Constructions and Strategies in Universal Dependencies

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

Abstract

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology? One way to find out is to investigate whether UD can adequately represent constructions of the world's languages, as described in William Croft's recent book *Morphosyntax*. This paper discusses how such an investigation could be carried out and why it would be useful.

1 Introduction

Universal Dependencies (UD) is a framework for morphosyntactic annotation, designed to be applicable to all human languages and to enable meaningful cross-linguistic comparisons. The two versions of the guidelines are described in Nivre et al. (2016) and Nivre et al. (2020); a longer description of the underlying linguistic theory can be found in de Marneffe et al. (2021); and annotated data for 168 languages[1] can be found together with additional documentation on the UD website.[2]

But can UD really handle the full range of morphosyntactic variation in the world's languages? And is it successful in revealing similarities and differences across these languages in a systematic fashion? One way to approach these questions is to review the UD framework through the lens of linguistic typology. An early attempt to do this can be found in Croft et al. (2017), where the authors review version 1 of the UD guidelines and propose a number of improvements for better alignment with typological research findings, some of which were integrated in version 2 of the guidelines. Since then, William Croft has published the book *Morphosyntax* (Croft, 2022), a comprehensive survey of constructions in the world's languages, which brings together the results of sixty years of research on typology and universals and thus provides an excellent basis for a new and more exhaustive review of the UD framework.

Croft's survey is based on two types of comparative concepts (Haspelmath, 2010; Croft, 2016): *constructions*, which are universal form-function pairings defined solely in terms of their function, and *strategies*, which are non-universal and defined by the pairing of a function with some cross-linguistically identifiable morphosyntactic form. Annotations in UD are not defined in terms of constructions and strategies, but for the framework to be universally applicable it must be possible to annotate all major constructions and strategies in the world's languages. And to support cross-linguistic comparisons, these annotations should ideally reflect systematic correspondances in constructions and strategies across languages. The purpose of this position paper is to motivate a more systematic study of these issues, by showing that we currently do not know to what extent UD satisfies these requirements, and to propose a research program to support this investigation.

The rest of the paper is organized as follows. In Section 2, I give a brief overview of the UD annotation framework, focusing on fundamental design principes; in Section 3, I outline the taxonomy of constructions and strategies in Croft (2022); and in Section 4, I discuss how constructions and strategies are annotated in UD. I conclude that, although the design principles of UD in some respects favor a clear representation of constructions and strategies, the correspondence between the two systems is far from perfect and merits further investigation.

2 The UD Annotation Scheme

The UD annotation scheme assumes that *words* are the basic units of morphosyntax. Words encode grammatical information internally through lexical stems and inflectional processes, but since the nature of these processes varies considerably

[1] UD v2.15, released November 15, 2024.
[2] https://universaldependencies.org

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 419–423
March 3-4, 2025

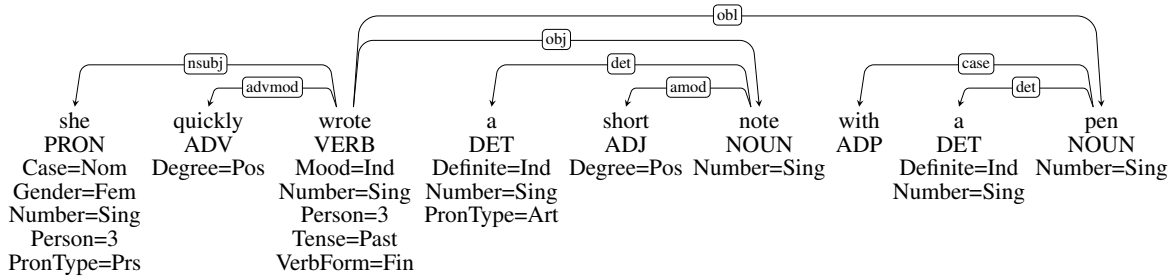

Figure 1: UD annotation of an English sentence.

