Slavic Languages in Universal Dependencies

Daniel Zeman

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Abstract. Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning and linguistic research from a language typology perspective. It is a merger and extension of several previous efforts aimed at finding unified approaches to parts of speech, morphosyntactic descriptions and syntactic dependency relations. In the present contribution we address the application of UD to Slavic languages. We devote the most space to peculiarities of pronouns, determiners, numerals and quantifiers. Other language features that are discussed include modal verbs, ellipsis, nominal predicates, and reflexive pronouns. Most of our examples are from Czech but the language features demonstrated are usually portable to other Slavic languages. We include examples from the other languages where appropriate.

1 Introduction

The general philosophy of the Universal Dependencies (UD) project[1] [14] is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. The first version of the standard was published in October 2014 and datasets for the first ten languages were released in January 2015; version 1.1 with eight additional languages was released in May 2015 and subsequent releases are currently planned after every 6 months.

Forseen applications include typological studies and cross-lingual transfer of parsing models. First experiments with the UD treebanks have already been published. For instance, [18] addressed statistical learnability of the UD dependency structures in comparison to other annotation styles; [1, 23] repeated cross-language parsing experiments that were previously done with unharmonized treebanks and the previous results were not conclusive, mostly because the diverse annotatin styles were not comparable. One interesting observation is that even small datasets can be useful. While bigger is definitely better, [16] found that a "treebank" of as few as 10 sentences gave better parsing accuracy than the best-performing unsupervised method.

UD is based on an evolution of several previous efforts to find a cross-linguistically valid annotation scheme of natural language morphology and dependency syntax. These efforts have contributed to various layers of UD:

The universal part-of-speech tags **(UPOS)** are based on the Google universal tagset [15], which has been extended and redefined from the original 12 to the current 17 tags; in addition, UD also defines a set of 17 **universal features** that can be used to describe lexical

[1] http://universaldependencies.github.io/docs/

and inflectional properties of words. These features are especially useful for morphologically rich languages. The core feature set is based on Interset [25], an interlingua for morphosyntactic tagsets. It is likely that new features or new feature values will be identified as new languages are added; therefore, the UD format allows additional language-specific features.

At the level of syntactic dependency relations, two related projects have independently tried to define a common scheme applicable to multiple languages: HamleDT (Harmonized Multi-Language Dependency Treebank) [17, 26, 27] comprises 36 languages in its version 3.0; the Universal Dependency Treebank (UDT) [13] has 11 languages in its version 2.0.

The annotation scheme used in UDT is based on Stanford Dependencies (SD) [7, 8], a popular syntactic representation that was first defined for English but later successfully adapted for various other languages. The early releases of HamleDT were based on Prague Dependencies (PD), essentially the annotation scheme of the Prague Dependency Treebank (PDT) [4]. The two projects started to converge when HamleDT 2.0 included a Stanford conversion of its trees, and became the largest collection of treebanks available in PD and SD [17]. Both teams participated in the formulation of the UD annotation guidelines and they are working on converting their data to UD; creators of treebanks for individual languages have joined the effort and either converted their existing data automatically or initiated new manual annotation. The 18 languages included in the UD 1.1 dataset [2] are Basque, Bulgarian, Croatian, Czech, Danish, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Irish, Italian, Persian, Spanish, and Swedish. With increasing coverage and popularity, UD could become a new de-facto standard in the not-so-far future.

The dependency relation inventory and guidelines of UD are based on SD and can be viewed as the next step in the evolution of SD towards a linguistically universal scheme. In the present contribution, we take a closer look at peculiarities of Slavic languages and how they can be handled in UD. We proceed from the first experiences with UD for Czech, and most examples we present come from Czech; we supplement them with examples from the other Slavic languages where appropriate.[2] One very relevant piece of previous work is [12], whose authors proposed several adjustments of SD for Slavic languages. As they based their work on the older (and now obsolete) version of the Stanford format, we will show that some of the issues they address have been solved in UD.

