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

Background: Universal Dependencies

- **Annotation framework for annotating dependencies and parts of speech consistently across languages**
- **Over 200 treebanks in over 100 languages → gigantic resource**

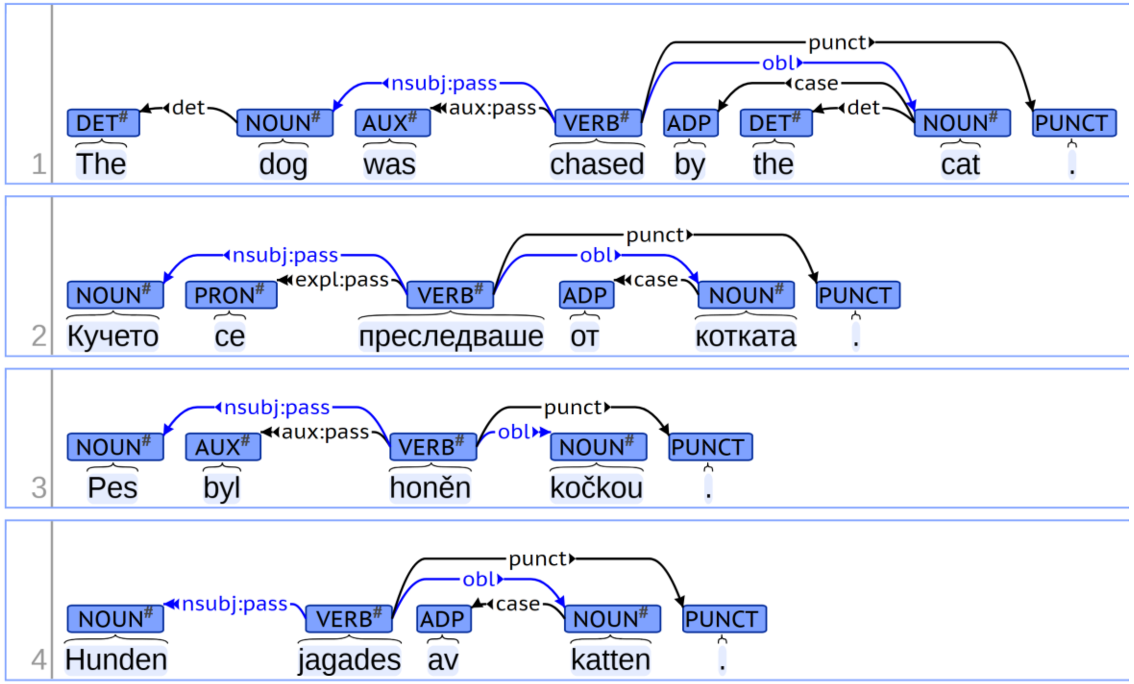
Enables you to find e.g. all noun phrases

However:

Some language structures are not in UD!

What if you want to look at all the questions?

→Solution: Construction Annotation in UD!



Constructions

- **Combinations of form and meaning**
- **Patterns with slots which can have morphological, syntactic, and semantic constraints**
- **Some favourite examples:**

`The X-er, the Y-er` ***The clearer the talk, the more you will understand***

`Jog someone‘s memory` ***I had to jog her memory on this***

`Article Adjective Numeral Noun` ***A beautiful five days in Austin***

`Let-alone` ***I can‘t do this sober, let alone drunk***

Constructions Typologically

According to Croft (2022): defining constructions by function, not form

→ One construction across languages (e.g. Interrogative), many different strategies (some shared across languages), e.g. wh-question

