Meaning Construction at the Syntax-Lexis Nexus

Nathan Schneider
Georgetown University

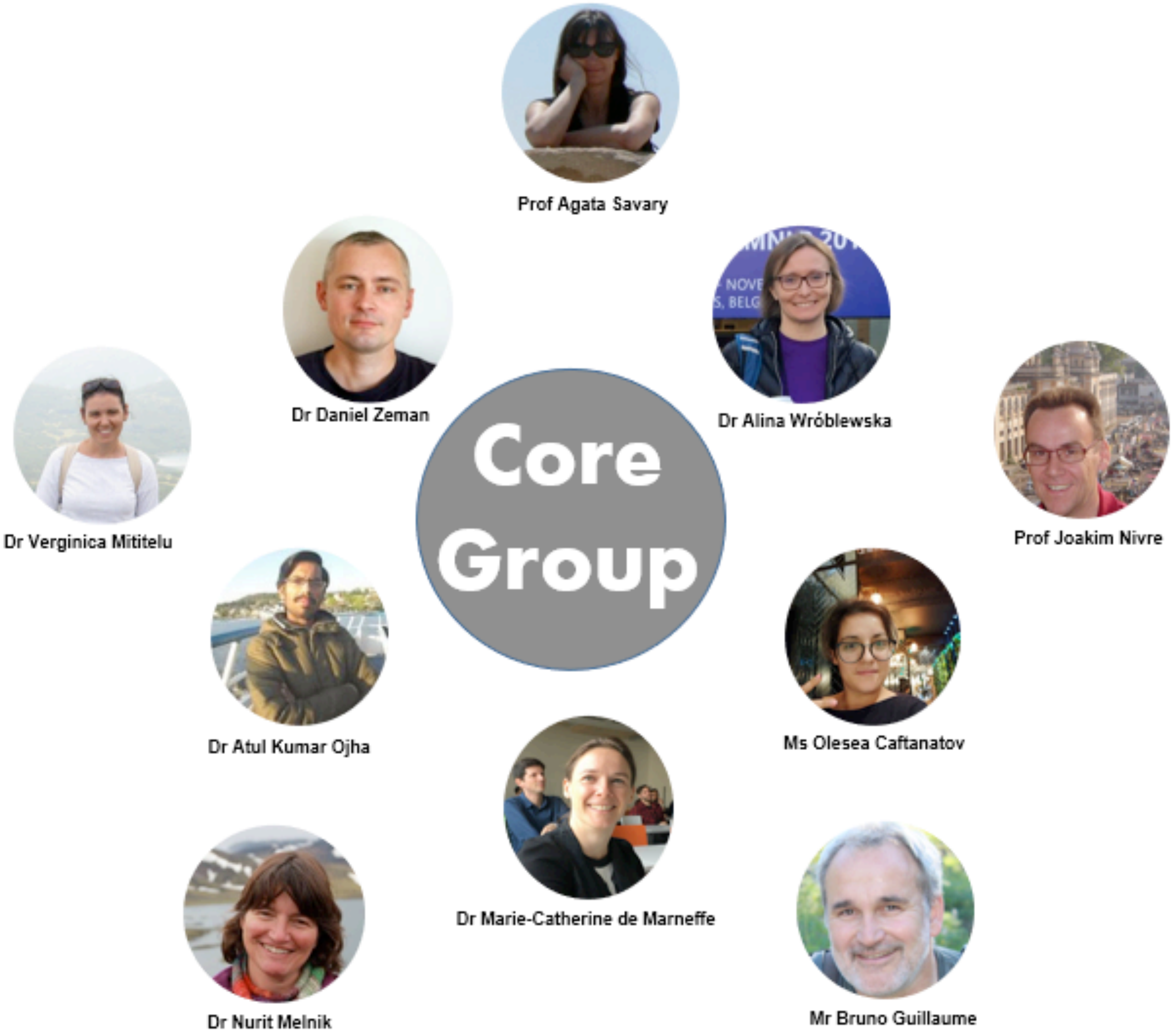
MWE 2025
May 4, Albuquerque

Image: ChatGPT

Shoutouts

UniDive

Universality, diversity and idiosyncrasy in language technology
COST Action



Core Group

Prof Agata Savary

Dr Daniel Zeman

Dr Alina Wróblewska

Dr Verginica Mititelu

Prof Joakim Nivre

Dr Atul Kumar Ojha

Ms Olesea Caftanatov

Dr Marie-Catherine de Marneffe

Dr Nurit Melnik

Mr Bruno Guillaume

Universal Dependencies

Outline

- This talk: moving beyond classical MWEs to look at situations where lexical and grammatical information interact in interesting ways
  - the syntax-lexis nexus
  - from my perspective as someone who does corpus annotation and works with language models

***no relation to LexisNexis®

- First: some background on MWEs & constructions

- Then: case studies in annotation with UD treebanks, and probing LMs

Linguistic Outlaws

- “Words and Rules” paradigm breaks down if you look closely
  - Some meanings come in packages larger than one word!

MWE Definition

Multiword expression (MWE): 2 or more orthographic words that are tightly associated

- **Strong MWEs:** idiomatic = not fully predictable in form and/or function
  - *non-* or *semi-compositional*: **ice cream**, **daddy longlegs**, **pay attention**
  - *unusual morphosyntax*: **by and large**
- **Weak MWEs:** statistically collocated or formulaic
  - p(**heavy rain**) > p(**strong rain**); **highly recommended**; **no amount of … can …**

STREUSLE corpus (Schneider et al., LREC 2014)

Noam Chomsky

daddy longlegs, hot dog

dry out the clothes

depend on, come across

pay attention (to)

no attention was paid (to)

put up with, give in (to)

under the weather

cut and dry

in spite of

pick up where they left off

easy as pie

You’re welcome.

To each his own.

The structure of this paper is as follows.

POS pattern	MWEs *contig.*	 *gappy*	most frequent types (lowercased lemmas) and their counts
N_N	331	1	**customer service: 31** oil change: 9 wait staff: 5 garage door: 4
^_^	325	1	santa fe: 4 dr. shady: 4
V_P	217	44	**work with: 27** **deal with: 16** **look for: 12** **have to: 12** ask for: 8
V_T	149	42	**pick up: 15** **check out: 10** show up: 9 end up: 6 give up: 5
V_N	31	107	take time: 7 give chance: 5 waste time: 5 have experience: 5
A_N	133	3	front desk: 6 top notch: 6 last minute: 5
V_R	103	30	**come in: 12** come out: 8 take in: 7 stop in: 6 call back: 5
D_N	83	1	**a lot: 30** **a bit: 13** a couple: 9
P_N	67	8	**on time: 10** in town: 9 in fact: 7
R_R	72	1	**at least: 10** at best: 7 as well: 6 of course: 5 at all: 5
V_D_N	46	21	**take the time: 11** do a job: 8
V~N	7	56	*do job: 9* *waste time: 4*
^_^_^	63		home delivery service: 3 lake forest tots: 3
R~V	49		***highly recommend: 43*** *well spend: 1* *pleasantly surprise: 1*
P_D_N	33	6	over the phone: 4 on the side: 3 at this point: 2 on a budget: 2
A_P	39		pleased with: 7 happy with: 6 interested in: 5
P_P	39		**out of: 10** due to: 9 because of: 7
V_O	38		**thank you: 26** get it: 2 trust me: 2
V_V	8	30	get do: 8 let know: 5 have do: 4
N~N	34	1	*channel guide: 2* *drug seeker: 2* *room key: 1* *bus route: 1*
A~N	31		*hidden gem: 3* *great job: 2* *physical address: 2* *many thanks: 2* *great guy: 1*
V_N_P	16	15	**take care of: 14** have problem with: 5
N_V	18	10	mind blow: 2 test drive: 2 home make: 2
^_$	28		bj s: 2 fraiser 's: 2 ham s: 2 alan 's: 2 max 's: 2
D_A	28		**a few: 13** **a little: 11**
P_P	25	1	all over: 3 even though: 3 instead of: 2 even if: 2
			: 2 play dumb: 1
V_P_N	14	6	go to school: 2 put at ease: 2 be in hands: 2 keep in mind: 1

STREUSLE corpus (Schneider et al., LREC 2014)

What’s Missing?