2 Existing treebanks

There are dependency treebanks of various sizes available for at least 8 Slavic languages. The oldest and largest of them is the Prague Dependency Treebank (PDT) of Czech [4]. It inspired annotation efforts for other languages, and about ten other languages have treebanks whose annotation style is very close to PDT, among them three Slavic languages: Slovak [19], Slovene [9, 11] and Croatian [3, 5]. Research teams in several other countries have created treebanks in different annotation scenarios, namely for Bulgarian [20], Russian [6] and recently also Polish [24]. In addition, the PROIEL project[3] provided syntactically annotated texts in Old Russian and Church Slavonic [10]; a new corpus called

[2] At the time this manuscript was submitted for review, Czech was the only Slavic language whose treebank had been converted to UD; later on, Bulgarian and Croatian were added.

[3] http://proiel.github.io/

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	[sk]	SNK	63,238	994K
Slovene	[sl]	SSJ500K	27,829	500K

Table 1. Dependency treebanks of Slavic languages. We use the ISO 639-1 language codes in brackets when referring to particular languages and treebanks throughout the paper.

TOROT for Old Russian and Church Slavonic has recently been launched in Tromsø.[4] The Russian and Slovak treebanks have no standard distribution channels so far; the other treebanks mentioned above are freely downloadable and available for non-commercial research purposes. Table 1 summarizes the Slavic treebanks and their sizes.

3 Pronouns and determiners

The UPOS tagset includes a tag for determiners, which is a category routinely distinguished in English and in Romance languages, but it is not used in the grammatical tradition of Slavic languages (among others). Determiners encompass definite and indefinite articles (which do not exist in Slavic languages, at least not as independent words), as well as other functional words; in Slavic grammars, these words are covered by the term *pronoun*.

The current definition of the borderline between pronouns and determiners in UD is drawn along syntactic properties, that is, it focuses on the function of the word rather than its form.[5] This principle essentially follows the recommendation of EAGLES (see Sections 8.3.1 of [21] and 6.2.2 of [22]). Pronouns are heads of noun phrases, while determiners are those function words that cannot stand alone but need a head (nominal, pronominal) to form an NP.

Many/`DET` *party-goers prefer wine to beer.*

Many/`PRON` *disagreed to the leader's speech.*

While this general guideline may look easy to apply at first glance, the matter is complicated by ellipsis. Consider the sentence

Moje auto je větší než tvoje. "My car is bigger than yours."

In contrast to English, Slavic languages do not use different word forms for self-standing possessive pronouns *(yours)* and for possessive determiners *(your)*. It is natural to view the sentence as an elliptical structure with deleted second instance of *auto*: *Moje auto je větší než tvoje auto.* "My car is bigger than your car." Therefore we propose for Slavic languages to classify all possessive pro-forms as personal possessive determiners. That is, their tag will be `DET` and their features will include `Poss=Yes|PronType=Prs`.

[4] http://site.uit.no/slavhistcorp/files/2015/04/Eckhoff.pdf

[5] There is an ongoing discussion in the UD community whether the definition can be modified and based more on lexical than on functional criteria.

Interrogative, relative, indefinite, negative and demonstrative pro-forms can be divided to those that never behave like determiners ([cs] *kdo, co, někdo, něco, nikdo, nic*) and those that could be determiners or pronouns *(jaký, který, čí, nějaký, některý, něčí, každý, žádný)*. The words from the latter group inflect similarly to adjectives; we may thus be tempted to classify them as determiners without looking at their context (if they appear without a noun, we would explain it by ellipsis). Unfortunately this analysis would be wrong at least for some occurrences of relative forms, which cannot be elliptic:

Muž, kterého jsem vám ukázal "The man whom (which) I showed you" cannot be expanded to *Muž, kterého *muže jsem vám ukázal* and the pronoun *kterého* cannot be attached to *muž* because *muž* is outside the relative clause in which the pronoun acts as the direct object.

For other pro-forms it is not clear whether they should be analyzed as elliptic. The Czech pronoun *každý* "every" occurs 1023 times in PDT and 76% of the occurrences are attributive (dependency labeled `Atr`),[6] which suggests they should be tagged `DET`. However, 24% occurrences independent of nouns seem quite a lot to get along with postulating an invisible deleted noun. A related word *všechen* "all" is even less pronounced: 64% attributive and 36% non-attributive.

