sciendo
DOI 10.2478/jazcas-2023-0043

THE *INTERCORP* PARALLEL CORPUS WITH A UNIFORM ANNOTATION FOR ALL LANGUAGES

ALEXANDR ROSEN
Institute of the Czech National Corpus, Faculty of Arts,
Charles University, Prague, Czech Republic

ROSEN, Alexandr: The InterCorp Parallel Corpus with a Uniform Annotation for All Languages. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 254 – 265.

Abstract: Recently, the language-specific morphosyntactic annotation of *InterCorp*, a large multilingual parallel corpus, has been replaced by the language-uniform morphosyntactic and syntactic annotation following the guidelines of the Universal Dependencies project. Because the corpus is used predominantly by human users via a token-based concordancer, the CONLL-U format produced by the *UDPipe* parser has been extended by attributes such as lemma of the token's syntactic head or morphosyntactic categories of the content verb's auxiliary. We conclude that despite some theoretical and practical issues, the new annotation is a promising solution to the issue of mutually incompatible tagsets within a single corpus.

Keywords: parallel corpus, Universal Dependencies, multilinguality, syntactic annotation, language-universal categories

1 INTRODUCTION

InterCorp[1] is a multilingual parallel corpus centred around Czech (Čermák – Rosen 2012; Rosen 2016). Since 2009 the corpus has been available with language-specific morphosyntactic annotation. In a recent release of *InterCorp* (v.13ud), a uniform annotation scheme was introduced, based on the guidelines of the *Universal Dependencies* project[2] (UD, de Marneffe et al. 2021). Texts in the corpus were tagged and parsed by *UDPipe*,[3] one of the UD tools (Straka 2018): in addition to morphosyntactic categories, the annotation offers also syntactic functions and structure.

Most users access *InterCorp* via *KonText*,[4] the standard token-based search interface of the Czech National Corpus.[5] To make the UD annotation more useful in *KonText*, the

[1] https://wiki.korpus.cz/doku.php/en:cnk:intercorp
[2] https://universaldependencies.org/
[3] https://ufal.mff.cuni.cz/udpipe/2
[4] https://www.korpus.cz/kontext/
[5] https://www.korpus.cz/

output of *UDPipe* (in the CONLL-U format)[6] has been extended by other attributes. Such attributes help to query the corpus and to generate statistics more efficiently.

After presenting *InterCorp* and its language-specific linguistic annotation in Section 2, we describe the specifics of the UD annotation as it was modified and extended for the use in *KonText* in Section 3. In Section 4 we explain how the annotation can be used in corpus queries. Finally, we point out some issues, sketch plans and conclude in Section 5.

2 INTERCORP

2.1 The contents

InterCorp consists of "Core", including mainly fiction, and "Collections": political commentaries by *Project Syndicate*[7] and *VoxEurop*,[8] legal texts from the *Acquis Communautaire* corpus,[9] proceedings of the European Parliament from the *Europarl* corpus,[10] film subtitles from the *Open Subtitles* database[11] and translations of the Bible, all equipped with bibliographical data.

Like the *Universal Dependencies* project, *InterCorp* is the result of collaborative efforts: academic staff and students of linguistic departments, mostly from Charles University's Faculty of Arts, are responsible for the choice of texts in the Core and much of their processing. Texts in the Core are also proofread for typos, sentence segmentation and correct alignment. Collections are not checked that carefully, and – moreover – they do not retain all texts from the original source, especially those that do not have any Czech counterpart.

For every text there is a single Czech version (original or translation), aligned with one or more foreign-language versions. In release 13, the Core numbers 328 mil. words in foreign texts and 114 mil. in Czech texts. For Collections, it is 1,223 mil. words (foreign) and 90 mil. (Czech).[12] Many texts are available only in few languages, which means that the corpus is far from balanced. Fig. 1 shows the share of text types for each language. *InterCorp* v.13ud contains the same texts as *InterCorp* v.13, the difference is in the UD-based linguistic annotation. Out of a total of 41 languages, 36 are annotated by *UDPipe.*[13]

[6] https://universaldependencies.org/format.html

[7] http://www.project-syndicate.org/

[8] http://www.voxeurop.eu/

[9] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[10] http://www.statmt.org/europarl/

[11] http://www.opensubtitles.org/

[12] At the time of writing, *InterCorp* v.16 is close to release. It is planned as the last version with language-specific tags, to be followed by UD-annotated versions.

