UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

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, Nathan Schneider

Constructions

- Combinations of **form** and **meaning** with varying schematicity, idiomaticity, and complexity
- Constructions can be patterns with slots with morphological, syntactic, or semantic constraints
  - The X-er, the Y-er
  - Jog X's memory

Most constructions are not represented in UD!

Goals

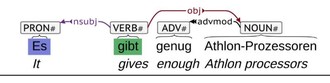
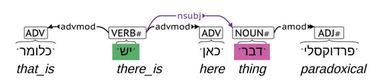
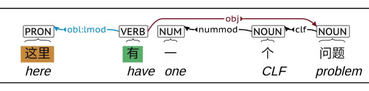
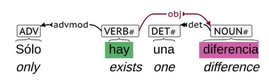
- New construction layer in UD
- Annotate more constructions/construction families
- Add more languages
- Use the layer for typological studies

Annotation Approach

Identifying the construction and adding

- Name
- Constructional anchor
- Elements

```
Who let the dogs out ?
1  Who  …  2  nsubj  …  CxnElt=2:Interrogative.WHWord
2  let  …  0  root   …
Cxn=Interrogative,Resultative|CxnElt=2:Interrogative.Clause,2:Resultative.Event
3  the  …  4  det    …  _
4  dogs …  2  obj    …  _
5  out  …  2  xcomp  …  CxnElt=2:Resultative.ResultState
6  ?    …  2  punct  …  _
```

Language	Instance	Query
German		pattern EXP[lemma="es"]; PRED[lemma="geben"]; PRED-[nsubj]->EXP;
Hebrew		pattern PRED[lemma="יש"]; PRED-[nsubj]->PIV; without LE[lemma="ל"]; PRED-[obl]->N; N-[case]->LE;
Mandarin		pattern PRED[form="有"]; PRED-[obl:lmod]->COD;
Spanish		pattern PRED[lemma="haber"]; PRED-[obj]->PIV; DET[upos=DET, Definite=Ind]; PIV-[det]->DET;

Existentials

Il y a une salle à l'étage (French)
It there has a room upstairs
There is a room upstairs.

Assert the existence (or non-existence) of an **entity (pivot)**, almost always indefinite, may specify a **location (coda)**

Identified using:

- Specific lexical items (e.g. Swedish *finnas*)
- Specific annotations (e.g. HebExistential=Yes)
- Only dependencies (→ false positives)
- Word order

Conditionals

Kommst du, gehe ich. (German)
Come.2SG you go.1SG I
If you come, I will go.

Complex sentence construction describing a broadly causal link between two states of affairs: the **protasis (condition)** and the **apodosis (consequence)**

Identified using:

- Conjunctions
- Word order
- Conditional circumfixes (e.g. Coptic)

Interrogatives

(ϵ-ı-ⲛⲁ-ϫϵ-ⲟⲩ ⲛⲁ-⸗) (Coptic)
E- i- na- je -ou na- f
FOC I- FUT say -what to- him
What shall I say to him?

Speech act construction expressing a request for information from the addressee

Identified using:

- Presence of WH-items
- Word order
- Question marks
- Existing sentence type annotations
- PronType = Int

Resultatives

我 敲 平 了 钉子 (Chinese)
wǒ qiāo píng le dīngzi
1SG hit flat PERF nail
I hammered the nail flat.

Expresses an event with two subevents: a **dynamic subevent** and a **resulting state subevent**

- Doesn't exist at all in some languages
- Doesn't exist as complex predicate in some languages (e.g. Hebrew)
- Indistinguishable from causatives and depictives without fixed word lists

N-Prep-N

מילה במילה (Hebrew)
mila be-mila
word in-word
word for word

Strategy, not construction
→ one form, multiple possible meanings

- Succession
- Comparison
- Opposition
- Proximity
- Quantification

Results

Lang.	Interrogative (§4)	Existential (§5)	Conditional (§6)	Resultative (§7)	NPN (§8)	total sent.	total tokens
English	1117; 769	472; 319 (F)	762; 375 (D)	H, D	21; 12	17k; 11k	254k; 187k
German	5483 (H)	3392 (H)	3291 (A,H)	D	40	190k	3.5m
Swedish	276	235	310 (H)	D	7	6k	96k
French	368	114 (F)	213 (F)	D	12	16k	400k
Spanish	580	160 (F)	502 (F)	D	37	18k	567k
Portuguese	337 (A)	340 (F)	106	D	7	9k	227k
Hindi	285	2058 (F)	350 (A)	D	?	16k	351k
Chinese	146	58 (F)	31	78 (D)	?	1k	9k
Hebrew	236; 22	113; 60	192; 56	D	9; 11	6k; 5k	160k; 140k
Coptic	150	80	185	D	2	2k	55k

Counts of identified construction instances by treebank, along with qualifications: definitional issues (D), UD annotation errors (A), occasional false positives (F), frequent false positives (F), unattested strategies (H). ? means that the existence of the productive construction is doubtful. The two numbers for EN and HE represent the two treebanks for each.

Resource

Full data and annotation guidelines are available on GitHub!