Based on this evidence, we propose that the ellipsis explanation, used for possessive determiners, is not extended to the other categories of pro-forms. Instead, the syntactic context should be consulted. If the word modifies a nominal and if there is morphological agreement, then it is a determiner; otherwise it is a pronoun.

4 Numerals and quantifiers

The morphological and syntactic behavior of Czech numerals is a complex matter. Small cardinal numerals *jeden* "one", *dva* "two", *tři* "three" and *čtyři* "four" agree with the counted noun in case (*jeden* also agrees in gender and number; *dva* also agrees in gender). They behave as if they modify the counted noun; they are similar to adjectives in this respect. Examples:

- ***Jeden** muž spal, **dva** muži hráli karty.* "One man slept, two men played cards."
- ***Jedna** žena spala, **dvě** ženy hrály karty.* "One woman slept, two women played cards."
- ***Jedno** kotě spalo, **dvě** koťata si hrála.* "One kitten slept, two kittens played."

In PDT, these numerals are attached to their counted nouns as `Atr` (attribute). UD will use the same structure, only the dependency will be labeled `nummod` (Figure 1).

Larger cardinals behave differently. They require that the counted noun be in the genitive case; this indicates that they actually govern the noun. Such constructions are parallel to nouns modified by other noun phrases in genitive. The whole phrase (numeral + counted

[6] This is just an approximation. In addition to the `Atr` label, we should also require that the determiner agrees with the modified noun in gender, number and case, and possibly also that it occurs before the noun. That way we would exclude genitive modifications such as *nabídka všech* "the offer by all".

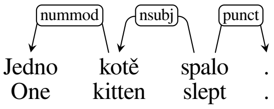

Fig. 1. One kitten slept.

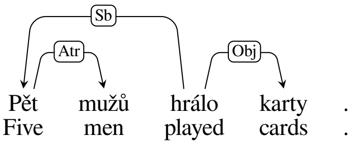

Fig. 2. Prague analysis of the numeral *pět* in nominative.

noun) behaves as a noun phrase in neuter gender and singular number (which is important for subject-verb agreement).

- ***Pět*** *mužů hrálo karty.* “Five men played cards.”
- *Skupina* ***mužů*** *hrála karty.* “A group of men played cards.”

In PDT, these numerals are analyzed as heads of the counted nouns, which are attached to the numeral as `Atr`, see Figure 2.

There are both advantages and drawbacks to this solution. On the one hand, it reflects well the agreement in case, gender and number. On the other hand, it is confusing that there are two different analyses of counted noun constructions, depending on the numeric value. Moreover, the numeral does not govern the noun in all morphological cases, as shown in Table 2.

Phrase Case	Example	Numeral Case	Noun Case
Nom	pět mužů	Nom	Gen
Gen	pěti mužů	Gen	Gen
Dat	pěti mužům	Dat	Dat
Acc	pět mužů	Acc	Gen
Voc	pět mužů	Voc	Gen
Loc	pěti mužích	Loc	Loc
Ins	pěti muži	Ins	Ins

Table 2. The morphological case of a counted phrase with a high-value numeral (first column) and the consequences for the case of the parts (note that these numerals have only two distinct morphological forms, resulting in homonymy). The example phrase is *pět mužů* “five men”.

We can say that the noun has the case of the whole phrase if it is dative, locative or instrumental. The numeral then agrees with the noun in case. The numeral forces the noun to the genitive case if the whole phrase is nominative, accusative or vocative (but the vocative usage is rather hypothetical). In genitive, the noun and the numeral agree with

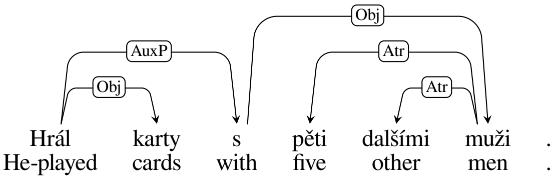

Fig. 3. Prague analysis of the numeral *pět* in instrumental.