across languages, there is no attempt to segment words into smaller units like morphs. Instead, the morphological annotation layer in UD combines coarse-grained part-of-speech tags with a rich inventory of morphological features, which together capture the information encoded in words without localizing it to smaller parts.[3]

Words also enter into syntactic relations with other words, and UD assumes that the information encoded in syntactic structure can be captured by a tree-structured representation consisting entirely of binary relations between words. A subset of these relations correspond to what grammarians would call dependency relations – asymmetric relations between a syntactic head and a dependent – but many of the relations that are necessary for a complete syntactic analysis are essentially symmetrical, even though the tree constraint forces one of the words to be (arbitrarily) chosen as the parent node. By way of illustration, Figure 1 shows the UD annotation of an English sentence.[4]

The syntactic analysis in UD assumes that all languages have *nominals*, which are the primary means of referring to entities, and *clauses*, which describe events (including actions and states). Both nominals and clauses can be further refined by *modifiers*, which describe attributes of entities or events. Figure 1 shows a main clause with the predicate *wrote* and three nominals: *she*, *a short note*, and *a pen*; there is also an adverbial modifier *quickly*, modifying the predicate *wrote*, and an adjectival modifier *short*, modifying the noun *note*.

A characteristic property of UD syntax is that it prioritizes direct relations between predicates, nominals and modifiers, rather than relations mediated by function words. Thus, in Figure 1, there is a direct relation from the predicate *wrote* to the noun *pen*, denoting the instrument of writing, while the preposition *with* is essentially treated as a case marker on the noun. This treatment is motivated by the observation that predicates, nominals and modifiers are more likely to be parallel across languages than function words, which often correspond to morphological inflection (or nothing at all) when comparing across many languages.

3 Constructions and Strategies

The most central concept in Croft's framework of morphosyntax is that of a *construction*, which is defined in the following way (Croft, 2022, p. 17):

> **construction:** any pairing of form and function in a language (or any language) used to express a particular combination of semantic content and information packaging

It is worth noting that the functional side of a construction consists of two components, a semantic content and a particular way of packaging the information, also known as a propositional act. This is exemplified in Table 1, which shows constructions defined by different combinations of semantic classes and propositional acts, with the most prototypical constructions being *nominal phrases*, which refer to objects, *adjectival phrases*, which express property modification, and *verbal clauses*, which express action predication.[5]

Constructions at the most abstract level are universal and defined only in terms of function. However, to enable cross-linguistic comparison of constructions also in terms of their form, Croft introduces the notion of a *strategy* (Croft, 2022, p. 19):

> **strategy:** a construction in a language (or any language), used to express a particular combination of semantic content and information pack-

[3]The morphological layer also includes lemmas, which are language-specific and will not be discussed here.

[4]For more information about tags, features, and relations, see https://universaldependencies.org.

[5]The prototypical constructions can be found along the diagonal from top left to bottom right in the first three rows of Table 1.

Semantic Class	Propositional Act		
	Reference	**Modification**	**Predication**
Object	Nominal Phrase **Head:** Noun	Possessive Modifier/Genitive Phrase	Predicate Nominal
Property	Property-Referring Phrase	Adjectival Phrase **Head:** Adjective	Predicate Adjectival
Action	Complement (Clause)	Relative Clause	Verbal Clause **Head:** Verb
All	Referring/Argument Phrase **Head:** Referent Expression	Attributive Phrase **Head:** Modifier	Clause **Head:** Predicate

Table 1: Grammatical constructions for combinations of three basic semantic classes and the three major propositional act (information packaging) functions (adapted from Croft (2022)).

aging (the 'what'), that is further distinguished by certain characteristics of grammatical form that can be defined in a crosslinguistically consistent fashion (the 'how')

To exemplify the notion of strategy, let us consider the *predicate nominal* construction, which is "a clause construction defined by the function of predicating an object concept of a referent – that is, asserting what object category the referent belongs to".[6] Two common strategies for this construction are exemplified in (1) and (2–3).