[13] The remaining 5 languages were not supported by *UDPipe* at the time v.13ud was getting ready for release. In this sense, the title (… FOR ALL LANGUAGES) is a misnomer, but we hope to remedy this in a future release.

2.2 Language-specific linguistic annotation

For linguistic annotation of a large multilingual corpus an opportunistic strategy of adopting available tools and annotation schemes is the only realistic option. Before UD had reached a practically useful stage, the only reasonable solutions were language specific. Tab.1 shows a sample result. The notational diversity may obscure the fact that many of the seemingly corresponding labels have mismatching denotations, due to different assumptions behind the designs of the tagsets rather than to cross-lingual differences. For example, the English tag IN is used both for prepositions and subordinating conjunctions, while all the other languages make this distinction. Or the Czech equivalent of a determiner is tagged as demonstrative pronoun (PD), undistinguished between attributive and substantive use, whereas its Polish counterpart is tagged as an adjective because it exhibits adjectival declension.

Incompatible tagsets can be mapped onto an interlingual taxonomy, as proposed e.g. by Rosen (2014), but once a uniform annotation scheme and the corresponding tools are available, a more viable solution is to apply language-universal annotation to all texts from scratch.

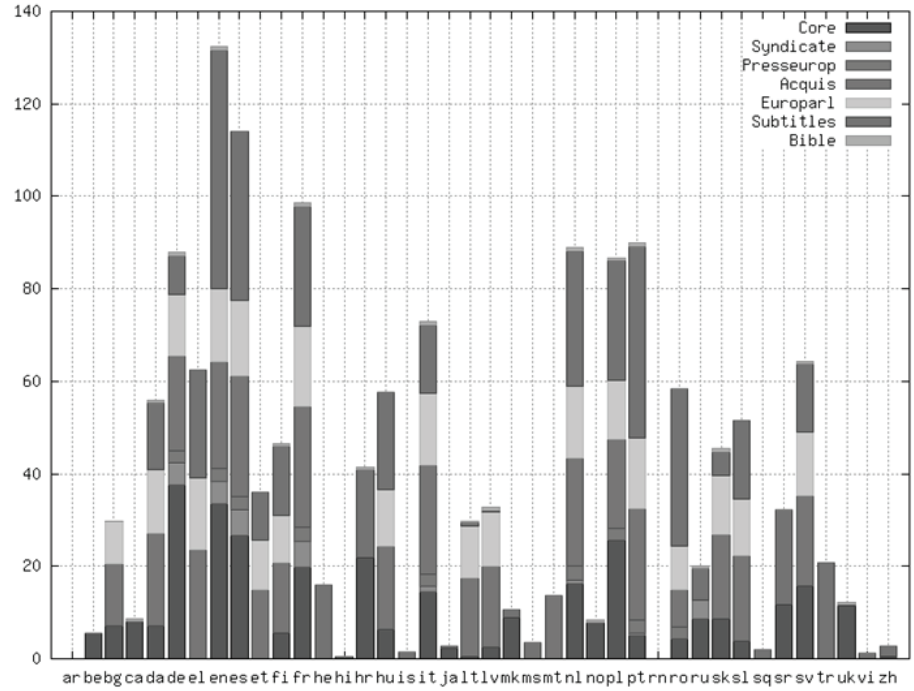


Fig. 1. *InterCorp v13*: languages and text types in mil. words[14]

[14] Language codes are given in ISO 639-1 except for Romany, represented as rn.