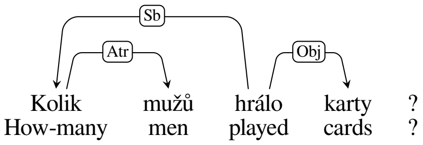

Fig. 4. Prague analysis of the quantifier *kolik* in nominative.

each other; but note that the numeral uses its inflected form, as in the other cases where it agrees with the noun.

In PDT, the genitive, dative, locative and instrumental cases are analyzed in parallel to the low-value numerals, i.e. the noun governs the numeral, see Figure 3.

Pronominal quantifiers behave as high-value numerals and govern the quantifed nouns:

- ***Kolik*** *mužů hrálo karty?* "How many men played cards?"
- ***Několik (mnoho, málo)*** *mužů hrálo karty.* "Several (many, few) men played cards."
- ***Tolik*** *mužů hrát karty jsem ještě neviděl.* "I have never seen so many men playing cards."

For Universal Dependencies we suggest to use the same tree shape for all the examples mentioned above. The counted noun will always be the head, and the numeral or quantifier will depend on it. Thus the structure will be parallel among similar phrases within one language, and also with the universal dependencies in non-Slavic languages. However, we use the UD mechanism of language-specific extended labels to preserve information about who governs the morphological case. There are four labels used and they are based on two UD labels: `nummod` and `det` (Table 3).

	Numeric	**Pronominal**
Noun governs	nummod	det:nummod
Numeral governs	nummod:gov	det:numgov

Table 3. Proposed language-specific dependency relation labels that distinguish quantifiers from other determiners, as well as the situations where the quantifier governs the case of the noun, from the situations where the quantifier agrees with the noun.

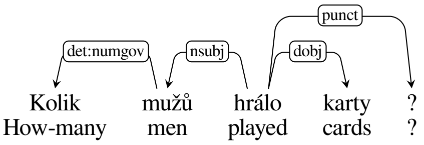

Fig. 5. UD analysis of a governing interrogative quantifier.

5 A verb or not a verb?

Verbal nouns ([cs] *čtení*, [ru] *чтение* "reading") are tagged as nouns, not as verbs. But even then they may have the feature `VerbForm=Ger` to distinguish them from other nouns. (Note that the `VerbForm` feature in UD is actually not constrained to verbs.)

The active (past) participles should always be verbs (these are the forms ending in *-l, -la, -lo* etc.) Note however that occasionally there are derived adjectives with the long adjectival ending, cf. [cs] *zkrachovalý* "bankrupt". These are tagged as adjectives, not as verbs. Passive participles and participial adjectives are told apart in a similar fashion. If the word is used as a modifier of a noun, it should be adjective. If it is used to form the periphrastic passive, it should be verb. This boundary differs across Slavic languages, cf.

[cs] *Město bylo* ***založeno***/`VERB` *Karlem IV.* "The city was founded by Charles IV."

[cs] *Město* ***založené***/`ADJ` *Karlem IV. vyhořelo.* "The city founded by Charles IV has burned down."

[sk] *Mesto bolo* ***založené***/`VERB|ADJ?` *Karolom IV.* "The city was founded by Charles IV."

[sk] *Mesto* ***založené***/`ADJ|VERB?` *Karolom IV. vyhorelo.* "The city founded by Charles IV has burned down."

In any case, all these word forms should also have the feature `VerbForm=Part`, regardless whether their main tag is `VERB` or `ADJ`.

Transgressives (adverbial participles) such as [cs] *pomáhajíc* "helping" or [ru] *будучи* "being" should have the feature `VerbForm=Trans` and the main tag `VERB`. They may also have the `Tense` feature to distinguish present and past transgressives.

6 Auxiliary verbs and modal verbs

Local equivalents of the verb *to be* are the most frequent Slavic auxiliaries, used to create periphrastic past, passive or conditional. The same verb can also be used as non-auxiliary (copula or main verb).

Some languages (e.g. Croatian) have a second auxiliary, *htjeti*, used to form the future tense. In northern Slavic languages the future is also formed using the verb *to be*.

