Invited Talk: William Croft, University of New Mexico

Using Typology to Develop Guidelines for Universal Dependencies

1. Linguistic Typology and Universal Dependencies

Language structures are incredibly diverse. Although typologists have discovered many language universals, a common saying in the field is that the only exceptionless language universal is that all language universals have exceptions. There are two major reasons for this diversity. First, language is a general-purpose communication system, and every subtly different thing we want to communicate has to be put into a (relatively) small number of words and constructions. Speakers of different languages do this in many different ways. Second, language change is gradual: constructions change their morphosyntactic properties one at a time, which increases structural diversity and blurs lines between construction types.

This is what typological theory would tell us. But for practical purposes, we have to carve up this continuum of language phenomena, and at any rate, the continuum is lumpy: the space of possible structures is dense in some regions and sparse in others. Hence there are better and worse ways to carve up the continuum.

Universal Dependencies represents one practical task that requires making such choices. UD aims to develop a syntactic annotation scheme used across languages that, if applied consistently, allows for comparison across languages, including languages not yet possessing UD resources (Nivre, 2015; Nivre et al., 2016).

Another practical task that requires making such choices is teaching a typologically-informed syntax course to undergraduates as their first syntax class. In both UD and teaching syntax, the aim is to develop a small set of annotations that can be applied more or less uniformly across languages, to capture similarities as well as reveal differences. This is how I became involved in UD. My focus has been on the syntactic dependency annotation of UD. There are different and more difficult issues in the POS tagging and morphological feature tagging of the UD enterprise, which I will not go into here.

2. Two basic principles for typological annotation of dependencies

Several basic principles guided my effort, and the two most important principles are described here; for more details, see Croft et al. (2017). The first is based on a distinction between *constructions* and *strategies* in crosslinguistic comparison. Constructions describe the class of grammatical structures in any language that is used to express a particular function. For example, *Ivan is the best dancer* is an instance of the predicate nominal construction, that is, the construction whose function is to predicate an object category of a referent.

Strategies are particular morphosyntactic structures, defined in a cross-linguistically valid fashion, that are used to express a function. For example, English uses an inflecting copula strategy for predicate nominals, that is, a word form distinct from the object word that inflects for at least some of the grammatical categories that ordinary predicates do. Other languages also use the inflecting copula strategy; but still other languages use an uninflected copula, or no copula at all, or inflect the object word. These are all different strategies.

The principle for designing a universal set of dependencies is that the structure of constructions should form the backbone of the dependency structure; strategies are secondary, although they have to be annotated when they are expressed by independent words, such as the English copula. UD's content-word-to-content-word principle basically conforms to this principle.

The second important principle is based on the hypothesis that constructions always involve the information packaging of the semantic content of the sentence, that is, the function of constructions has to be defined in terms of both semantic content and information packaging. For example the predicate nominal construction involves packaging an object concept as a predication.

The principle that emerges from this hypothesis is that universal dependencies are, to a great extent, describing information-packaging relations, not semantic relations. That is, information packaging functions are much more isomorphic to syntactic structures than semantic classes or semantic relations. Information packaging functions are less variable across languages than semantics, especially lexical semantics. UD minimizes reliance on semantics in defining UD dependencies and in applying them to specific languages, so UD basically conforms to this principle as well.

3. UD dependencies: inventory and guidelines

The principles described in the preceding section, and other principles described in Croft et al. (2017), led me to a set of universal dependencies that is quite close but not identical to the set of universal dependencies in UD (version 2). These differences are relatively minor, although I will discuss one of them in this talk. The much bigger issue is the development of guidelines for consistent annotation of the many different constructions and the many different strategies that languages use, both for languages for which there exist UD resources and for new languages which may be added.

What is the best way to do this? Constructions, as defined in the preceding section, are not enough: they are defined by function, whereas we need to carve breaks in the range of strategies used to express function. The basic idea is to find typological universals constraining the distribution of strategies over constructions in such a way that the universals reveal the "cleanest" breaks and the best strategies to use as uniform guidelines across languages.

This will be a "good news, bad news" story. The "good news" is that some current practice, based mainly on Western grammatical tradition and the Western European languages that make up most of the UD treebanks, are justified in a broader typological perspective, and allow for uniform guidelines. The "bad news" is that some current practices, and some distinctions among UD dependencies, are not very well justified typologically. In some of these cases, the dependencies I use in teaching syntax differ from the current version of UD.

I believe that for the most part, the good news exceeds the bad news. The most important conclusion is that detailed guidelines are necessary, and ideally should be typologically justified. An overview of the typological variation and typological universals constraining that variation—and justifying distinctions we need to make—will appear in my forthcoming textbook for the advanced syntax class I teach (Croft, In preparation).