Lang	Preposition	Determiner	Adjective	Noun
bg	R	Pde-os-n	Ansi	Ncnsi
cs	RR-6	PDXP6	AAFP6---3A	NNFP6---A
de	APPR	ART	ADJA	NN
en	IN	DT	JJS	NNS
es	PREP	ART	NC	ADJ
et	P--s3		A-p-s3	Nc-s3
fr	PRP	DET:ART	ADJ	NOM
it	PRE	PRO:demo	NOM	ADJ
no	prep	det	adj	subst
pl	prep:loc:nwok	adj:sg:loc:m3:pos	adj:sg:loc:m3:pos	subst:sg:loc:m3
pt	SPS	DA0	NCFS	AQ0
ru	Sp-l	P--pl	Afp-plf	Ncmpln
sk	Eu6	PFip6	AAip6z	SSip6
sl	Sl	Pd-nsg	Agpfsg	Ncnsl

Tab. 1. A phrase such as *in the remotest suburbs*, annotated by language-specific tags[15]

2.3 *Universal Dependencies* as the obvious choice for consistent linguistic annotation

Language-universal annotation moves a part of the burden of enforcing consistency from the corpus onto the annotation scheme. As a result, the scheme and its application face the linguistic diversity of multilingual data.

Language-universal annotation assumes a single set of categories for all languages. It seems that this is what UD does. However, de Marneffe et al. (2021) and the Introduction to the UD Guidelines[16] describe the design of UD as a "very subtle compromise" between six competing criteria, only one of which concerns cross-lingual universals: "UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families." Yet, according to the guidelines, each language is described by the single universal set of categories. Only when they are too coarse, an additional language-specific distinction can be used. This is at odds at with the mainstream typological literature (Croft et al. 2017; Haspelmath 2007; Haspelmath 2010). The compromise is therefore less subtle than presented. However, given the practical goals of the project, the sacrifice of some typological neutrality seems to be justified.

As a more down-to-earth issue, some UD treebanks do not fully comply with the guidelines, either due to specific (mis)interpretation of the guidelines or to failures in applying the guidelines consistently during annotation. The tools trained on the treebanks also rely on the size of the treebanks, their domains and annotation

[15] Tags such as PDXP6 are truncated for the sake of brevity.

[16] https://universaldependencies.org/introduction.html

quality. This is the main reason why some results may be disappointing when compared with those of language-specific tools, often trained on larger data without syntactic annotation.[17] However, both the UD treebanks and the tools have seen significant improvements since the first UD-annotated release of *InterCorp*. An experiment with recent versions of *UDPipe* and Czech treebanks promises a rise in success rates by almost 10 points.[18]

The following section refer to the annotation guidelines on the UD project website (UD Guidelines[19]), including a detailed description of word types (Universal POS tags[20]), morphological categories (Universal features[21]) and syntactic functions (Universal Dependency Relations[22]).

3 THE HAPPY RELATIONSHIP BETWEEN *INTERCORP* AND *UNIVERSAL DEPENDENCIES*

3.1 Annotation adopted from the UD scheme:

- The UD word class and UD morphological categories are listed separately as values of the `upos`[23] and `feats`[24] attributes. Language-specific tags are values of the `xpos` attribute.
- Each word is assigned its syntactic function (`deprel` – see Section 3.5) and a pointer to its syntactic head in the dependency tree (`head`).

3.2 Modification of the UD annotation scheme in *InterCorp* v.13ud:

- Some common categories from the `feats` list are listed also as regular attributes: `case`, `number`, `gender`, `person`.
- Fused forms are represented as single tokens, with their components still accessible. For more details see Section 3.4 below.
- Each word is also annotated with references to important properties of its head (lemma, part of speech and morphological categories), see Section 3.6.

---

[17] In *InterCorp* v.13ud, annotation errors can be found even in the morphosyntactic annotation of Czech, a language with large and high-quality UD treebanks. E.g. a syncretic adjectival form ending in *–ní* after the preposition *na* 'on' (taking accusative or locative), does not agree in case (accusative or locative) with its immediately following noun head in 259 hits, (1 ipm), as in *na východní hranici* 'on the eastern border'. The same sequence is tagged correctly in 75,591 hits (292 ipm), which means that only 0.34% of such constructions are mistagged. However, there is no mistagged constructions like this in v.13.