In contrast to the Universal Dependencies applied to English and other Germanic languages, we do not recommend treating modal verbs as auxiliaries. Modal verbs are a subset of verbs that take an infinitive of another verb as complement: [cs] *můžu přijít*, [ru] *ты можешь взять / ty možeš' vzjat'*, [bg] *може да бъде избиран / može da băde izbiran*. The morphological paradigms of Slavic modal verbs are slightly restricted but not as much

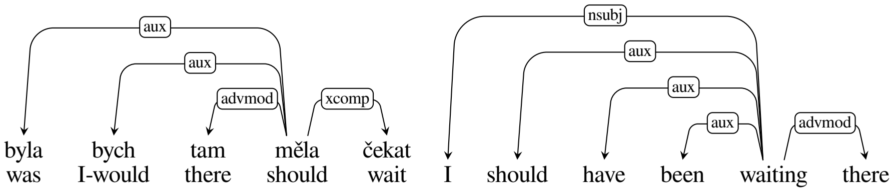

Fig. 6. Combination of modal and auxiliary verbs in Czech and English. English modals are treated as auxiliaries, Czech modals are treated as main verbs.

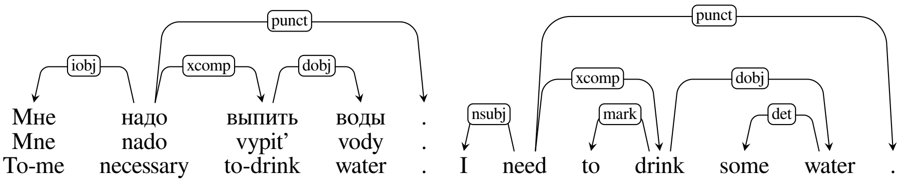

Fig. 7. Modal predicative adverb in Russian and its English translation: *I need to drink some water.*

as in English. They do not form passive participles[7] and most of them also do not have imperative forms. The set is not identical to English. For instance, the Czech verb *chtít* "to want", if used with infinitive and not with a direct object, counts as a modal verb, while its English equivalent does not. There is not much to be gained from treating the modal verbs in the same way as the auxiliary *to be*. It seems more natural to keep the modal dependency structures parallel to those of phase verbs and verbs of control, which also take an infinitival argument. That is, the infinitive will be attached to the modal verb as `xcomp`: see Figure 6.

There is usually just one modal verb to one content verb. However, two modal verbs may co-occur even if it is very rare: [cs] *bude muset chtít pracovat* "he will have to want to work". Treating modals as content verbs has the advantage of capturing scope and hierarchy between the two modals in this example. Furthermore we also want to be able to capture the scope of negation and other modifiers: [cs] *nemohl jsem přijít* "I was not able to come", *mohl jsem nepřijít* "I was able not to come" and *nemohl jsem nepřijít* "I was not able/allowed not to come" are three semantically different expressions.

In addition to modal verbs, modality can also be expressed by modal adverbs, adjectives or nouns. In some cases they are derived from the same roots as modal verbs. These non-verbal modal expressions are particularly pervasive in Russian but other languages have them as well. Again, analyzing modal verbs as content words results in annotation that is parallel to the annotation of non-verbal modal expressions (Figure 7).

7 Reflexive pronouns and verbs

Most of the time the reflexive pronoun is attached to a verb. In the case of transitive verbs, the reflexive pronoun is just another form of object (labeled `dobj` or `iobj`). The test is here

[7] But note that some of them have homonyms that are not used modally and that can form the passive.

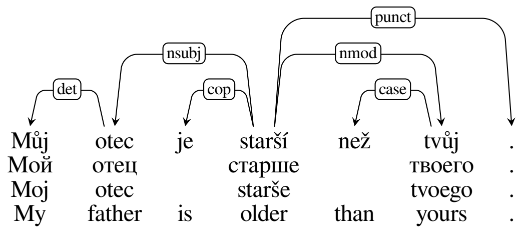

Fig. 8. Comparison with a nominal [cs, ru, ru, en]. Note that Russian omits both the comparative conjunction and the copula.

whether it can be substituted with a normal personal pronoun. If it cannot be substituted, then we are dealing with an inherently reflexive verb ([cs] *smát se* “laugh”). We cannot use an object relation for these reflexives; we suggest to use a language-specific extension of the UD label `expl` (expletive) between the verb and the pronoun: `expl:reflex`.[8] Finally, the reflexive pronoun may also be used to form the so-called reflexive passive ([cs] *to se snadno řekne* “it is said easily (= easier than done)”). The language-specific label `auxpass:reflex` should be used in this case.

