A Constructicon for Universal Dependencies

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

RISE Research Institutes of Sweden
Department of Computer Science
joakim.nivre@ri.se

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

We propose to build a *constructicon* for Universal Dependencies (UD), consisting of (i) an inventory of universal linguistic *constructions*; (ii) for each construction, an inventory of common *strategies* for realizing that construction in the world's languages; and (iii) for each construction-strategy pair, a cross-linguistically applicable UD analysis and representative examples from different languages. After providing some background and motivation for the project, we illustrate the idea with a case study of one construction: predicate nominals.

2 Background and Motivation

The UD framework is designed to support cross-linguistically consistent morphosyntactic annotation for the world's languages (Nivre et al., 2016, 2020). A key idea is to bring out similarities (and differences) across languages by maximizing the amount of parallel structures. This is achieved by giving priority to direct syntactic relations between content words, such as verbs, nouns and adjectives. This increases the probability of finding parallel structures across languages, since function words in one language often correspond to morphological inflection (or nothing at all) in other languages. This is illustrated in Figure 1, showing a simplified UD analysis of two equivalent sentences in English and Finnish, which are similar concerning relations connecting verbs and nouns (highlighted in red), but differ in that English uses function words like *the* and *from* to encode grammatical information, while Finnish uses morphological inflection (represented by features like Case=Nom). (For an in-depth discussion of the UD framework, see de Marneffe et al. (2021).)

The UD framework has been claimed to harmonize well with findings from linguistic typology (Croft et al., 2017), but it is limited by the fact that it only annotates overt morphosyntactic *strategies* and not the underlying universal *constructions*, in the terminology of Croft (2022). As a consequence,

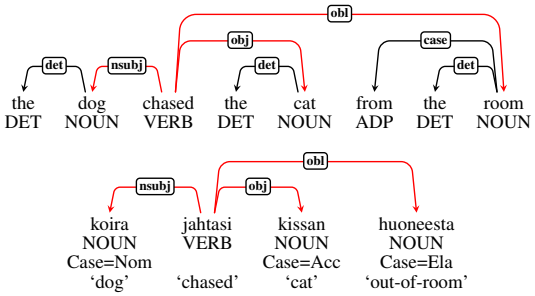

Figure 1: Simplified UD annotation for equivalent sentences in English (top) and Finnish (bottom).

UD fails to capture common constructions across languages when strategies diverge beyond the encoding of grammatical information through morphology or function words. A typical example is the *possession* construction, expressing that A is in possession of B, which can be realized by very different strategies, including those paraphrased below (Croft, 2022):

1. The location strategy: "B is at A"
2. The with-strategy: "A is with B"
3. The topic-strategy: "as for A, B exists"
4. The have-strategy: "A has B"

In such cases, there is nothing in the UD annotation that captures the common construction realized by different strategies across languages.

A constructicon for UD is a way of organizing the annotation guidelines by constructions and strategies. Such a resource can serve a number of complementary purposes in computational linguistics. For the UD project itself, it will provide better support for adding new languages to UD, improve cross-linguistic annotation consistency, and help identify gaps or deficiencies in the existing guidelines. For the wider community interested in linguistic typology, the constructicon can be combined with annotated data sets from UD to construct a fine-grained morphosyntactic typology based on the statistical distribution of strategies across constructions and languages. Such a typology would go beyond the categorical typological classification that has been dominant so far, in