These guys took Customer_Service 101 from a Neanderthal.

- FORM: **X 101**, where **X** is a concept or skill that can be learned
- FUNCTION: name of the *most introductory course* on the topic of **X** in an institution of higher learning (based on a naming convention common at U.S. universities)
  - Idiomatic construction requires “101” even though some universities count from 100
- STREUSLE does not annotate X+101 as an MWE because only “101” is fixed, so misses this idiom.

The X-101 Construction

These guys took Customer_Service 101 from a Neanderthal.

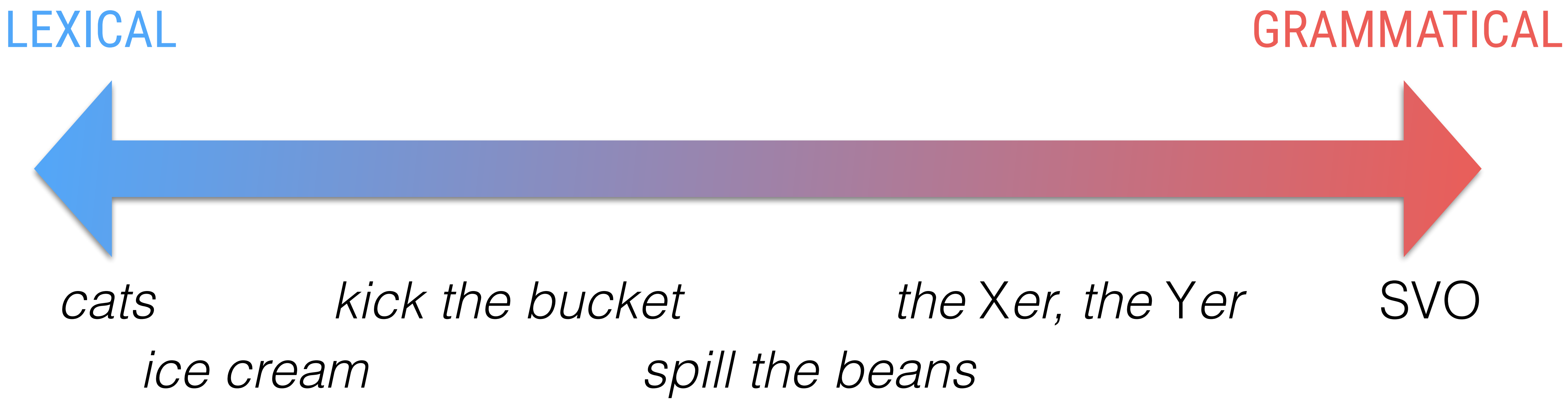
- FORM: **X 101**, where **X** is a concept or skill that can be learned
- FUNCTION: name of the *most introductory course* on the topic of **X** in an institution of higher learning (based on a naming convention common at U.S. universities)
    - Idiomatic construction requires “101” even though some universities count from 100
- STREUSLE does not annotate X+101 as an MWE because only “101” is fixed, so misses this idiom.

Constructions

- In Construction Grammar frameworks,
  construction = any symbolic pairing of **form** and **meaning**
  - Form may be a morpheme, word, multiword expression, syntactic construction, …
  - Knowing a language entails knowing a network of constructions, and how they can be deployed to produce/interpret utterances