Note that in Russian by convention the verb is written together with the reflexive element as one word (*смеяться / smejat’sja* “laugh”). The general UD approach is to cut off clitics (split the token into two **syntactic words**). An often cited example from Spanish is *vámonos* “let’s go” that should be split to *vamos nos,* lit. *go us,* and each part analyzed as a separate word. This approach could be ported to Russian in cases where the clitic *-ся / -sja* can be substituted by an irreflexive pronominal object (*изменять+ся / izmenjat’+sja* “change oneself” would be parallel to *изменять его / izmenjat’ ego* “change him”), and for reflexive passives. However, it does not seem a good idea to extend this approach to inherently reflexive verbs such as *смеяться / smejat’sja*, where the reflexive morpheme does not have its own syntactic function.

8 Comparative constructions

The UD guideline for comparisons is that the comparative complement is attached to the adjective or adverb that denotes the feature being compared. If the complement is a clause, the relation is labeled `advcl`. If it is a bare nominal, it is labeled `nmod`. Some Slavic languages use a comparative conjunction parallel to English *than*: [cs] *Můj otec je starší než tvůj.* “My father is older than yours.” [cs] *Ten hotel je větší, než jsme čekali.* “The hotel is bigger than we expected.” In other languages, the conjunction is not used and the complement is in genitive: [ru] *Мой отец старше твоего. / Moj otec starše tvoego.* “My father is older than yours.” Some Slavic languages use periphrastic comparative of adjectives while others largely prefer the morphological comparative. See Figures 8 and 9 for illustration.

[8] In the Czech and Croatian UD 1.1 data, we used variants of the `compound` relation to express that the two tokens actually form one lexeme. This was revised at a UD workshop in August 2015. In order to make the data more similar to other languages (including Bulgarian), we accepted the `expl`(etive)-based solution.

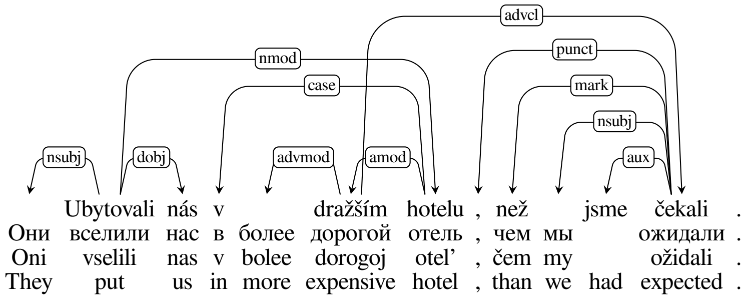
advcl
nmod
punct
case
mark
nsubj
nsubj
dobj
advmod
amod
aux
Ubytovali nás v dražším hotelu , než jsme čekali .
Они вселили нас в более дорогой отель , чем мы ожидали .
Oni vselili nas v bolee dorogoj otel’ , čem my ožidali .
They put us in more expensive hotel , than we had expected .

Fig. 9. Comparison with a clause [cs, ru, ru, en]. Note that Czech uses morphology to form a comparative adjective, while Russian and English form it periphrastically. Also note that Czech is a pro-drop language and omits the subjects.

9 Conclusion

We briefly introduced the concept of Universal Dependencies and listed a number of morphological and syntactic phenomena that occur in Slavic languages and their treatment in UD may not be apparent or straightforward. For each of the issues we discussed its context and proposed how it should be treated in UD. Even though in theory the UD mechanism of language-specific extensions enables treating them differently for different Slavic languages, it would go against the general spirit of UD. We argue that most of these features apply (even if with some variation) in most Slavic languages and thus they should be treated in all these languages in a unified way.