Challenge

UD mainly annotates morphosyntactic form → annotation is nontrivial

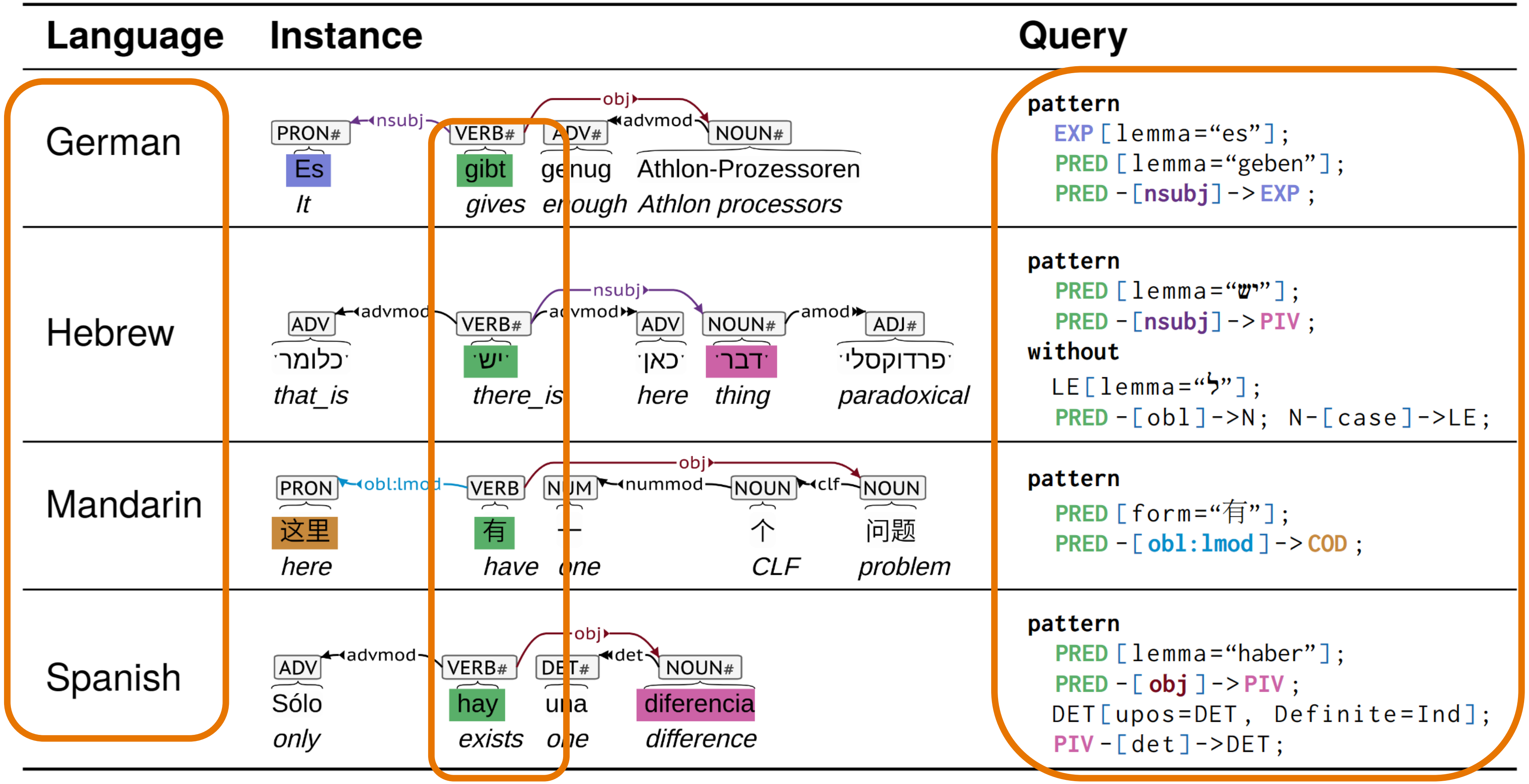
Hypothesis

many constructions can be annotated with reasonable precision by writing one or more rules operating on top of existing UD annotation

A Construction Annotation Layer in UD

```
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  …    _
```

Semi-Automatic Annotation



Language	Instance	Query

German	nsubj, obj, advmod
	PRON# VERB# ADV# NOUN#
	Es gibt genug Athlon-Prozessoren
	It gives enough Athlon processors
		pattern
		EXP [lemma=“es”];
		PRED [lemma=“geben”];
		PRED -[nsubj]-> EXP;

Hebrew	advmod, nsubj, advmod, amod
	ADV VERB# ADV NOUN# ADJ#
	כלומר יש כאן דבר פרדוקסלי
	that_is there_is here thing paradoxical
		pattern
		PRED [lemma=“יש”];
		PRED -[nsubj]-> PIV;
		without
		LE[lemma=“ל”];
		PRED -[obl]->N; N-[case]->LE;