(1) Иван танцор
Ivan.NOM dancer.NOM
'Ivan is a dancer'

(2) Ivan är dansare
Ivan COP dancer
'Ivan is a dancer'

(3) Ivan is a dancer
Ivan COP a dancer

The Russian example in (1) uses a *zero* strategy (Stassen, 1997), which simply juxtaposes the referring expression Иван with the noun танцор in nominative case expressing the object concept. By contrast, the Swedish and English examples in (2) and (3) both use a *verbal copula* strategy (Stassen, 1997), where predication is mediated by a copula verb. The notion of strategy allows us to abstract over language-specific constructions and say that Swedish and English use the same strategy, while the Russian strategy is different.

4 Constructions and Strategies in UD

How are constructions and strategies represented in UD? At first sight, it may appear that they are not represented at all, because the UD annotation is centered on properties and relations of words. However, as noted in Section 2, the UD scheme systematically distinguishes clauses, nominals and modifiers. For example, a word with an incoming relation labeled *nsubj* must be the head of a nominal phrase, and a word with an incoming relation labeled *advcl* must be the head of a (subordinate) clause. So there is an almost perfect correspondence between the basic structures posited by UD – nominals, modifiers, and clauses – and the three major propositional acts in Croft's framework: reference, modification, and predication.[7] In addition, the UD principle of prioritizing direct relations between predicates, nominals and modifiers often reveals constructional parallelism across languages that use different strategies for a given construction.

To illustrate this, let us return to the predicate nominal construction and consider the UD annotation of (1–3) in Figure 2. All three representations share a structure NOUN $\xrightarrow{\text{nsubj}}$ X, where X can be replaced by any category that can be the head of a referring expression. This captures the fact that the predicate nominal construction involves using a noun as a predicate, which would have been less clear if the copula verb had been treated as the head of the clause in Swedish and English. Moreover, the fact that Swedish and English uses the same strategy is captured by the presence of the structure NOUN $\xrightarrow{\text{cop}}$ AUX, which contrasts with the absence of such a structure in Russian. In general, strategies often correspond to relations involving function words (like the *cop* relation).

The predicate nominal example suggests that UD representations can be decomposed into distinct substructures corresponding to constructions and strategies. Unfortunately, this is not true in

[6] https://comparative-concepts.github.io/cc-database/

[7] The only discrepancy is that Croft's notion of modification is restricted to modification of referring expressions, whereas the UD concept also includes adverbial modifiers and modifiers of modifiers.

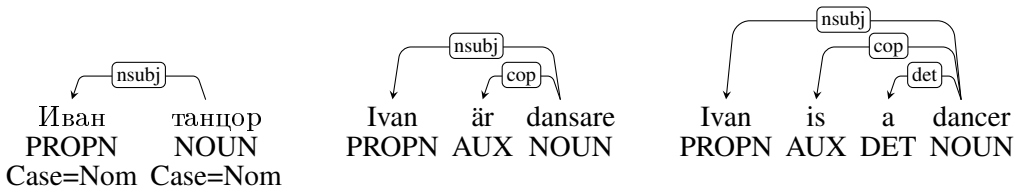

Figure 2: Simplified UD annotation for predicate nominal constructions in Russian, Swedish and English.

Figure 3: Simplified UD annotation for presentational possession constructions in Finnish and English.

the general case. First of all, it is clear that UD representations are more coarse-grained than constructions and strategies, so there will often be a one-to-many mapping from the former to the latter. For example, the substructure that is characteristic of the predicate nominal construction in Figure 2 would also be characteristic of an *equational* construction, as exemplified by *Ivan is the winner*, which in Croft's framework is a distinct construction, even though the two constructions often share strategies through a process known as recruitment.