Lexicon–Grammar as a Spectrum

Construction Grammar posits continuity between lexicon and grammar



LEXICAL

GRAMMATICAL

cats

ice cream

kick the bucket

spill the beans

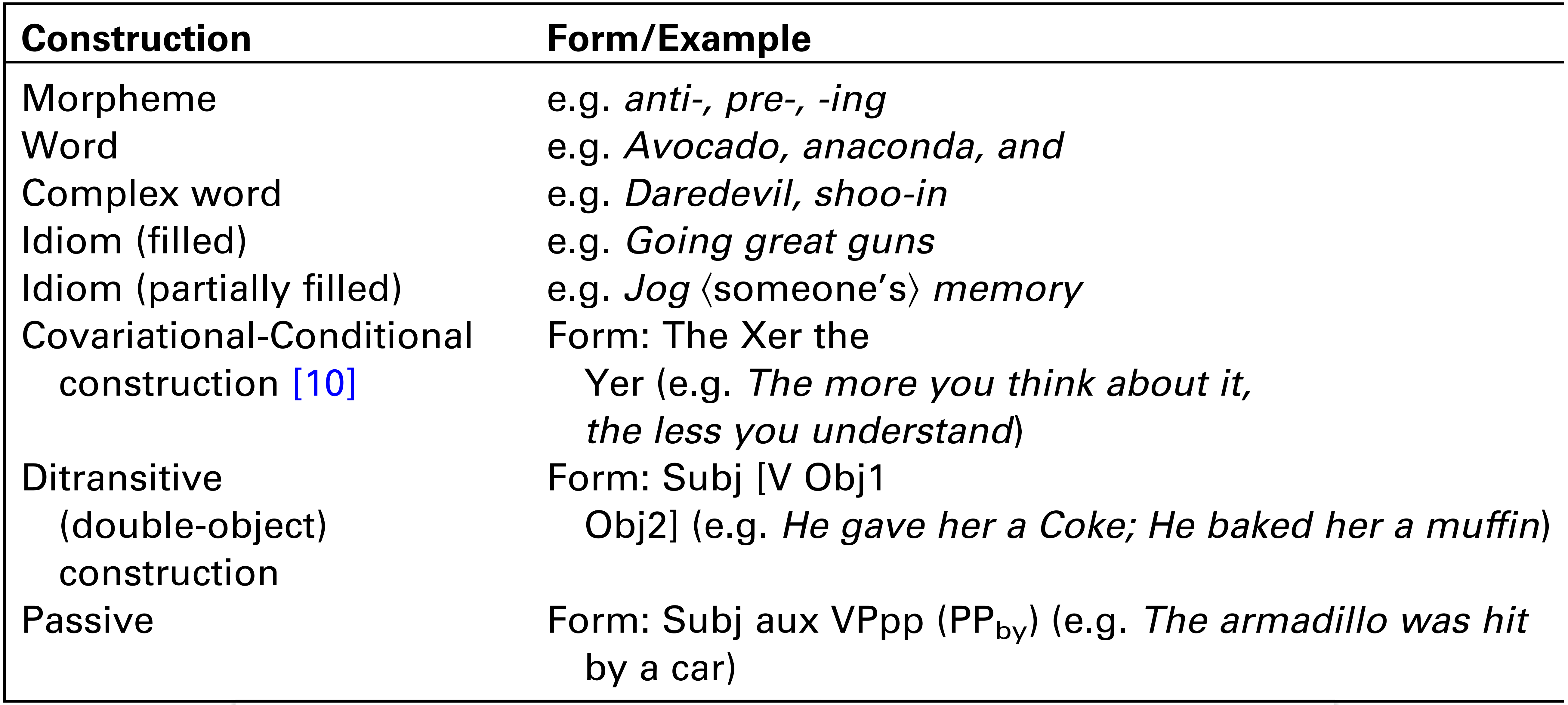
the Xer, the Yer

SVO

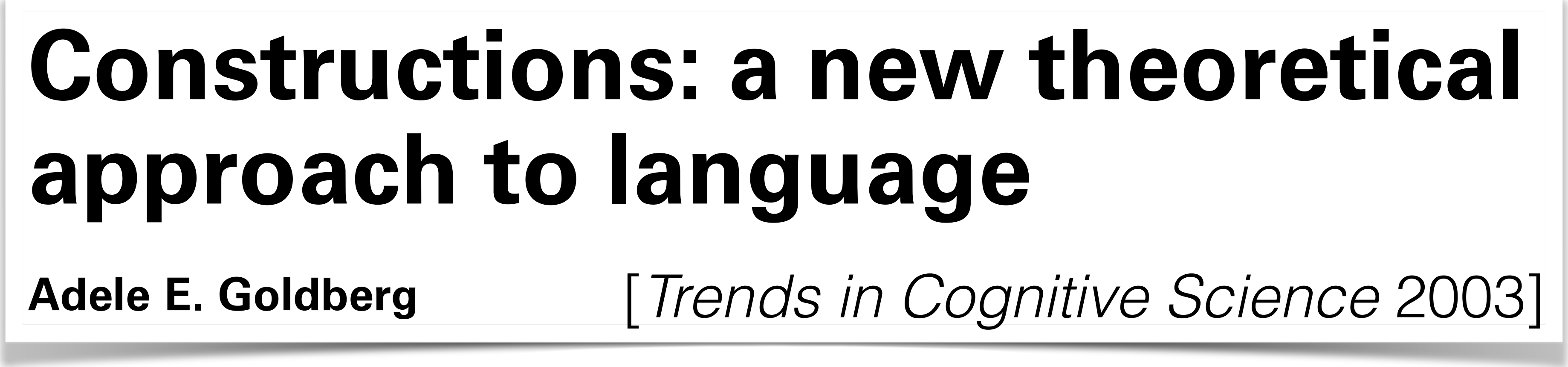
construction = conventionalized **form/function** pairing of any grammatical shape, level of abstractness

constructicon = structured inventory of constructions characterizing knowledge of a language

Constructions



Construction	Form/Example
Morpheme	e.g. *anti-, pre-, -ing*
Word	e.g. *Avocado, anaconda, and*
Complex word	e.g. *Daredevil, shoo-in*
Idiom (filled)	e.g. *Going great guns*
Idiom (partially filled)	e.g. *Jog* ⟨someone's⟩ *memory*
Covariational-Conditional construction [10]	Form: The Xer the Yer (e.g. *The more you think about it, the less you understand*)
Ditransitive (double-object) construction	Form: Subj [V Obj1 Obj2] (e.g. *He gave her a Coke; He baked her a muffin*)
Passive	Form: Subj aux VPpp (PP$_{by}$) (e.g. *The armadillo was hit* by a car)



Constructions: a new theoretical approach to language

Adele E. Goldberg

[*Trends in Cognitive Science* 2003]

Constructions in Typology

- Constructions can be used as a device to account for *intra*linguistic form–meaning mappings

- How to compare and contrast across languages? Need **comparative concepts** to map (Haspelmath, Croft)

  - Consider predication of a person's age:

    - ✳ English recruits the copular construction for adjectival predication:

      I **am** 20 years old

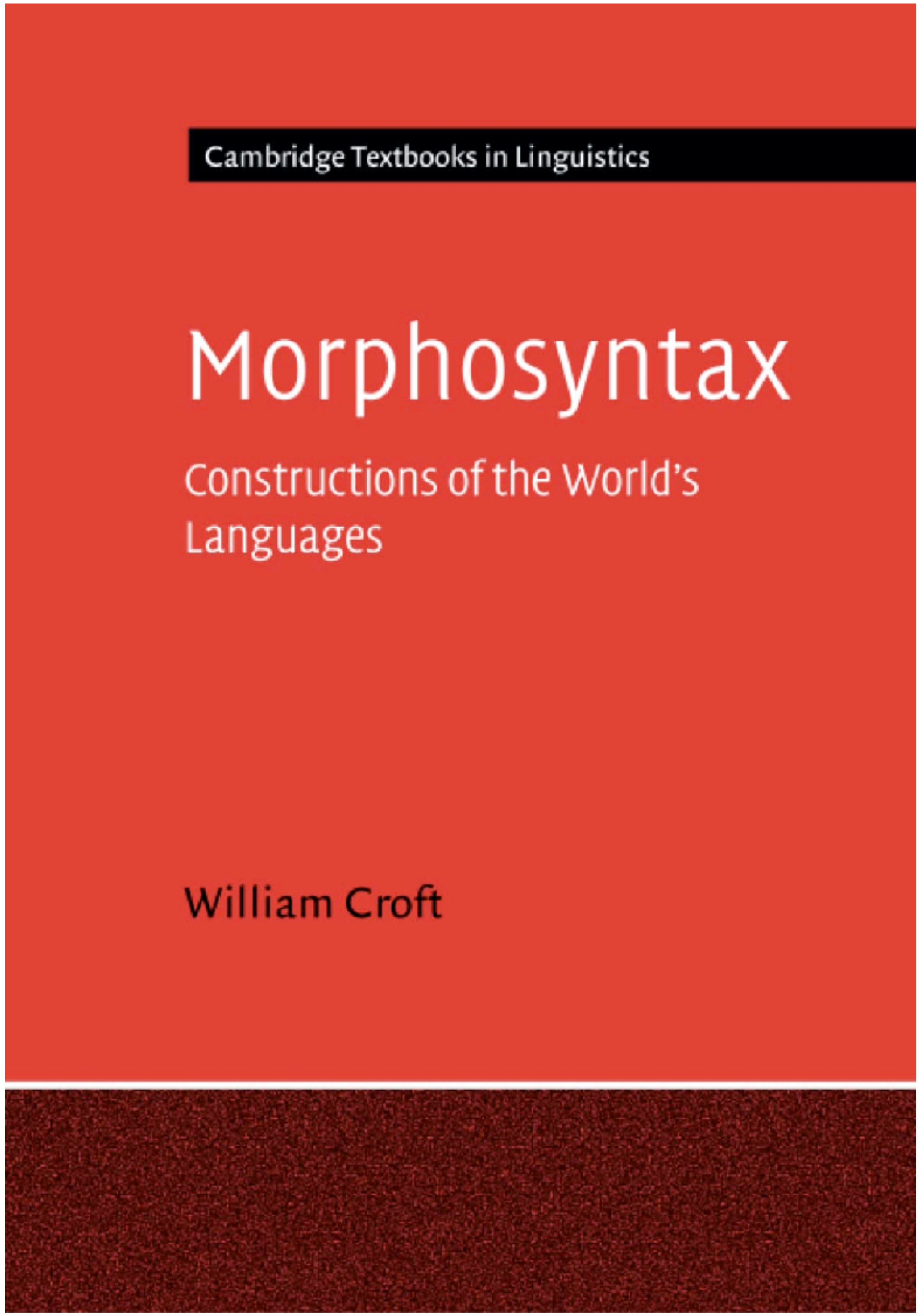
    - ✳ French recruits a verbal possession construction:

      J'**ai** 20 ans lit. 'I **have** 20 years'

  - We can define the comparative concept of 'age predication construction' and distinguish two *strategies* (copular vs. possessive verb)

Constructions in Typology

- Croft (2022) *Morphosyntax*: a deep dive into morphosyntactic constructions of the world's languages



Cambridge Textbooks in Linguistics

Morphosyntax

Constructions of the World's Languages

William Croft

Questions

1. How can we apply **general** syntactic categories to instances of **idiosyncratic** constructions?

2. How can we **annotate** instances of constructions in a **crosslinguistic** way?

3. How well do LMs implicitly capture constructions' **form and meaning**?

Beautiful Washington D.C.

“There was **bumper to bumper** traffic all the way into the city”

Image Credit: https://www.routific.com/us-cities-with-worst-traffic

Slide from Wesley Scivetti

Door-to-door Salesman

- In the previous example, “bumper to bumper” references the closeness of bumpers to one another.
- Is that was “door to door” means here?
- Not really, it seems to reference the movement of a salesperson from one door to the next.