Mandarin	obl:lmod, obj, nummod, clf
	PRON VERB NUM NOUN NOUN
	这里 有 一 个 问题
	here have one CLF problem
		pattern
		PRED [form=“有”];
		PRED -[obl:lmod]-> COD;

Spanish	advmod, obj, det
	ADV VERB# DET# NOUN#
	Sólo hay una diferencia
	only exists one difference
		pattern
		PRED [lemma=“haber”];
		PRED -[obj]-> PIV;
		DET[upos=DET, Definite=Ind];
		PIV -[det]->DET;

Table 1: Existential/presententional construction instances in selected languages and the Grew queries used to identify them. The predicate (PRED), pivot (PIV), coda (COD) and expletive subject (EXP) construction elements and the nsubj, obj and obl:lmod dependency relations are color-coded in the trees and queries.

Overview

- **5 constructions: Interrogative, Existential, Conditional, Resultative, NPN (strategy)**
- **10 languages: English, German, Swedish, French, Spanish, Brazilian Portuguese, Hindi, Chinese, Hebrew, Coptic**
- **Full rules and annotated data on Github!**
- **In the paper: typological overview, annotation efforts, and takeaways for each construction**

Interrogatives

- **Speech act construction expressing a request for information from the addressee**

  (ⲉ-ⲓ-ⲛⲁ-ϫⲉ-ⲟⲩ ⲛⲁ-ϥ)

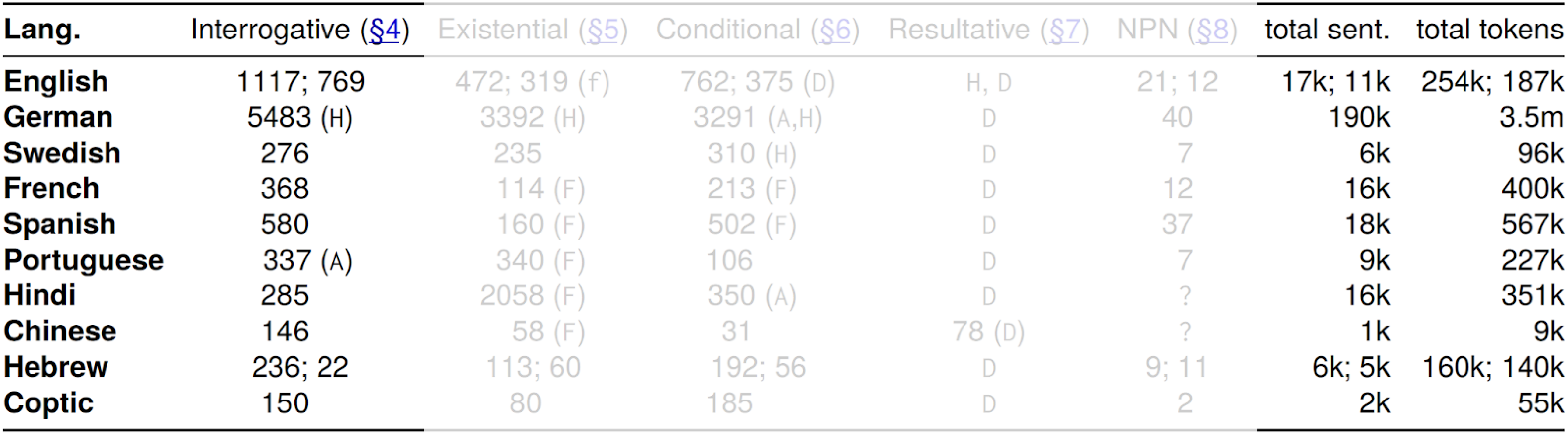
  E- i- na- je -ou na- f

  foc I- fut say -what to- him

  What shall I say to him?

- **Annotation**
  - presence of WH items (what, who, etc.)
  - word order
  - question marks
  - existing sentence type annotations
  - PronType=Int

Annotated Interrogatives



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.