[18] Daniel Zeman, p.c.

[19] https://universaldependencies.org/guidelines.html

[20] https://universaldependencies.org/u/pos/index.html

[21] https://universaldependencies.org/u/feat/index.html

[22] https://universaldependencies.org/u/dep/index.html

[23] For a list of the 17 UD word classes see also UD Parts of speech (in the *InterCorp* wiki (https://wiki.korpus.cz/doku.php/en:pojmy:ud#parts_of_speech).

[24] See Section 3.3 below.

- If a content word occurs with a function word (preposition, auxiliary verb, copula, subordinate conjunction, determiner, classifier), the content word includes some properties of the function word (see Section 3.7).[25]
- Annotations differ between languages in the number of categorial attributes and in links to function words.[26]

3.3 Morphological categories

Morphological categories are embedded under the `feats` attribute as a list of `<category_name>=<category_value>` pairs, separated by "|", as in (1).[27] Their choice and values are determined by a part of speech and language, but identical or comparable categories and their values have the same name in all languages.[28]

(1) `feats="Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing"`

3.4 Multi-part tokens

Some tokens, *fused words* in the UD parlance, consist of multiple parts, corresponding to different nodes in the syntactic structure. In English this concerns contractions such as *isn't* or *cannot*. Their orthographic form is preserved, the individual parts are separated only in the annotation (e.g. in `lemma`), with "|" as the separator, resulting in `lemma="be|not"`.

A part of the fused token, such as *n't* in English, may have a different form when occurring as a separate word: *not*. The Czech 2nd person singular auxiliary enclitic *s*, fused with the reflexive *se* into *ses*, corresponds to *jsi*. Both variants (*n't* and *not*, *s* and *jsi*) are represented in the annotation of *isn't* or *ses*: the `iword` attribute shows the original form – `iword="is|n't"` or `iword="se|s"`, while the `sword` attribute shows the unabbreviated version – `sword="is|not"` or `sword="se|jsi"`.[29] The example *Oč mu vlastně jde?* 'What's he really up to?' in Fig. 2 (from the *KonText* tree display of the concordance) concerns a fused form *oč*, consisting of the preposition *o* 'about' and an abbreviated form of the pronoun *co* 'what' preposition. The full form of the pronoun appears as the `sword` value after a click on the node for *č*.

[25] In UD, copula is a function word dependent of the nominal predicate. In *This was not a good moment in the history of English cuisine* the noun *moment* is the root node of the whole tree.

[26] See Description of the list of attributes in the *InterCorp* wiki (https://wiki.korpus.cz/doku.php/en:pojmy:ud#description_of_the_list_of_attributes).

[27] As a reviewer observed, the double use of the "=" operator is unfortunate. Within the feats list the operator means strict equality but after `feats` (a CQL attribute) it means membership in the following list of attribute-value pairs. While the latter use is given by CQL, the former is inherited from the CONLL-U format and could be changed in a future release.

[28] See Universal Features or Other categories in the *InterCorp* wiki (https://universaldependencies.org/u/feat/index.html or https://wiki.korpus.cz/doku.php/en:pojmy:ud#other_categories).

[29] It is this unabbreviated form which is available in the output of *UDPipe*.