Image Credit: https://www.routific.com/us-cities-with-worst-traffic

Slide from Wesley Scivetti

NPN

- English (+ many other languages) have constructions that consist of a noun, a preposition, and the same noun again:
  - bumper to bumper
  - door to door
  - day to day
  - day by day
  - day after day
  - review assignment upon review assignment
- Meanings related to juxtaposition, quantity, iteration (depending on preposition and noun)

NPN: Jackendoff (2008)

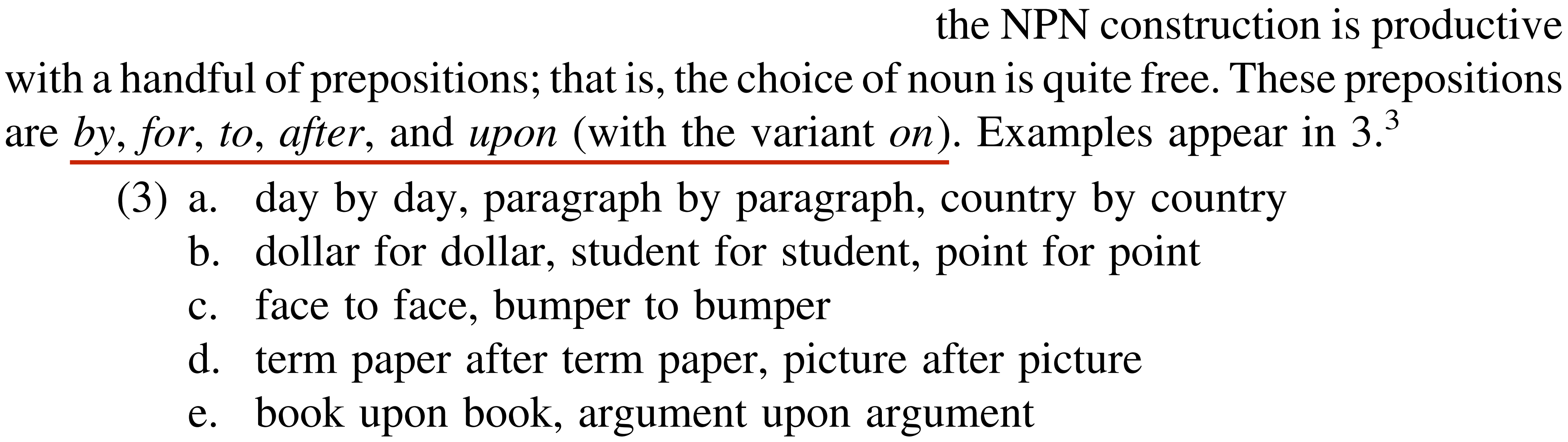
- “The basic insight of construction grammar is that languages can contain numerous offbeat pieces of syntax with idiosyncratic interpretations.”

- NPN construction, e.g. *day by day* is an example of idiosyncratic syntax+semantics
  - Actually a **family** of constructions

Idiosyncratic Restrictions on Form

Though an NP could consist of “N P N”, instances of the NPN construction don’t appear in NP-like contexts (more like PP contexts)

In addition to some fixed expressions like “tongue in cheek” and “head over heels”,



the NPN construction is productive with a handful of prepositions; that is, the choice of noun is quite free. These prepositions are *by*, *for*, *to*, *after*, and *upon* (with the variant *on*). Examples appear in 3.[3]

(3) a. day by day, paragraph by paragraph, country by country
b. dollar for dollar, student for student, point for point
c. face to face, bumper to bumper
d. term paper after term paper, picture after picture
e. book upon book, argument upon argument

Idiosyncratic Restrictions on Form

(4) a. No mass nouns: *water after water, *dust for dust
b. No determiners: *the man for the man, *a day after a day, *some inch by some inch
c. No plurals: *men for men, *books after books, *weeks by weeks
d. No postnominal complements or modifiers: *father of a soldier for father of a soldier, *day of rain to day of rain, *inch of steel pipe by inch of steel pipe
e. —except with *after* and *upon*: day after day of rain
f. Prenominal adjectives: day after miserable day, tall boy by tall boy

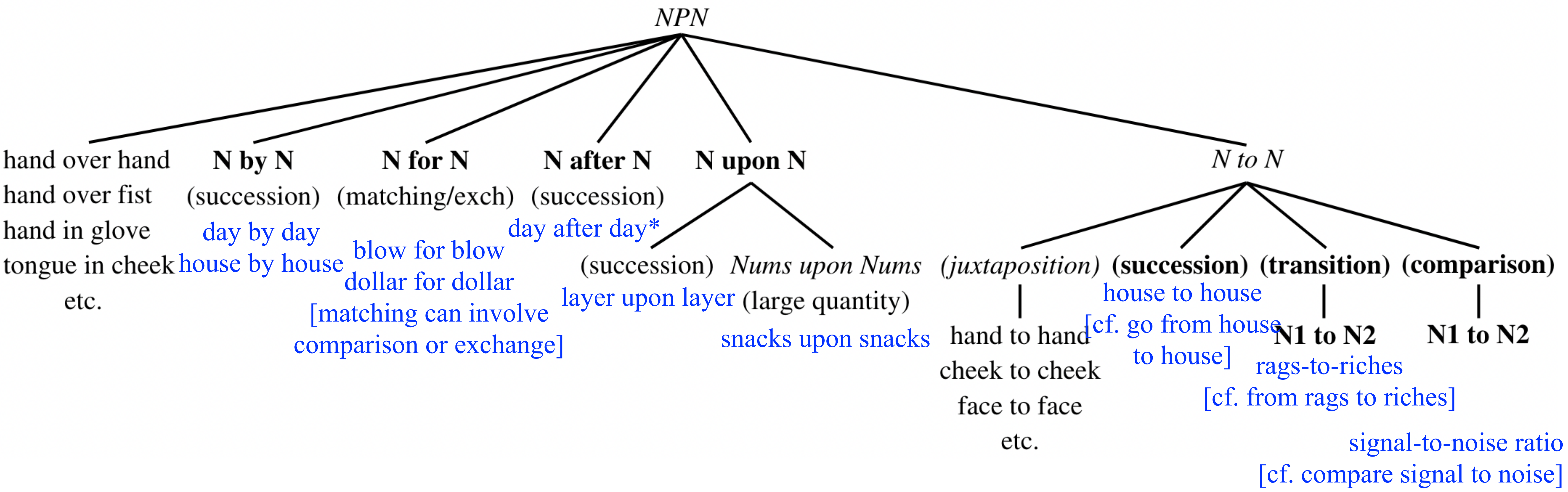
NPNs: Meaning

- NPN constructions are really a family of constructions with similar forms but different meanings

- Commonly attested meanings (Jackendoff 2008, Roch et al. 2010, Sommerer and Baumann 2021)
  - **Succession/Iteration**
    - “The plan changes **day to day**”
  - **Comparison/Exchange**
    - “They’re the best **pound for pound** boxer”
  - **Juxtaposition/Close Contact**
    - “The two stood **chest to chest**”
  - **Intensification/Quantification**
    - “I graded **essays upon essays**”

Slide from Wesley Scivetti

Cxn Network (Hierarchy)



NPN

hand over hand
hand over fist
hand in glove
tongue in cheek
etc.

N by N
(succession)
day by day
house by house

N for N
(matching/exch)
blow for blow
dollar for dollar
[matching can involve
comparison or exchange]

N after N
(succession)
day after day*

N upon N

(succession)
layer upon layer

Nums upon Nums
(large quantity)
snacks upon snacks

N to N

(juxtaposition)
hand to hand
cheek to cheek
face to face
etc.