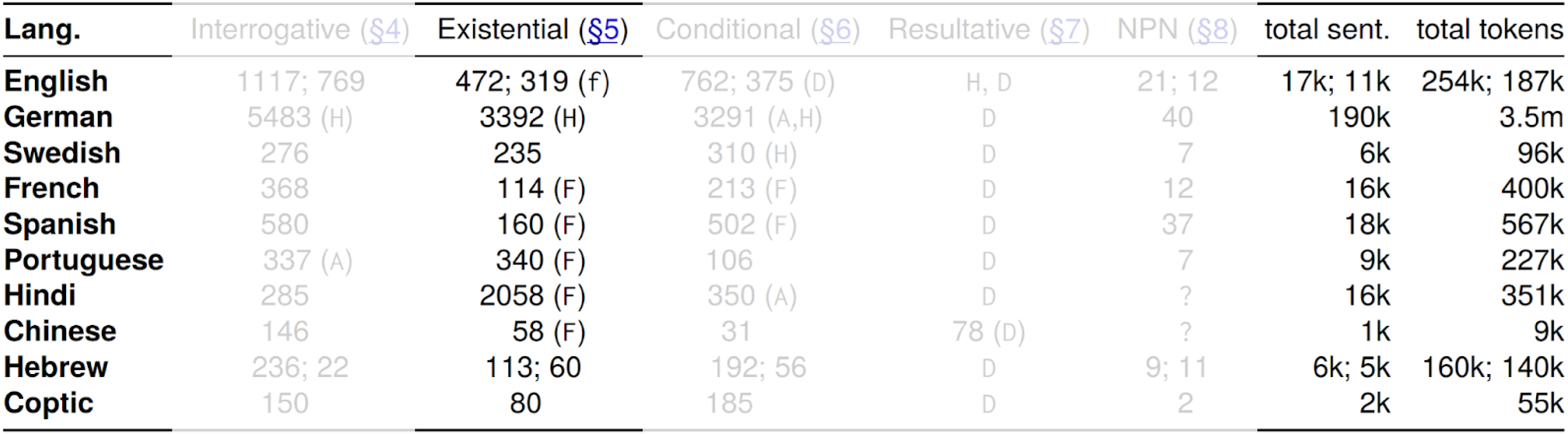
Existentials

- **Assert the existence (or non-existence) of an entity (pivot), almost always indefinitive, usually specificied in a location (coda)**
- **Il y a une salle à l'étage**

  It there has a room upstairs

  There is a room upstairs.
- **Formally indistinguishable from presentatives: *There's a yak on the road***
- **Diachronically and synchronically related to (sharing strategies with) possessives (e.g. French), predicational locative (e.g. Hebrew) and auxiliaries (e.g. Spanish)**
- **Identified with**
  - specific lexical items (e.g. Swedish)
  - specific annotations (e.g. HebExistential=Yes)
  - only dependencies (resulting in false positives)

Annotated Existentials



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.