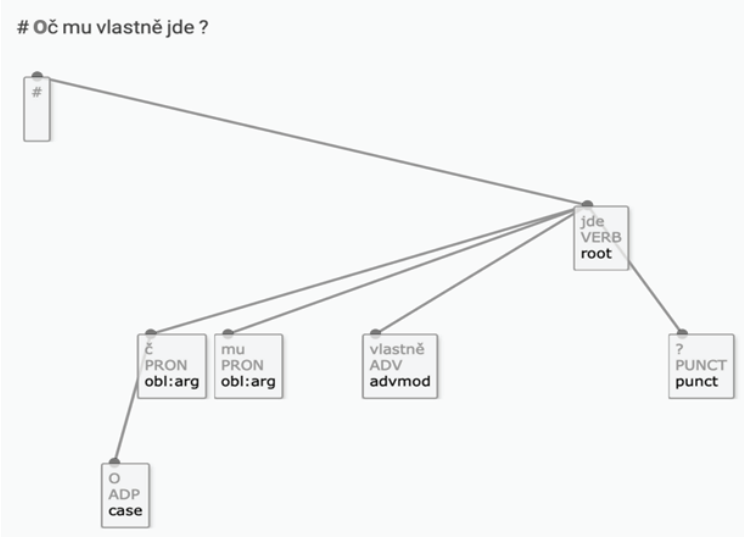


Fig. 2. Syntactic tree including a Czech fused form *oč*, consisting of *o* 'about' + *co* 'what'

3.5 Syntactic functions

Each token specifies its dependency relation (deprel) and a reference to its syntactic governor (head).[30] The 37 relations are cross-classified in terms of their relation to the head (core arguments, non-core dependents, dependents of a noun) and their structural categories (nominals, clauses, modifiers, function words), e.g. iobj (indirect object, a core nominal: *Can I buy* you *a drink?*), advcl (adverbial clause, a non-core clausal dependent: *The country will pay a heavy price* if the president's obsessions prevail for long.), amod (adjectival modifier, a modifier of a noun: *The* sustainable *future of humanity is at stake*.), or cop (copula, a function word as a non-core dependent: *Where's the rest of your luggage?*). Some deprels may have subtypes, e.g. acl:relcl indicates an attribute expressed by a relative clause.

3.6 References to syntactic heads

Each token specifies attributes referring to the token's head (its absolute position – head, or relative to the token – parent). Other attributes of the head include its lemma (p_lemma), POS (p_upos), morphological categories (p_feats), and syntactic function (p_deprel).[31]

[30] See Universal Dependency Relations (https://universaldependencies.org/u/dep/index.html) for an overview and detailed descriptions of syntactic functions used in UD, or Syntactic functions (https://wiki.korpus.cz/doku.php/en:pojmy:ud#syntactic_functions) as they are used in *InterCorp*.

[31] Cf. similar attributes in other syntactically annotated corpora available in the *KonText* browser (https://www.korpus.cz/kontext/query?corpname=syn2020 or https://lindat.mff.cuni.cz/services/kontext/corpora/corplist?keyword=ud).

3.7 References to function words

In UD, function words include auxiliaries, pre- and postpositions, conjunctions, determiners, copulas and classifiers. All function words depend on their content word host. Their types are specified by their deprel attribute.

It is often useful to interpret their properties as features of their content word heads.[32] This is why content words have attributes specifying properties of dependent function words. For example, the lemma of a preposition is shown by the attribute case_lemma, categories of an auxiliary by aux_feats, categories of a copula by cop_feats, word class of a determiner by det_upos, lemma of a marker by mark_lemma. The content word may also include a more detailed specification of the function word's type, e.g. aux_type="pass" or det_type="numgov".[33]

A single content word can govern several function words, e.g. three for *she would have been pleased*. The values, separated by "|", appear concatenated in a single content word attribute. The feats from multiple auxiliaries, concatenated into a single value may include some conflicting values or repetitions. For example, in the sentence *who would have guessed that*, the aux_feats of the content verb *guessed* are composed of the feats of the auxiliary verbs *would* (Mood=Ind|Person=3|Tense=Past|VerbForm=Fin) and *have* (VerbForm=Inf).[34]

3.8 Coordination

In UD, the first conjunct depends on the head of the coordination. Its deprel determines the syntactic function of the whole coordination. The other conjuncts always depend on the first conjunct. Their deprel is specified as conj. Conjunctions depend on the following conjunct with the deprel of cc.