(succession)
house to house
[cf. go from house
to house]

(transition)
N1 to N2
rags-to-riches
[cf. from rags to riches]

(comparison)
N1 to N2
signal-to-noise ratio
[cf. compare signal to noise]

*growing intensity:
We worked day after week after month
*We worked month after week after day

NPN Annotation

- NPNs are rare
- Shallow matching will give false positives
  - *problem of sticking plastic to plastic*
- We want a way to annotate
  - where an instance of the NPN construction occurs, and
  - what its meaning is

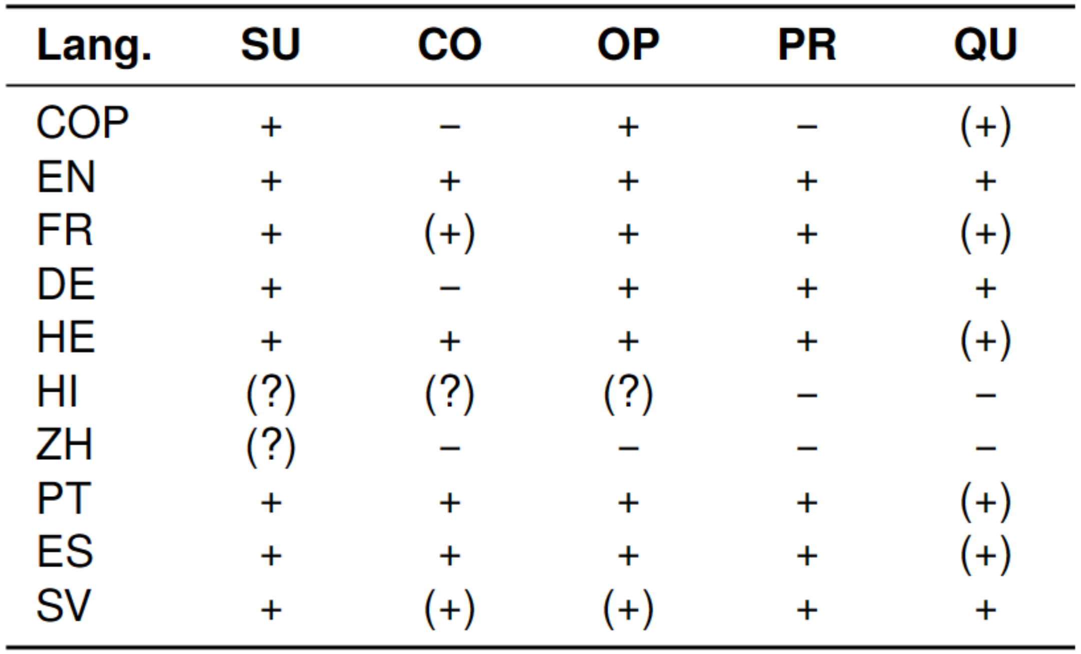
NPN Datasets

- **UCxn** project: investigated NPNs in 10 languages (Weissweiler et al. LREC-COLING 2024)
  - Found treebank attestations in 8 languages by querying UD treebanks

NPNs

- **Strategy → one form, multiple possible meanings**
- **Day after day, shoulder to shoulder, box upon box**
- **Easy to automatically annotate**

Analysis of attested meanings



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

- 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

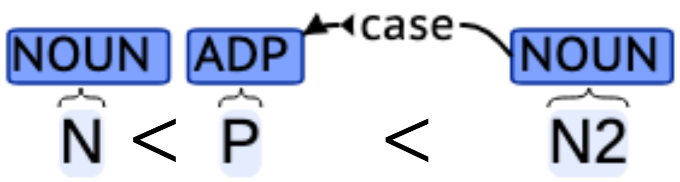
UCxn: Weissweiler et al. (2024)

Slide from Leonie Weissweiler

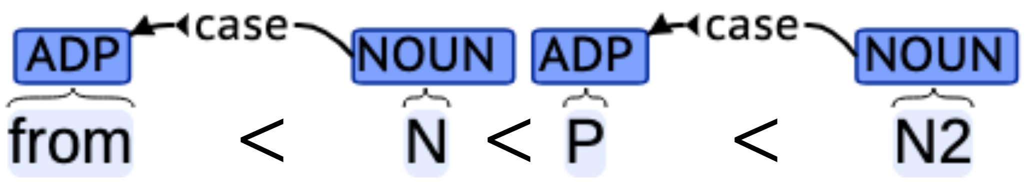
Syntactic Querying

- Dependency syntax already provides the scaffolding for recognizing instances of the formal pattern

- Queries like the following can be adapted to different languages:



without:



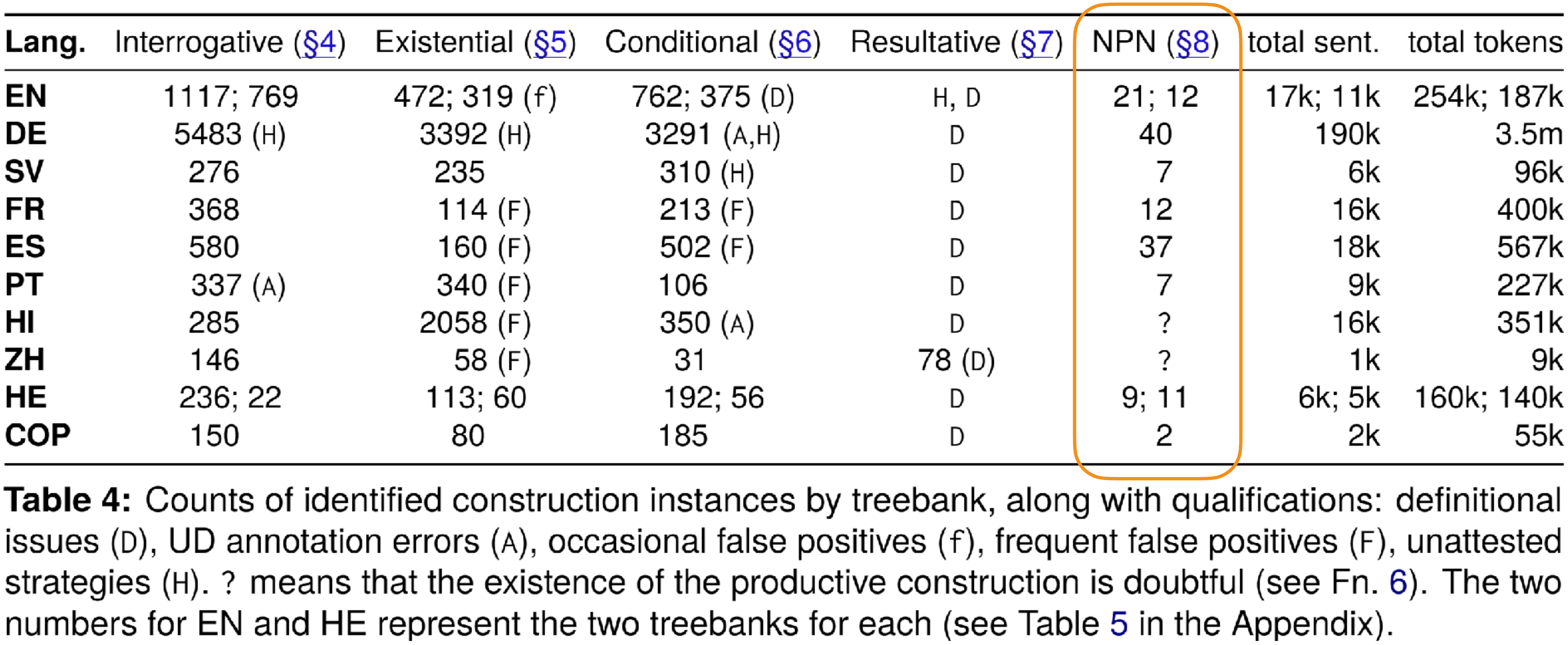
(N, P, N2 are variables; < indicates successive words)

Grew Query (English)

```
rule npn {
    pattern {
        _anchor_ [ xpos = re"N.*"];
        N2 [xpos = re"N.*" ];
        _anchor_.lemma = N2.lemma;
        _anchor_ -> N2;
        P[upos="ADP"];
        _anchor_ < P;
        P < N2;
    }
    without {
        N -[case]-> P2;
        P2 < N;
        P2 [lemma="from"];
    }
    without { X -[fixed]-> _anchor_ }

    commands { _anchor_.Cxn="NPN";
               _anchor_.CxnElt="N1";
              P.CxnElt="P";
              N2.CxnElt=“N2"; }
}
```

UCxn V1: A New Resource



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

Table 4: 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 (see Fn. 6). The two numbers for EN and HE represent the two treebanks for each (see Table 5 in the Appendix).