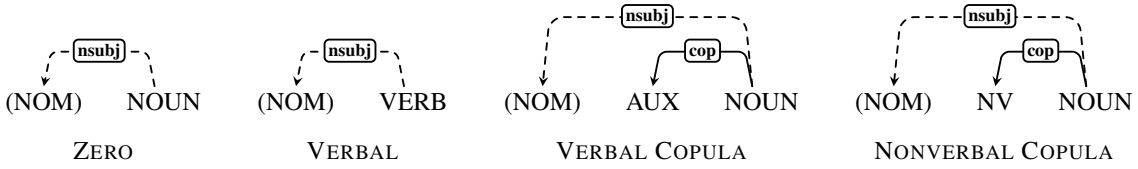

Figure 2: UD annotations for the four strategies used to realize the predicate nominal construction. From left to right: zero, verbal, verbal copula, nonverbal copula. NOM = any nominal part of speech (NOUN, PRON, PROPN); NV = any nonverbal part of speech. Dashed arcs indicate syntactic relations that may be missing in case of pro-drop.

line with recent work on quantitative typology by Futrell et al. (2015), Levshina (2019), Gerdes et al. (2021), Yan and Liu (2023), among others. This classification could be useful not only for linguistic typology but also for typologically informed approaches to NLP, as it would make available more fine-grained language representations that could be leveraged by multilingual NLP models. Finally, a UD constructicon can form the basis for a more systematic evaluation and understanding of such models, complementing more surface-oriented work on probing and targeted syntactic evaluation (Marvin and Linzen, 2018; Goldberg, 2019; Tenney et al., 2019; Hewitt and Manning, 2019; Hu et al., 2020; Kulmizev et al., 2020). An interesting research question in this context is whether multilingual language models learn cross-lingual abstractions at the level of strategies or constructions (or both).

3 Case Study: Predicate Nominals

To illustrate the idea of a UD constructicon, we discuss one universal construction, the *predicate nominal* construction[1] (Croft, 2022), which is used to predicate of some argument that it belongs to a category of objects. A typical example in English is *Sam is a doctor*, which predicates of some individual named *Sam* that he/she belongs to the category of doctors. This is a non-prototypical predication construction, which is realized using at least four different strategies across the world's languages (Croft, 2022):

1. Zero strategy: A noun is used as a predicate without any overt marking or linking element.
2. Verbal strategy: A noun is inflected like a verb when used as a predicate.
3. Verbal copula strategy: A noun is combined with a linking element in the form of a verb.
4. Nonverbal copula strategy: A noun is combined with a linking element that is not a verb.

The four strategies are exemplified in (1–4) below.[2]

(1) aga bawa taleng-duap [Waskia]
my brother policeman
'my brother is a policeman'

(2) ni-ticitl [Classical Nahuatl]
1SG-doctor
'I am a doctor'

(3) elle est médécin [French]
she is doctor
'she is a doctor'

(4) eia la taua [Nakanai]
3SG DEM spirit
'he is a spirit'

A UD constructicon should describe how each of these strategies is to be annotated in UD, as shown in Figure 2. What is common to all strategies is a nominal subject relation (*nsubj*) linking the predicate nominal to its subject, which is itself a nominal (that is, a phrase whose head is of category NOUN, PRON and PROPN). The NOM element is put in brackets and the *nsubj* relation is dashed to indicate that the subject may not be overtly realized in the case of pro-drop. The verbal strategy differs from the zero strategy in that the predicate is assigned the part-of-speech tag VERB in virtue of the verbal inflection. The verbal and nonverbal strategies both have an obligatory linking element, connected to the predicate with a *cop* relation, but differ in the part-of-speech tag assigned to this category (AUX vs. any nonverbal category).

4 Conclusion

To improve the support for cross-linguistically consistent annotation in the UD framework, we propose to build a systematic inventory of constructions and strategies for the world's languages, with concrete annotation guidelines and corpus examples. We call this resource a constructicon for UD.

[1] An alternative name for this construction is *object predication*.

[2] Note that example (2) contains no overt subject because of pro-drop.

References

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75.

Marie de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47:255–308.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling)*, pages 91–100.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: A Journal of General Linguistics*, 6(1).

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR*, abs/1901.05287.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4077–4091.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23:533–572.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of the 5th International Conference on Learning Representations*.

Jianwei Yan and Haitao Liu. 2023. Basic word order typology revisited: A crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*.