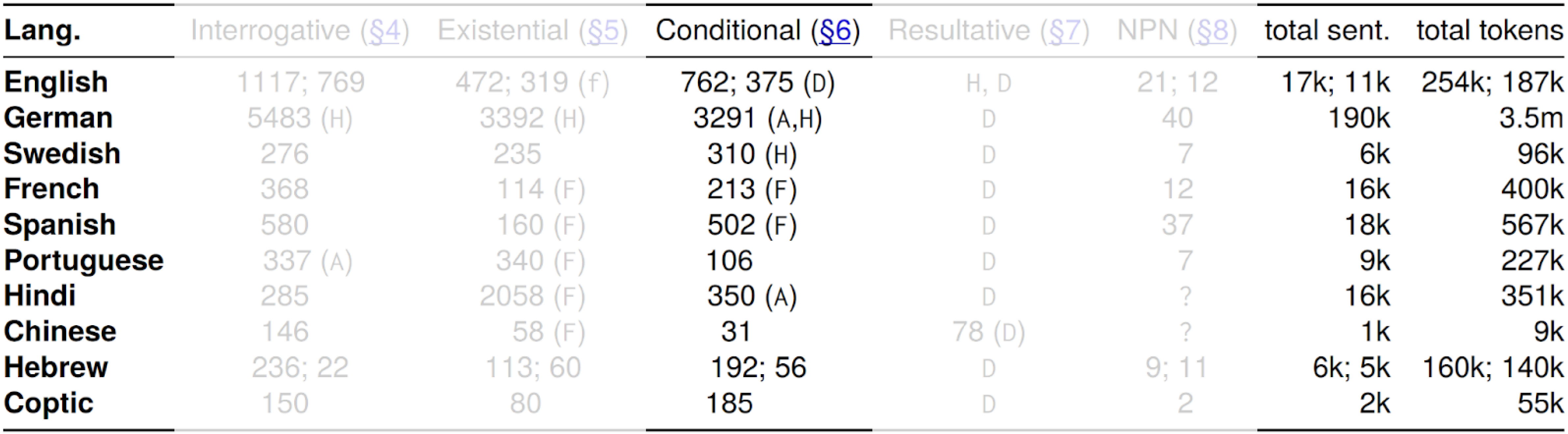
Conditionals

- **Complex sentence construction describing a broadly causal link between the two states of affairs, the protasis (condition) and the apodosis (consequence)**
- **Kommst du, gehe ich.**

  Come.2SG you go.1SG I

  If you come, I will go.
- **Identified using**
  - Conjunctions
  - Word order
  - Conditional circumfixes (e.g. Coptic)
- **Unavoidable false positives due to shared strategies with other constructions**

Annotated Conditionals



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.

Resultatives

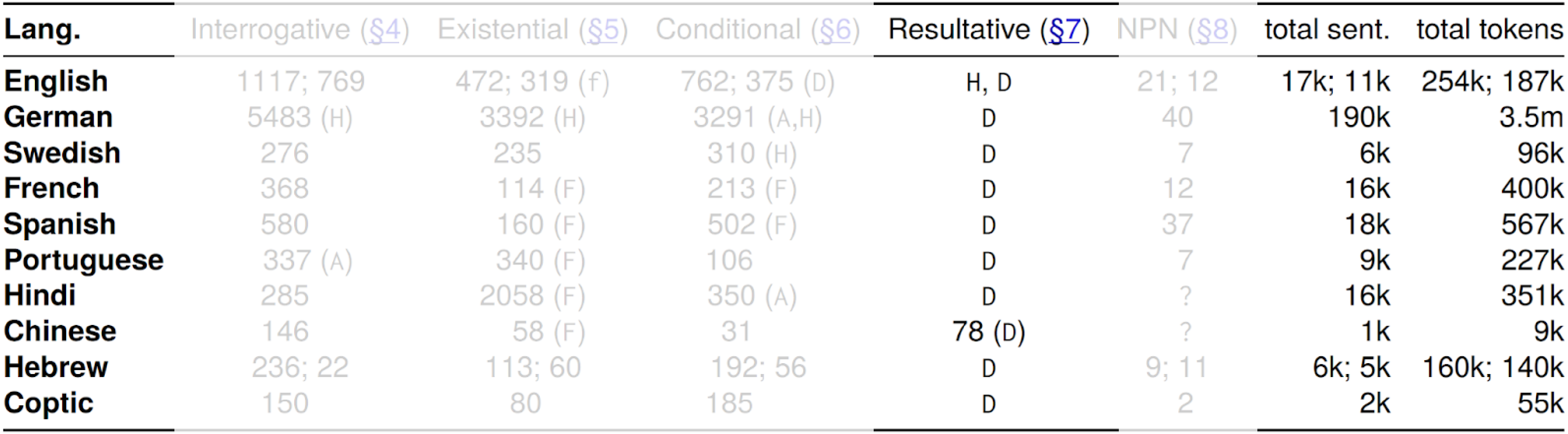
- **Expresses an event with two subevents: a dynamic subevent and a resulting state subevent**
- **我 敲 平 了 钉子**

  wǒ qiāo píng le dīngzi

  1SG hit flat PERF nail

  I hammered the nail flat.
- **Difficulty: non-conventionalised ways of expressing this, like *The door was red as a result of their painting* (not annotated here)**
- **Some languages (e.g. Hebrew) do not have a complex predicate for resultatives**
- **Results are often indistinguishable from**
  - Causatives: *This makes it possible*
  - Depictives: *I left the door open*

