Expansion of UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

Leonie Weissweiler
LMU Munich & MCML, Germany
weissweiler@cis.lmu.de

Abstract content

1. Project Phase

The project is based on a paper with the same name, which I will present at LREC-COLING this year. The initial project covered 10 languages and 5 constructions. While it is the long-term goal to make this a community project in the same way that UD is, and ideally even to move it to its own layer, a specific workflow for how to expand the project both in terms of languages and in terms of constructions has not been established. This is what I want to focus on in the training school.

2. Resource Description

The notion of a *construction* is an important concept in grammar as it allows for an analysis of patterns of form and function within languages as well as systematic comparisons across languages. Consider the WH-interrogatives in English and Coptic. While English uses a combination of WH-words and word order to encode such questions, Coptic typically leaves WH-words in situ, meaning they occur in the same position as non-interrogative pronouns:

(1) e- i- na- je **-pai/-ou** na- f [cop]
FOC- I- FUT- say **-it/-what** to- him
'I shall say **it** to him.' /
'**What** shall I say to him?' (ϵ-ι-ⲛⲁ-ϫϵ-ⲟⲩ ⲛⲁ-ϥ)

The notion of a WH-interrogative construction is a shared level of abstraction that underlies the differences between the languages: both languages have conventionalized morphosyntactic means to convey that a piece of information is being sought.

Meaning-bearing grammatical constructions such as interrogatives, conditionals, and resultatives are an object of study within and across languages, and many of these have been the focus of semantic/pragmatic annotation schemes, usually involving manual annotation. The goal of UCxn is to annotate them on a large scale across many languages in UD treebanks as automatically and accurately as possible. In the initial paper, we demonstrated how UD treebanks can be enriched with a layer identifying these larger constructions in a typologically informed way so as to enable crosslinguistic comparisons and typological studies. The paper has presented a case study of five construction families and ten languages to illustrate the challenges and opportunities of this approach.

This goal is challenging because holistic constructions are often not reflected in syntactic labels used in treebanks, which aim to break sentences down into minimal grammatical parts. The UD framework, for example, annotates the individual components of a construction but not the larger whole: there is no 'interrogative clause' label in UD. There are other challenges as well. For example, there are many non-canonical and elliptical ways of asking questions in English (e.g., *Can you tell us where?*) and some questions look identical to exclamations, e.g., *What stunning views*. Thus, defining constructions (or families of related constructions) in crosslinguistically comparable ways, determining what is within scope for annotation in a particular language, and reckoning with ambiguity are all significant challenges.

Despite these challenges, we see constructional annotation as a *worthy* mission for the multilingual computational linguistics community, because the empirical work will deepen understanding of constructional phenomena across languages and provide data for further typological studies. It is also a *viable* way forward, because the work will draw on the rich ecosystem of UD treebanks and tools in order to add and refine constructional descriptions over time. In addition to offering fuller grammatical descriptions of the treebanked sentences, construction annotations may be used to improve the intra- and interlingual consistency of UD guidelines and data. To compare across languages, it is necessary to identify patterns larger than a single word or grammatical relation, and to do so in a way that is sensitive to different *morphosyntactic strategies* exhibited by different languages (Croft, 2016, 2022). **UCxn** is grounded in ideas from Construction Grammar and linguistic typology. The original project annotated treebanks in 10 languages for selected constructions by constructing graph pattern queries and matching them against UD trees. Technical specification, queries and annotated corpora available at github.com/LeonieWeissweiler/UCxn.

Identifying Constructions Constructions are defined crosslinguistically in terms of their *function*, but UD annotates morphosyntactic *form*. Our hypothesis is that, in many cases, we can search for the morphosyntactic *strategies* associated with a construction using UD morphosyntactic annotations and extract tokens of the construction from a treebank with reasonable accuracy.

We test this hypothesis using Grew (Guillaume, 2021), which allows us to specify search queries with constraints on sentences and their UD annotations. For each construction, a language may have multiple Grew patterns corresponding to multiple morphosyntactic strategies. Grew can be combined with Arborator-grew (Guibon et al., 2020) to annotate the trees that it finds.

Annotation Atop UD The new annotation layer, UCxn, represents construction instances in UD treebanks. UCxn information is incorporated directly into CoNLL-U files, which support arbitrary key-value annotations via the MISC field (10th column). UCxn has introduced the key `Cxn`, located on the syntactic head token of the construction from the UD tree perspective, i.e., the highest-ranking node involved in the construction according to the UD tree, or the earliest such node in case of ties. Construction names are given possibly hierarchical names if subtypes are identifiable, such as `Interrogative-Polar-Direct` below, to reflect queries at different levels of granularity.

1	You	you	PRON	...	_
2	have	have	VERB	...	Cxn=Interrogative-Polar-Direct
3	a	a	DET	...	_
4	pencil	pencil	NOUN	...	_
5	?	?	PUNCT	...	_

3. Benefits of Training School

This project would benefit from the training school because I would be able to learn more about how to create annotation guidelines for UD. I did not know much about UD and annotation in general when the UCxn project started, and I think the writing of the guidelines would greatly benefit from several days of intensive work in which I could consult experts.

4. Open Questions

The UCxn paper has presented a case study of annotating constructions in UD treebanks. We developed automatic annotation queries for ten languages and five construction families, and developed UCxn as a framework for representing them in UD treebanks. Overall, we find that annotating constructions is feasible with a mix of automatic and manual efforts, and that with typologically-based construction definitions, the annotations support crosslinguistic quantitative studies. The next step is to scale up the UCxn approach to more languages and constructions, possibly with the aid of construction parsers (and/or UD parsers to produce larger-scale silver treebanks for investigating rare constructions). Two key questions that need to be answered are 1) how can we expand to more constructions and 2) how can we expand to more languages? I will now describe some potential approaches to each.

More Constructions The initial project covered five constructions, of which four were widely annotated across the ten languages. This represented a pilot study to test the feasibility of creating an entire layer of constructions, and to specify the way in which they would be annotated. As the long-term goal is to annotate many more constructions, and we are hopeful that the community would be interested in this, it will be useful to create a roadmap of constructions to be added. With what we have learned about the feasibility of each construction, and what is needed to annotate them, I aim to create a sorted list of constructions that can hopefully be tackled next. This would make it easier for community members to focus their efforts without first having to choose a viable construction. I would work out a list and conduct feasibility tests on the first items with the help of the community at the training school. I could then take this list back for review from the UCxn author group.

More Languages The languages initially included in UCxn were a sample that was trying to be as diverse as possible, while being bound to the limitation that collaborators with detailed knowledge of each language were needed. This resulted in ten languages for the initial study, which was a good start but neither the breadth or the diversity that we would ideally desire. Adding more languages will require recruiting experts in those languages, but that means that we need to prepare guidelines for those experts so that they can add a new language with as little effort as possible, while maintaining high standards of consistency across languages. To this end, I would like to use the summer school to enhance the current annotation guidelines into a specific guide for adding new languages, and also try to recruit annotators for new languages.

5. Keywords

grammatical constructions, treebanks, Universal Dependencies, typology, corpus annotation

Bibliographical References

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

William Croft. 2022. *Morphosyntax: Constructions of the world's languages*. Cambridge University Press.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.