10 Acknowledgements

The author is grateful for the valuable comments provided by the anonymous reviewers.

The work on this project was supported by the Czech Science Foundation grant no. GA15-10472S. This 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).

References

[1] Agić, Ž. and Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of BSNLP/RANLP 2015 (in press)*.

[2] Agić, Ž., Aranzabe, M. J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A. T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., Lynn, T., Manning, C., Martínez, H. A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.1, http://hdl.handle.net/11234/LRT-1478. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

[3] Agić, Ž. and Ljubešić, N. (2014). The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1724–1727, Reykjavík, Iceland. European Language Resources Association (ELRA).

[4] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, v. (2013). Prague dependency treebank 3.0, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

[5] Berović, D., Agić, Ž., and Tadić, M. (2012). Croatian dependency treebank: Recent development and initial experiments. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1902–1906, İstanbul, Turkey. European Language Resources Association (ELRA).

[6] 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.

[7] de Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual.

[8] de Marneffe, M.-C., Silveira, N., Dozat, T., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavík, Iceland. European Language Resources Association (ELRA).

[9] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., and Žele, A. (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).

[10] Haug, D. T. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. In Choukri, K., editor, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*.

[11] Krek, S., Erjavec, T., Dobrovoljc, K., Holz, N., Ledinek, N., and Može, S. (2015). Training corpus – dependency treebank annotation, http://eng.slovenscina.eu/tehnologije/ucni-korpus.

[12] Marszałek-Kowalewska, K., Zaretskaya, A., and Souček, M. (2014). Stanford typed dependencies: Slavic languages application. In *PolTAL 2014, LNAI 8686*, pages 151–163. Springer International Publishing Switzerland.

[13] 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.

[14] Nivre, J. (2015). Towards a universal grammar for natural language processing. In *CICLing 2015: 16th International Conference on Computational Linguistics and Intelligent Text Processing*, Cairo, Egypt. Springer International Publishing Switzerland.

[15] Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 2089–2096, İstanbul, Turkey. European Language Resources Association (ELRA).

[16] Ramasamy, L. (2014). *Parsing under-resourced languages: Cross-lingual transfer strategies for Indian languages*. PhD thesis, Charles University in Prague, Praha, Czechia.

[17] Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., and Mariani, J., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, Reykjavík, Iceland. European Language Resources Association.

[18] Silveira, N. and Manning, C. (2015). Does Universal Dependencies need a parsing representation? An investigation of English. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 310–319, Uppsala, Sweden.

[19] Šimková, M. and Garabík, R. (2006). Sintaksičeskaja razmetka v slovackom nacional'nom korpuse. In *Trudy meždunarodnoj konferencii Korpusnaja lingvistika – 2006*, pages 389–394, Sankt-Peterburg, Russia. St. Petersburg University Press.

[20] Simov, K. and Osenova, P. (2005). Extending the annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 173–184, Barcelona.

[21] Teufel, S. (1996a). ELM-DE: EAGLES specifications for German morphosyntax. Lexical specification and classification guidelines, technical report EAG-CLWG-ELMDE/F.

[22] Teufel, S. (1996b). ELM-EN: EAGLES specifications for English morphosyntax. Lexical specification and classification guidelines, technical report EAG-CLWG-ELMEN/F.

[23] Tiedemann, J. (2015). Cross-lingual dependency parsing with Universal Dependencies and predicted pos labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349, Uppsala, Sweden.

[24] Wróblewska, A. and Woliński, M. (2012). Preliminary experiments in Polish dependency parsing. In Bouvry, P., Kłopotek, M., Leprévost, F., Marciniak, M., Mykowiecka, A., and Rybiński, H., editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 279–292. Springer Berlin Heidelberg.

[25] 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 Sixth International Language Resources and Evaluation Conference, LREC 2008*, pages 28–30, Marrakech, Morocco. European Language Resources Association (ELRA).

[26] 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.

[27] Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). Hamledt: To parse or not to parse? In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2735–2741, İstanbul, Turkey. European Language Resources Association (ELRA).