Annotated Resultatives

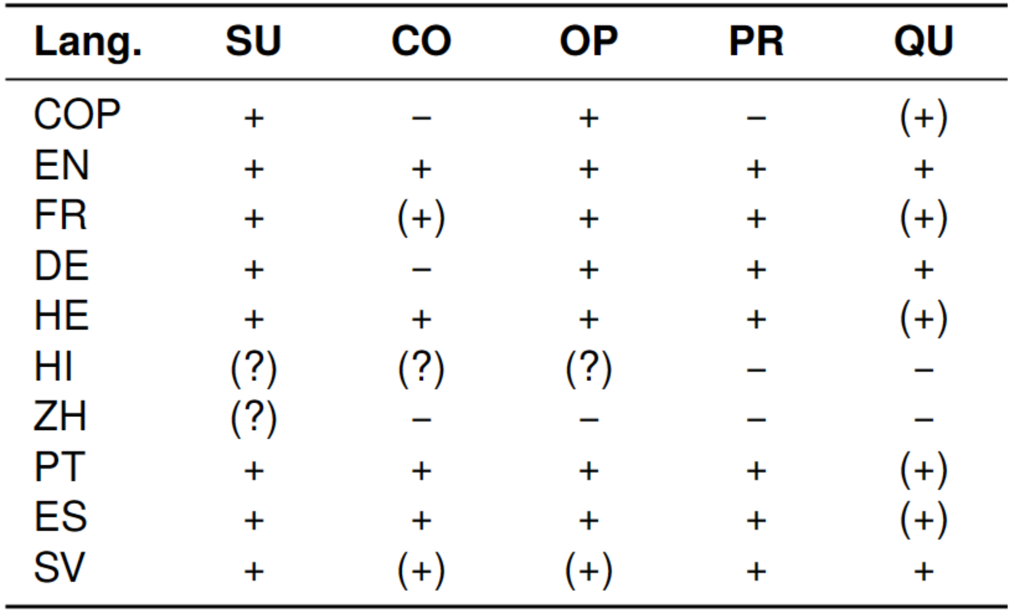


Lang.	Interrogative (§4)	Existential (§5)	Conditional (§6)	Resultative (§7)	NPN (§8)	total sent.	total tokens
English	1117; 769	472; 319 (f)	762; 375 (D)	H, D	21; 12	17k; 11k	254k; 187k
German	5483 (H)	3392 (H)	3291 (A,H)	D	40	190k	3.5m
Swedish	276	235	310 (H)	D	7	6k	96k
French	368	114 (F)	213 (F)	D	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.

NPNs

- Strategy, not a construction → one form, multiple possible meanings
- Day after day, shoulder to shoulder, box upon box
- Easy to automatically annotate



Lang.	SU	CO	OP	PR	QU
COP	+	–	+	–	(+)
EN	+	+	+	+	+
FR	+	(+)	+	+	(+)
DE	+	–	+	+	+
HE	+	+	+	+	(+)
HI	(?)	(?)	(?)	–	–
ZH	(?)	–	–	–	–
PT	+	+	+	+	(+)
ES	+	+	+	+	(+)
SV	+	(+)	(+)	+	+

