaldependencies.org/conll17/

[14] http://ufal.mff.cuni.cz/udpipe

and unbalanced. It may not be easy to find human resources and funds to complete the disagreement resolution in the near future; however, we believe that there is room for checking additional sentences semi-automatically.

There are 7,564 sentences (95K tokens) where there was just one disagreement point between the annotators. Some of the mismatches mentioned in Section 2 may not be important for UD conversion or may be easily fixable. [8] notice that one of the most frequently confused pair of relations is `AuxT` (reflexive pronoun of an inherently reflexive verb) and `AuxR` (reflexive pronoun used to form reflexive passive). Both of them would be subtypes of `expl` in UD. Another frequent mismatch is `AuxX` vs. `Coord`. More research would be needed but if it signals inconsistent encoding of coordination, it could be normalized automatically.

Observations of this kind will hopefully help to speed up the completion of the remaining annotated data. Once all of them are added, the future releases of the Slovak Dependency Treebank will be four times bigger than the current one.

6 Acknowledgements

The work was supported by the grant 15-10472S of the Czech Science Foundation. We have been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

The author wishes to thank Mária Šimková, Katarína Gajdošová and the entire team of the Ľudovít Štúr Institute of Linguistics for their work on the treebank and for the permission to release it.

References

[1] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague dependency treebank 3.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

[2] Boguslavsky, I., Iomdin, L., Petrochenkov, V., Sizov, V., and Tsinman, L. (2013). A case of hybrid parsing: Rules refined by empirical and corpus statistics. In Gerdes, K., Hajičová, E., and Wanner, L., editors, *Computational Dependency Theory*, volume 258, pages 226–240. IOS Press, Amsterdam, Netherlands.

[3] Gajdošová, K., Šimková, M., et al. (2016). Slovak dependency treebank. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1822.

[4] Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task – automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1989.

[5] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, Sofija, Bulgaria.

[6] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia.

[7] Popel, M., and Žabokrtský, Z. (2010). TectoMT: modular NLP framework. *Advances in Natural Language Processing*, pages 293–304.

[8] Šimková, M., and Garabík, R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006*, pages 389–394, Санкт-Петербург, Russia, St. Petersburg University Press.

[9] Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

[10] Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.

[11] Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the 6th International Language Resources and Evaluation Conference, LREC 2008*, pages 28–30, Marrakech, Morocco.

[12] Zeman, D. (2015). Slavic languages in Universal Dependencies. In *Natural Language Processing, Corpus Linguistics, E-learning*, pages 151–163, Lüdenscheid, Germany. RAM-Verlag.

[13] Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

[14] Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Fernandez Alcalde, H., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics. Vancouver, Canada.