More importantly, it is not hard to find constructions where the UD representations completely fail to capture constructional parallelism. One example is the *presentational possession* construction, defined as "a presentational information packaging of the possession relation in which a possessum is introduced into the discourse, anchored by the possessor"[8] and exemplified in Figure 3 with examples in Finnish and English. Finnish here uses a *locational possessive* strategy (Stassen, 2009), in which the possessum (*kirja* 'book.NOM') is expressed in a subject phrase, and the possessor (*hänellä* 'her.ADESS') in an oblique (locative) phrase, with a linking copula verb (*on* 'be.3SG.PRES'). By contrast, English uses a *have-possessive* strategy (Stassen, 2009), where the possessor is expressed in a subject phrase (*she*), and the possessum in an object phrase (*a book*), connected by a full transitive verb (*has*). A closer comparison of the examples reveals that the two representations have next to nothing in common, which could capture the common construction, and also that the two strategies in this case involves syntactic relations like *nsubj* and *obj*, which in the predicate nominal example were considered elements of the construction.

[8]https://comparative-concepts.github.io/cc-database/

5 A Research Program for UD

Which of the two cases discussed above is typical? Are UD annotations mostly decomposable into parts corresponding to constructions and strategies, with a few anomalous cases like the presentational possession construction? Or is it the latter that is the norm, and the former the exception? At this point, we simply do not know, and this is the main motivation for proposing a research program that systematically investigates how constructions and strategies can be represented in UD, using the survey in Croft (2022) as a starting point. More precisely, I propose to develop a *constructicon* for UD, consisting of the following components:

- An inventory of universal constructions.
- For each construction, an inventory of common strategies for realizing that construction in the world's languages.
- For each construction-strategy pair, a cross-linguistically valid UD analysis and representative examples from different languages.

Why should we build such a resource and how can we hope to construct it? Starting with the *why*, I believe that a UD constructicon could help us improve cross-linguistic annotation consistency by providing a complementary view of the UD guidelines, which is holistic and onomasiological. It is holistic because it starts from complete constructions rather than particular syntactic relations, and it is onomasiological because it goes from function

to (cross-linguistically identifiable) form. This would in particular benefit the annotation of new languages, where guidelines could be developed systematically by first identifying what strategies are used for different constructions. It would also provide better support for construction-based annotation on top of UD, as proposed in Weissweiler et al. (2024). Last but not least, it would help us find out to what extent UD can represent constructions and strategies systematically and transparently across languages and thereby identify shortcomings in the current guidelines.

Returning to the question of *how* to build the constructicon, we can fortunately bootstrap the process by taking the first two components – the inventories of constructions and strategies – directly from Croft (2022), or rather from MoCCA, the database of comparative concepts that is being developed from the glossary of the book (Lorenzi et al.).[9] We can then concentrate on constructing valid UD analyses for all construction-strategy pairs, starting with the most prototypical construction types – reference, modification and predication – and proceeding to non-prototypical cases with more complex variation patterns. Examples for all constructions can be found in Croft (2022), which contains at least one concrete example for every construction-strategy pair discussed in the book. This should be supplemented with examples from existing UD treebanks, which will allow us to assess the cross-linguistic annotation consistency for different constructions and strategies.

6 Conclusion

In this paper, I have reopened the question of whether UD is an adequate annotation framework from the point of view of linguistic typology, previously raised by Croft et al. (2017). I have argued that one way of answering this question is to study more systematically how constructions and strategies, in the sense of Croft (2022), can be represented in UD, and I have proposed that this can be done by building a constructicon for UD.

Acknowledgments

Thanks to Bill Croft for valuable comments on a draft of this paper and to members of the UniDive COST Action (CA21167) for useful discussions. Swedish Research Council grant no. 2022-02909.

[9] https://comparative-concepts.github.io/cc-database/

References

William Croft. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.

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.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86:663–687.

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig, and Ely A. Matos. MoCCA: A model of comparative concepts for aligning constructicons. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.

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

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.

Leon Stassen. 1997. *Intransitive Predication*. Oxford University Press.

Leon Stassen. 2009. *Predicative Possession*. Oxford University Press.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932.