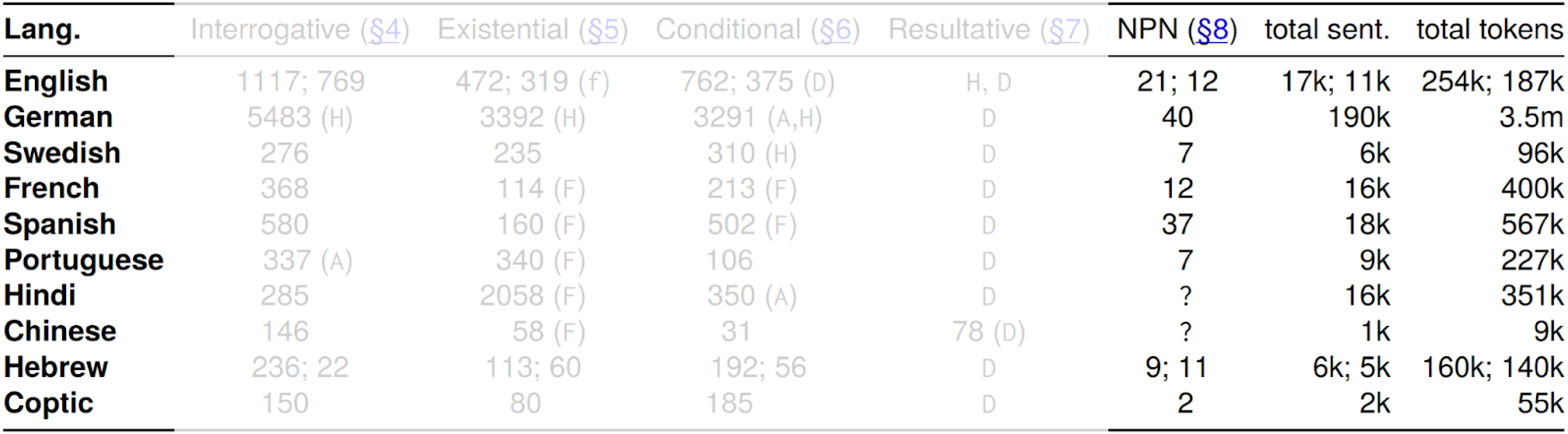
Analysis of attested meanings

- Succession: hour after hour
- Comparison: man for man
- Opposition: brother against brother
- Proximity: hand in hand
- Quantification: snacks upon snacks

(+) possible but not attested in treebanks

(?) existence unclear

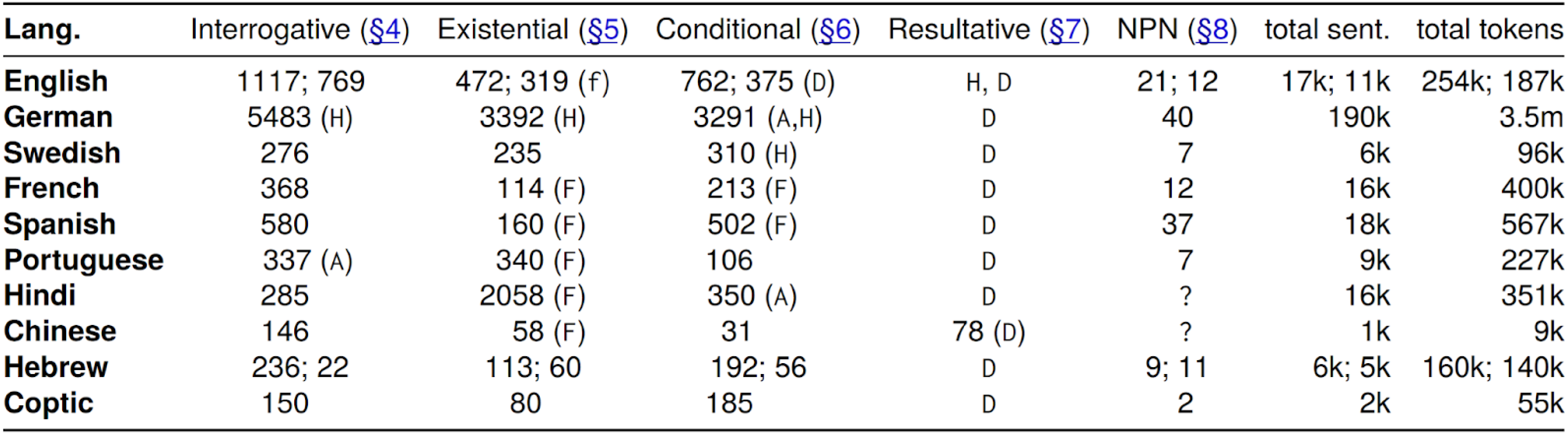
Annotated NPNs



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.

UCxn V1: A New Resource



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.

Summary

- **Pilot study of the feasibility of annotating constructions in UD, using UD**
- **Successfully annotated four out of five constructions for ten languages**
- **Developed annotation guidelines for future constructions and languages**
- **Check out the UCxn GitHub!**

We‘d love to add more languages and constructions!

Feel free to contact us

Leonie.weissweiler@gmail.com - @LAWeissweiler

We‘re here, come talk to us!