4. Some examples of how typology can be used to develop guidelines for UD

UD distinguishes between core grammatical roles (*subj, dobj, iobj*) from oblique roles (*obl*). In practice, however, this is difficult. We cannot rely on semantic roles (patient, instrument, etc.) because voice, argument structure alternations and applicative constructions change the syntactic roles of participants. Hence we must look elsewhere.

There are three strategies used for encoding core and oblique arguments: case marking (adpositions and affixes), indexation (agreement) and word order. The categories of case markers vary a lot, and there are mismatches across strategies. How safe is it to rely on these strategies for annotating core vs. oblique?

Fortunately, there are two universals that support the identification of core vsòblique arguments:

- *If case marking is zero, then the argument is overwhelmingly likely to be core.*
- *If the predicate indexes the argument, then the argument is overwhelmingly likely to be core.*

There are exceptions, but the point is that they are rare. So we can assume that if the argument phrase has zero-coded case marking and/or is indexed on the verb, it is core, without having to rely on semantic roles. The universals are one-way conditionals: some core arguments have overt case marking, and others are not indexed on the predicate. But it is usually clear which case-marked arguments are core.

An example which represents not so good news is when there are mismatches in strategies for arguments. Two common examples are so-called "dative subjects", common in South Asian languages, and "patient subjects" (passives). There is a diachronic typological universal governing the acquistion of subjecthood (Cole et al., 1980; Croft, 2001):

- *Nonsubject arguments may become subjectlike, first by word order, then indexation, then case marking.*

Unfortunately, this universal implies a gradient of strategies from nonsubject to subject, and does not offer guidelines as to when to decide when an argument is a subject, or still is not a subject. However, it is unlikely that the constructions with mismatches are common. In the case of mismatches, I would suggest that if an argument uses any morphological strategy associated with subject status—that is, case marking or indexation—then it should be annotated as subject. The universal indicates that such mismatches will have subject-like indexation but nonsubject case marking.

Other cases of a gradient of strategies are found in several common paths of grammaticalization (Heine and Kuteva, 2002; Lehmann, 2002). These cases also involve a reversal of headedness in UD, which is problematic in a dependency grammar (heads are in boldface):

- **Verb** + Complement → Auxiliary + **Verb**
- **Relational Noun** + Noun → Adposition + **Noun**
- **Verb** + Noun → Adposition + **Noun**
- **Quantity** + Noun → Quantifier + **Noun**

As with the acquisition of subjecthood, it is likely that the intermediate cases are crosslinguistically not that common. I would suggest, as with subject annotation, that once a construction acquires the first typical strategy for the more grammaticalized construction, it should be annotated like the more grammaticalized construction.

Some semantic roles, such as recipient, are sometimes core and sometimes oblique across languages; and they are sometimes both in the same language, in which case they are described as object-oblique alternations: *I showed the policeman my driver's license/I showed my driver's license to the policeman*. In typology, these are called different strategies for encoding the recipient, specifically alignment strategies (Haspelmath, 2011).

If they are simply different strategies, then perhaps they should be annotated the same way in a universal scheme like UD. But in fact a construction should be defined by both semantic content and information packaging (Croft, 2016, In preparation). Encoding a participant role as object or oblique arguably does differ in information packaging. In most languages, only one option exists, object or oblique. But the crosslinguistic variation is due to competing motivations: for example, a recipient is a less central event participant, yet it is almost always human and hence of greater salience. So I conclude that one should follow the language's structure in annotating a semantic role as object or oblique.

In the equivalent German sentence, the recipient role is in the Dative case, whlie the theme role is in the Accusative case: *Ich zeigte dem Polizisten* [Dative] *meinen Führerschein* [accusative]. Many Germanists analyze the Dative as an object, despite the oblique-like case marking. This is justified by the fact that the dative noun phrase occurs without a preposition. Yet there is a language universal that suggests this is the right choice, albeit for a different reason (Siewierska, 1998; Levin, 2008):

- *Constructions with a dative coding of the recipient distinct from the allative or locative coding are crosslinguistically in complementary distribution to constructions with the same coding of recipient and theme.*

Hence, even a language-specific annotation choice may be typologically justified, though in this case the rule should be whether the dative is distinct from allative or locative, not whether the dative noun phrase is accompanied by a preposition.

Modifiers are a more complex case. Modifiers come in many different semantic types: definiteness (articles), deixis (demonstrative), cardinality (cardinal numerals), quantification (quantifiers), properties (adjective), actions (relative clauses, participles) and possession (genitive) and other noun-noun relations. UD distinguishes a subset of those semantic types: *det, nummod, amod, nmod* and *acl*. Modifiers also use a wide range of strategies: gender/number agreement, case marking, classifiers, and linkers (more grammaticalized, invariant markers of a relation). However, all the different strategies are found across almost all modifier types, although there is typological evidence that noun modifiers and relative clauses tend to stand apart. In this case, I have lumped together all modifiers into a single `mod` dependency, except for *nmod* and *acl*.