A reference to the "effective head" is used to identify the head of the whole coordination: the e_id attribute refers to the effective head's absolute position, the eparent attribute to its position relative to the token.[35]

---

[32] Cf. the recent addition of auxiliary-related features to the Czech tagset (Jelínek et al. 2022). In addition to the head-related attributes (see Section 3.6), information about the head's function word dependents is sometimes useful but not easily accessible in *KonText.* The solution along the lines of *ibid.* could be an interpretation of *all* dependent function words (e.g. as analytical verb forms or prepositional phrases), stored in the content word head and accessible as another head-related attribute from the head's non-functional daughters.

[33] See https://universaldependencies.org/cs/dep/aux-pass.html, or https://universaldependencies.org/cs/dep/det-numgov.html.

[34] In the next UD-annotated release, categories from different function words concatenated in a single content word's list will be separated by two vertical bars.

[35] One more attribute, useful especially in the annotation of non-initial conjuncts, should specify its "real" deprel, i.e. the syntactic function of the whole coordination, specified in the first conjunct. This attribute (e_deprel) is due to appear in the next UD-annotated release.

4 QUERYING THE UD-ANNOTATED *INTERCORP*

4.1 Basic query

Parts of fused words should not be separated by a space, as in *is n't* or *był bym*, but entered as a whole. The same applies to the value of the `word` attribute in an advanced query. A query for *is* or *n't* will not show concordances including the form *isn't*.

4.2 Advanced query for lemmas and tags

Word class and other categories are listed separately as values of the attributes `upos` and `feats`. The values are identical for all languages. E.g. the proper query for proper names is `[upos="PROPN"]`.

Each of the other categories can be specified separately: the Czech form *radostmi* 'joy-PL-INS' is one of the answers to the query (2). The order of categories in the query is irrelevant. The value of `feats` can also be treated as a string of characters using regular expressions, as in (3). Here the order of categories in the query should correspond to their order in the corpus. Some of the categories in `feats` are listed also outside the list as categorial attributes. The query (4) gives the same result as the two queries above.

(2) `[upos="NOUN" & feats="Gender=Fem" & feats="Number=Plur" & feats="Case=Ins"]`

(3) `[upos="NOUN" & feats=".*Case=Ins.*Gender=Fem.*Number=Plur.*"]`

(4) `[upos="NOUN" & gender="Fem" & case="Ins" & number="Plur"]`

Categorial attributes can also be used to generate frequency lists or in global conditions,[36] e.g. to require morphological agreement of two or more words.[37] Such attributes appear on the light brown background in Attribute list by language[38] or in *KonText* in the lower part of the table shown in `View / Corpus-specific settings...`

For most languages the `xpos` attribute provides a "legacy" language-specific tag, usually identical to the tag used in the previous releases of *InterCorp*. For Czech, the queries (2)–(4) can also be entered as in (5).

(5) `[xpos="NNFP7.*"]`

[36] https://www.sketchengine.eu/documentation/cql-global-conditions/

[37] Note that for technical reasons the names of the categorial attributes are all in lower case, including names such as `VerbForm` (in `feats`), rendered as `verb_form`, or `NumType`, rendered as `num_type`. The attribute values, such as `Fem`, retain the initial upper case character, but are enclosed in double quotes, like other attribute values outside `feats`.

[38] https://wiki.korpus.cz/lib/exe/fetch.php/cnk:intercorp:ud_ic_atributy.pdf

Values of upos and of any category from the feats list can also be entered using the "Insert tag" function of *KonText*. A menu helps to select upos and/or feats categories. The choice is determined by their actual occurrence in the corpus and may reflect incorrect combinations.[39]

4.3 Query for syntactic functions

Syntactic function is specified by the deprel attribute (see Section 3.5 above). E.g. a query meant to show the occurrences of the verb *run* as the head of an adnominal clause is entered as [lemma="run" & deprel="acl"]. Results include examples such as *Everyone of the rabbits was seized by the instinct to run away.* or *Some people have the idea that rabbits spend a good deal of their time running away from foxes.* When querying a deprel that may have a subtype, a possible subtype should be taken into account, i.e. deprel="acl.*" is preferable to deprel="acl".