Finally, one of the more challenging problems is distinguishing subordinate clauses from nominalizations (or in the case of participles, adjectivalizations). Constructions using all the structure of main clauses—tense-aspect-modality (TAM) inflections, indexation of core arguments, main clause-like case marking of core arguments—such as *I am surprised that he fired Flynn* are clearly subordinate clauses. But there is a wide range of constructions lacking some or all of the typical structure of main clauses, as in *His firing Flynn surprised me* or *Him firing Flynn was surprising*. Also, the terminology in grammatical description here is very confusing: there are special terms such as infinitives, gerunds, masdars, and converbs; but many descriptions use the term "nominalization" for all sorts of non-clause-like constructions.

Fortunately, there is a reliable grammatical criterion that has two significant typological universals associated with it, which allows us to consistently distinguish subordinate clauses from nominalizations. The grammatical criterion is that an event nominalization allows for "reasonably productive" case marking (Comrie, 1976). The two universals are (Cristofaro, 2003):

- *If a verb form can take case affixes or adpositions, then with overwhelming frequency it does not inflect for TAM like a main clause verb (it either has no TAM inflections, or uses special TAM forms).*
- *If a verb form can take case affixes or adpositions, then with overwhleming frequency it does not express person indexation/agreement like a main clause verb (it either has no person indexation, or special person indexation forms different from those in main clauses).*

In other words, external case marking of verb forms coincides with non-clauselike TAM inflection and person indexation. Again,this is a one-way conditional: subordinate clauses may lack the TAM or indexation of main clauses. Case marking of dependent arguments of the verb, however, does not conform to these universals and so cannot be used reliably to distinguish subordinate clauses from nominalizations. But case marking of the verb form can be used reliably and consistently as a guideline to distinguish subordinate clauses from nominalizations (or adjectivalizations, for participial modifiers).

Deciding whether a verb form allows "reasonably productive" case marking is not always easy, since dependent constructions denoting actions do not take the full range of case forms, and infinitives are often historically derived from allative case marking, such as English *I began **to** work*. But case marking of the verb form is a consistent and typologically justified criterion.

Finally, there is an asymmetry in strategies between complement clauses and adverbial subordinate clauses that can be used for guidelines to distinguish complement relations (UD *scomp, ccomp, xcomp*) from adverbial ones (UD *advcl*):

- *If the subordinating conjunction is relational, that is, expresses contrastively a semantic relation between the matrix clause and the subordinate clause, then the subordinate clause is overwhelmingly likely to be an adverbial clause.*
- *If the subordinating conjunction is a linker, so does not express a specific semantic relation, then the subordinate clause is overwhelmingly likely to be a complement (or relative clause).*

If a verb form that semantically looks like a complement appears to take case marking, then it is likely that either it is part of a paradigm of case-marked verb forms and hence is an event nominal, or the putative case marking no longer contrasts meaningfully with another form, as in English infinitival *to*, and so should be analyzed as a linker governing a complement clause.

These examples indicate how typological universals about the relationship between functions of constructions—semantic content and information packaging—and the grammatical strategies used to express those functions can help in constructing guidelines for applying Universal Dependencies across languages in a consistent fashion.

References

Peter Cole, Wayne Harbert, Gabriella Hermon, and S. N. Sridhar. The acquisition of subjecthood. *Language*, 56:719–743, 1980.

Bernard Comrie. The syntax of action nominals: a cross-language study. *Lingua*, 40:177–201, 1976.

Sonia Cristofaro. *Subordination*. Oxford: Oxford University Press, 2003.

William Croft. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, 2001.

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

William Croft. *Morphosyntax: constructions of the world's languages*. Cambridge: Cambridge University Press, In preparation.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. Linguistic typology meets universal dependencies. In Markus Dickinson, Jan Hajič, Sandra Kübler, and Adam Przepiórkowski, editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75. CEUR Workshop Proceedings, 2017.

Martin Haspelmath. On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology*, 15:535–67, 2011.

Bernd Heine and Tania Kuteva. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press, 2002.

Christian Lehmann. *Thoughts on grammaticalization: a programmatic sketch, Vol. I*, volume 9, Arbeitspapiere des Seminars für Sprachwissenschaft der Universität. Erfurt: Seminar für Sprachwissenschaft der Universität, 2002.

Beth Levin. Dative verbs: a crosslinguistic perspective. *Linguisticæ Investigationes*, 31:285–312, 2008.

Joakim Nivre. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational linguistics and intelligent text processing*, pages 3–16. New York: Springer, 2015.

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 Daniel Zeman. Universal Dependencies v1: a multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1659–1666. European Language Resources Association, 2016.

Anna Siewierska. Languages with and without Objects: the Functional Grammar perspective. *Languages in Contrast*, 1:173–90, 1998.