When the deprel in a query targets a coordinated structure, only the first conjunct is found. Non-initial conjuncts are marked as deprel="conj". To query the "true" deprel of a non-initial conjunct (annotated by deprel="conj"), the p_deprel attribute should be used instead. This attribute shows deprel of the token's head. For example, a query for all indirect objects, including all conjuncts, should use the disjunction operator "|" as in (6).[40] See Section 3.8 above for details.

(6) [deprel="obj" | deprel="conj" & p_deprel="obj"]

4.4 Display of syntactic structure

After clicking on the syntax tree icon preceding a concordance line, the syntactic structure of the sentence is displayed. For each node, the word form, upos and deprel of the token appears. More attributes appear after clicking on the node.

Fused words are split into multiple nodes. After clicking on such a node, its full form (the sword attribute) and the entire token (word) also appear. In the text line above the structure and in the structure, the relevant strings and nodes are highlighted under the cursor in parallel (see Fig. 2 above).

5 CONCLUSION

UD seems to be a good solution for a multilingual corpus when the goal is to avoid heterogeneous annotation. The impossibility of language-universal annotation facing the typological diversity of human languages does not seem to seriously damage the

[39] In the next UD-annotated release, the "Insert tag" function will be extended to cover also syntactic functions (see Section 4.3 below).

[40] We plan to avoid the need to specify queries in such a clumsy way in the new release. See footnote 35 above about the e_deprel attribute. This attribute would duplicate the deprel value of each token, except for non-initial conjuncts, where the value of the first conjunct's deprel would appear.

practical gains of a consistent annotation. It seems that some problems in the use of UD categories for language-specific phenomena are due to the authors' failure to follow the guidelines properly rather than to the inherent impossibility to use the categories cross-linguistically. There is an ample space for improvements in the size of some treebanks, quality of their annotation and the type of texts included. Also, the development of better tools has surely not reached the limits. Together with the progress in the acquisition of larger, more varied and better annotated data, the perspectives are promising, despite the still higher accuracy of some language-specific taggers.

Besides providing a solid foundation for comparative and other cross-lingual studies and applications, UD offers the bonus of syntactic annotation, turning *InterCorp* into a parallel treebank.

ACKNOWLEDGEMENTS

The research is part of the project *Czech National Corpus*, supported by the Ministry of Education, Youth and Sports of the Czech Republic as one of the *Large Research, Development and Innovation Infrastructures* (LM2023044).

References

Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic Typology Meets Universal Dependencies. In M. Dickinson – J. Hajič – S. Kübler – A. Przepiórkowski (eds.): Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), pages 63–75. Bloomington, Indiana University. Accessible at: http://ceur-ws.org/Vol-1779/05croft.pdf.

Čermák, F., and Rosen, A. (2012). The Case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 13(3), pages 411–427.

Haspelmath, M. (2007). Pre-established categories don't exist – Consequences for language description and typology. Linguistic Typology, 11, pages 119–132.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. Language, 86(3), pages 663–687. Accessible at: https://doi.org/10.1353/lan.2010.0021.

Jelínek, T., Petkevič, V., and Skoumalová, H. (2022). Hodnoty slovesných morfologických kategorií v korpusu SYN2020 – atribut verbtag. Časopis pro moderní filologii, 104(1), pages 89–109.

Marneffe de, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pages 255–308. Accessible at: https:doi.org/10.1162/coli_a_00402.

Rosen, A. (2014). A 3D taxonomy of word classes at work. In L. Veselovská – M. Janebová (eds.): Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure, Vol. 4, pages 575–590. Olomouc, Palacký University. Accessible at: http://olinco.upol.cz/assets/olinco-2014-proceedings.pdf.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In E. Gruszczyńska – A. Leńko-Szymańska (eds.): Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora, Vol. 1, pages 21–40. Instytut Lingwistyki Stosowanej. Accessible at: https:doi.org/10.13140/RG.2.1.2808.7444.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207. Accessible at: https:doi.org/10.18653/v1/K18-2020.