UCxn: Weissweiler et al. (2024)

Slide from Leonie Weissweiler

NPN Datasets

- **UCxn** project: investigated NPNs in 10 languages (Weissweiler et al. LREC-COLING 2024)
  - Found treebank attestations in 8 languages by querying UD treebanks

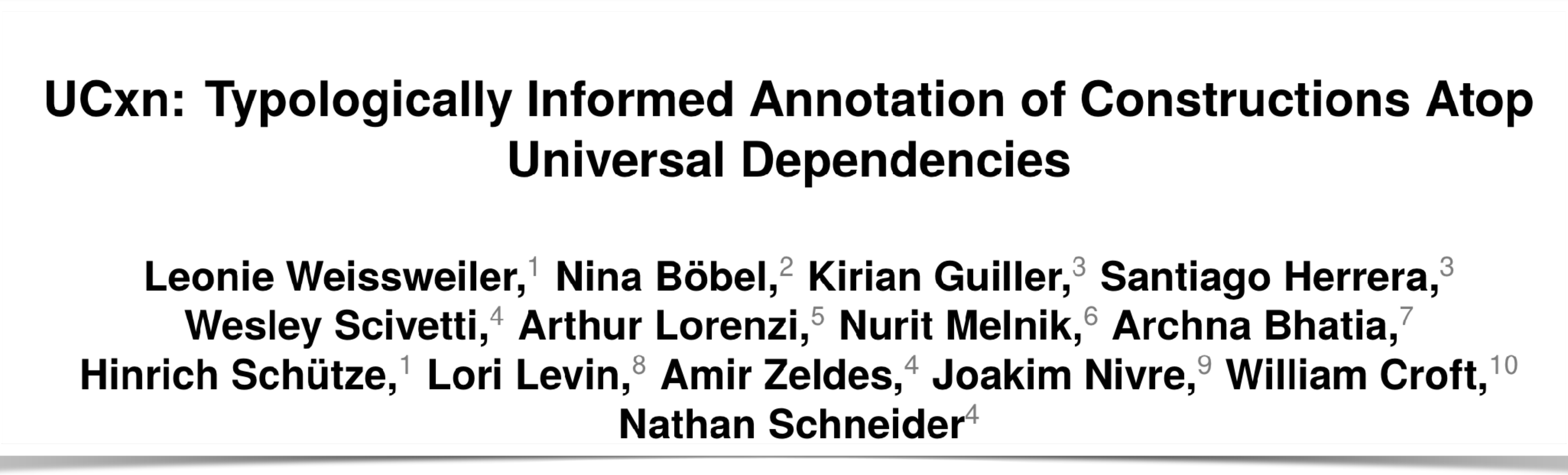
- Semantic disambiguation of a larger sample of **N-*to*-N** instances in English (Scivetti & Schneider 2025 preprint)
  - Sampled from COCA: 6600 true instances + 450 distractors
  - Labeled as SUCCESSION vs. JUXTAPOSITION meaning

Typological Construction Annotation via Comparative Concepts

- Beyond NPN, which is a *strategy*, the UCxn project defined 4 other constructions in **functional** terms and investigated them across UD corpora in 10 languages:
  - Interrogatives
  - Existentials
  - Conditionals
  - Resultatives
- These are towards the grammatical end of the spectrum, but they still have meanings/functions!

UCxn: Weissweiler et al. (2024)

UCxn: A multilingual collaboration!



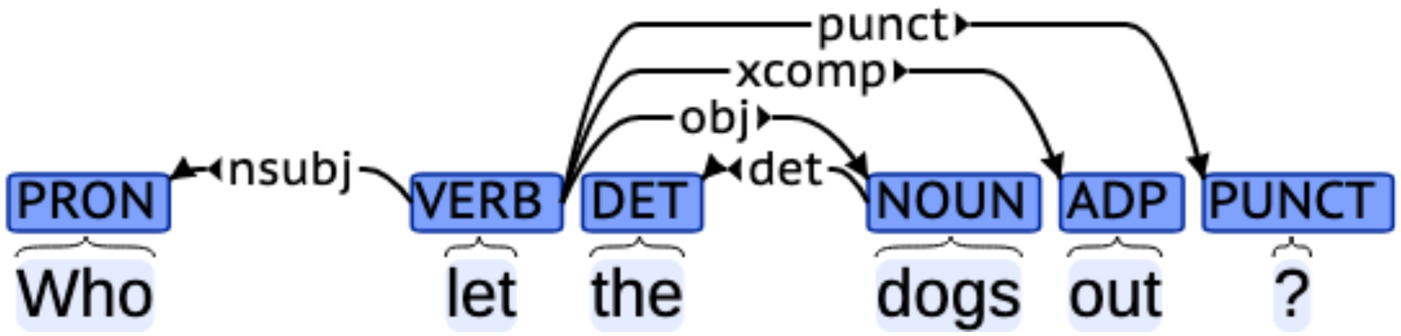
UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

Leonie Weissweiler,[1] Nina Böbel,[2] Kirian Guiller,[3] Santiago Herrera,[3] Wesley Scivetti,[4] Arthur Lorenzi,[5] Nurit Melnik,[6] Archna Bhatia,[7] Hinrich Schütze,[1] Lori Levin,[8] Amir Zeldes,[4] Joakim Nivre,[9] William Croft,[10] Nathan Schneider[4]

English, German, Swedish, French, Spanish, Brazilian Portuguese, Hindi, Chinese, Hebrew, Coptic

Interrogatives

- What would appear in a UD treebank:

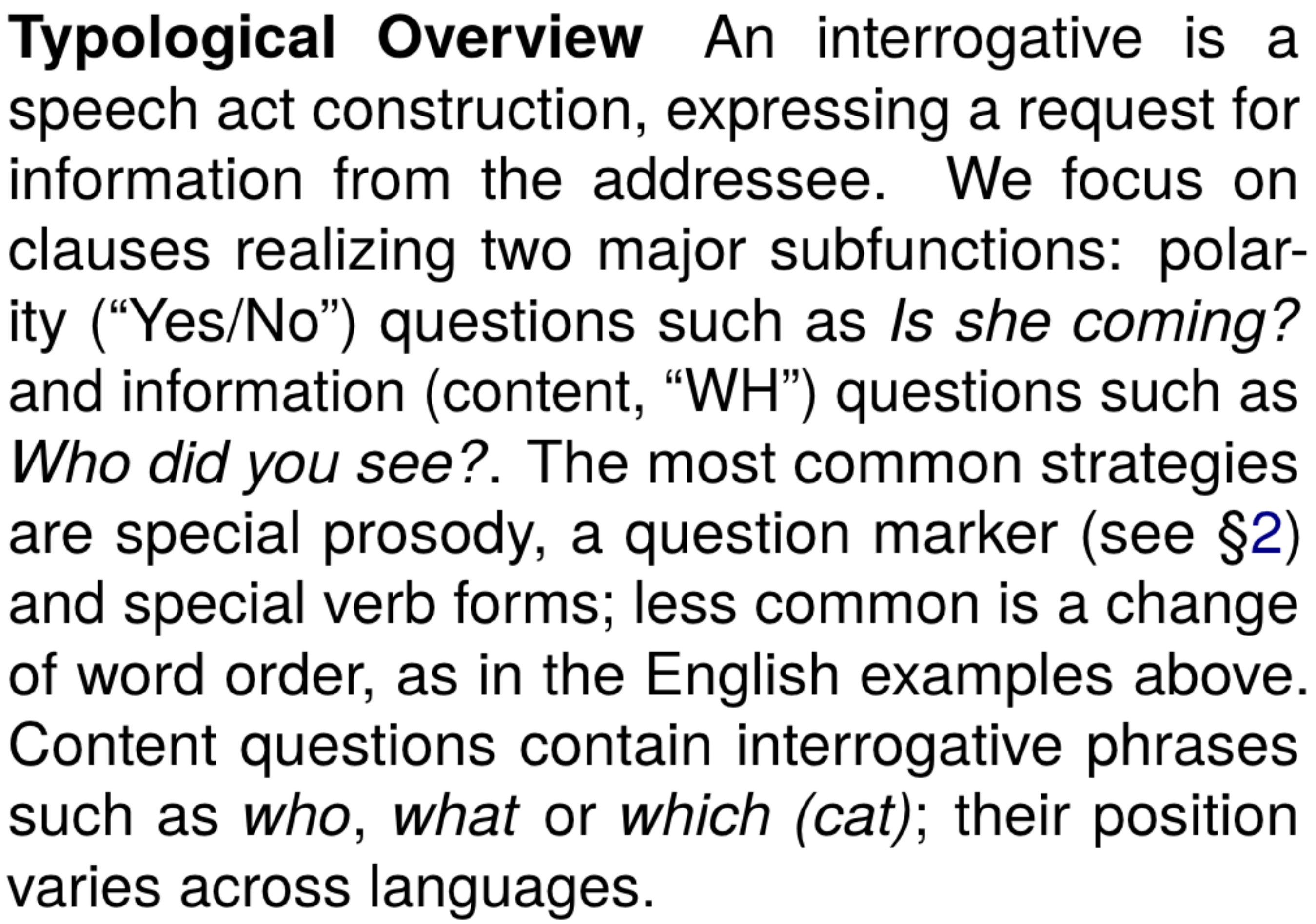


- Nowhere does it say that this is a question, because that's indicated by a combination of morphosyntactic features (word order, WH-words etc.) and orthographic features (punctuation).

- We adopt functional definitions of different kinds of interrogatives and write language-specific queries to match formal strategies.

UCxn: Weissweiler et al. (2024)

Interrogatives

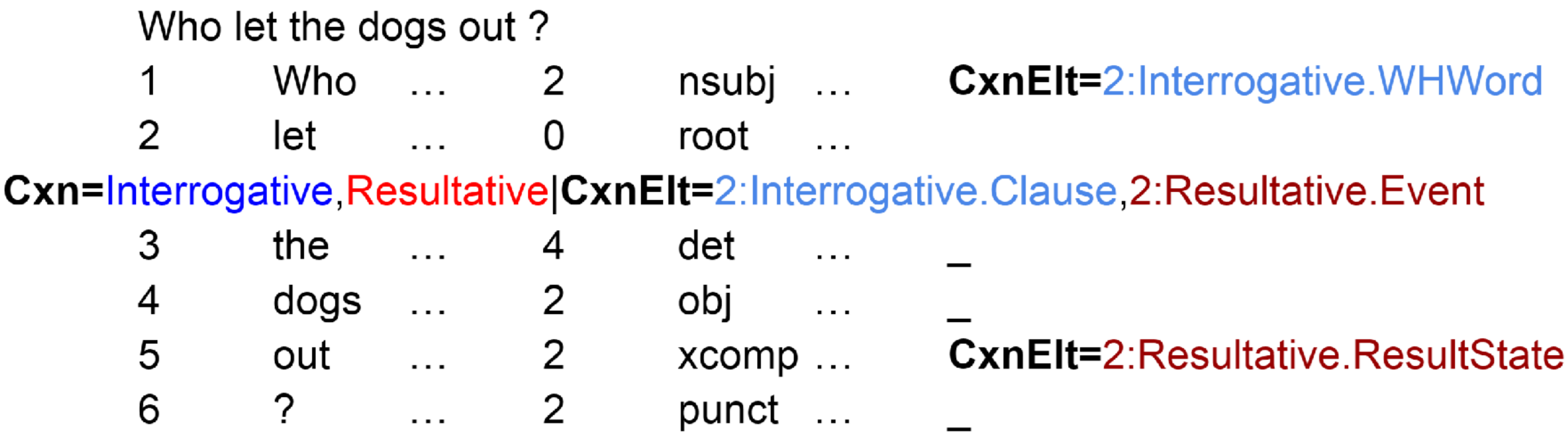


Typological Overview An interrogative is a speech act construction, expressing a request for information from the addressee. We focus on clauses realizing two major subfunctions: polarity ("Yes/No") questions such as *Is she coming?* and information (content, "WH") questions such as *Who did you see?*. The most common strategies are special prosody, a question marker (see §2) and special verb forms; less common is a change of word order, as in the English examples above. Content questions contain interrogative phrases such as *who*, *what* or *which (cat)*; their position varies across languages.

UCxn: Weissweiler et al. (2024)

Interrogatives

- UCxn guidelines offer a set of names for construction subtypes and elements of the construction, and a standard for marking instances in CoNLL-U format:

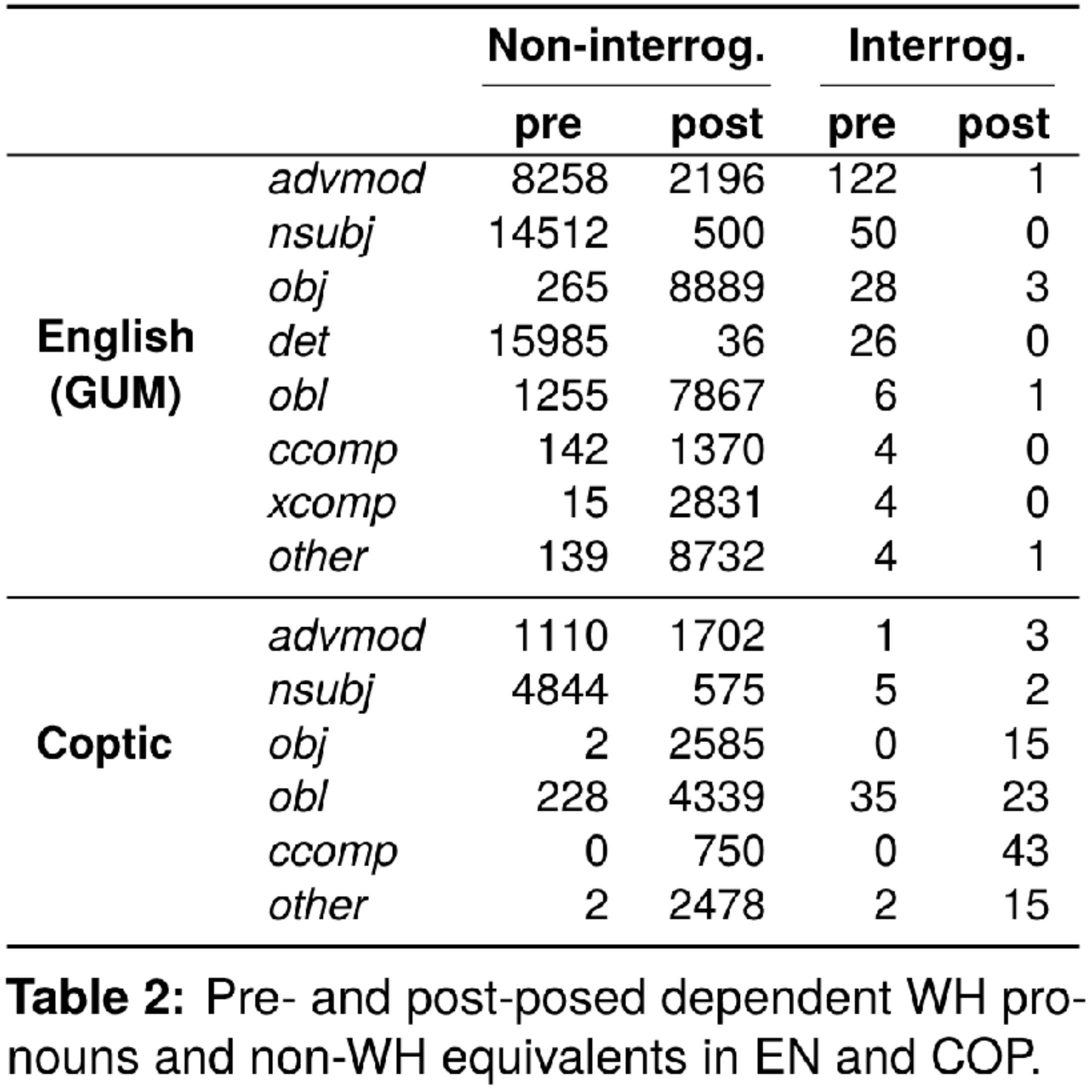


```
        Who let the dogs out ?
        1       Who     …       2       nsubj   …       CxnElt=2:Interrogative.WHWord
        2       let     …       0       root    …
Cxn=Interrogative,Resultative|CxnElt=2:Interrogative.Clause,2:Resultative.Event
        3       the     …       4       det     …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

UCxn: Weissweiler et al. (2024)

Quantitative Comparisons

- Annotations allow for fine-grained comparison of strategies, e.g. how WH words are realized in English vs. Coptic:

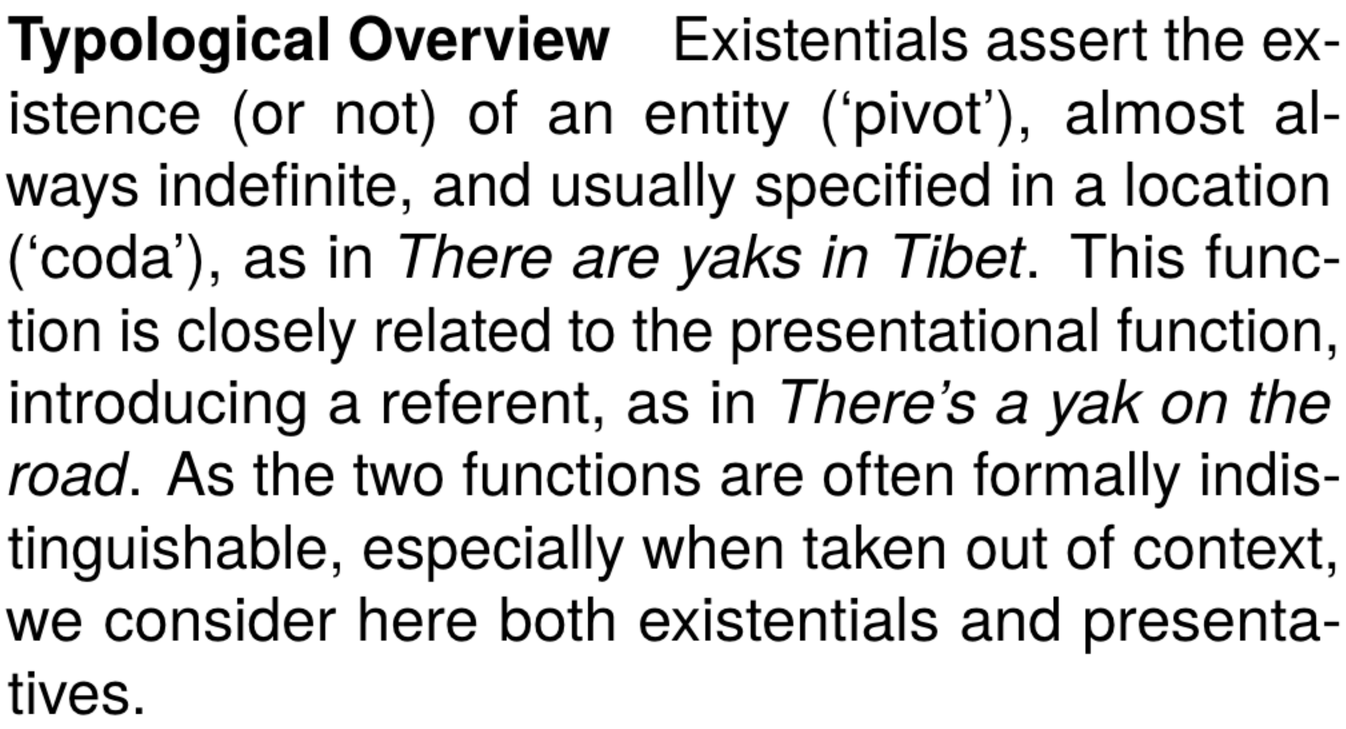


		Non-interrog.		Interrog.	
		pre	post	pre	post
English (GUM)	*advmod*	8258	2196	122	1
	nsubj	14512	500	50	0
	obj	265	8889	28	3
	det	15985	36	26	0
	obl	1255	7867	6	1
	ccomp	142	1370	4	0
	xcomp	15	2831	4	0
	other	139	8732	4	1
Coptic	*advmod*	1110	1702	1	3
	nsubj	4844	575	5	2
	obj	2	2585	0	15
	obl	228	4339	35	23
	ccomp	0	750	0	43
	other	2	2478	2	15

Table 2: Pre- and post-posed dependent WH pronouns and non-WH equivalents in EN and COP.

UCxn: Weissweiler et al. (2024)

Existentials/presentationals

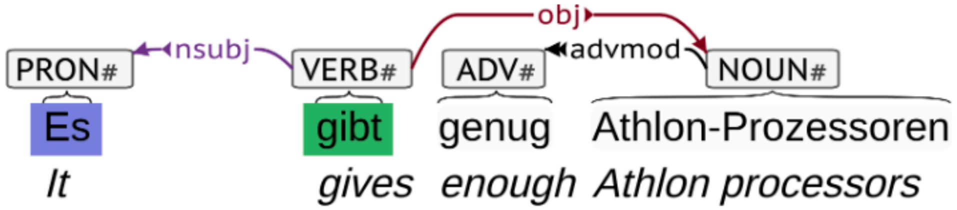


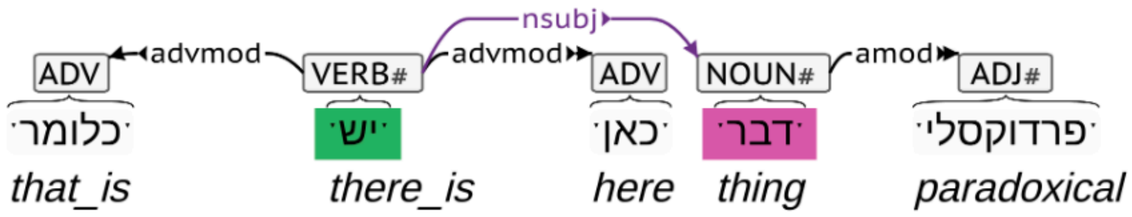
Typological Overview Existentials assert the existence (or not) of an entity (‘pivot’), almost always indefinite, and usually specified in a location (‘coda’), as in *There are yaks in Tibet*. This function is closely related to the presentational function, introducing a referent, as in *There’s a yak on the road*. As the two functions are often formally indistinguishable, especially when taken out of context, we consider here both existentials and presentatives.

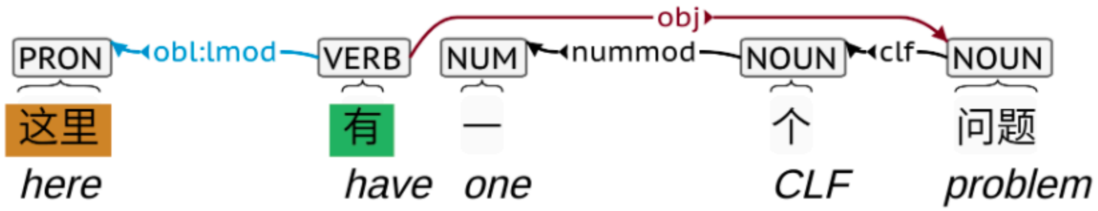
UCxn: Weissweiler et al. (2024)

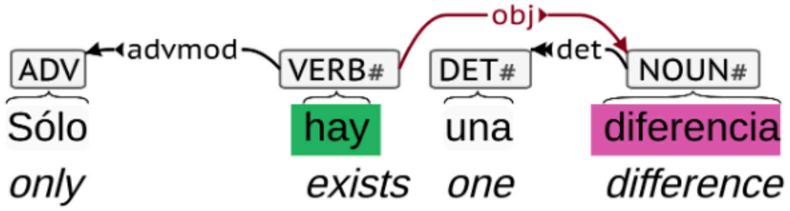
Existentials/presentationals

Language	Instance	Query

German	 obj
	nsubj 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 ;

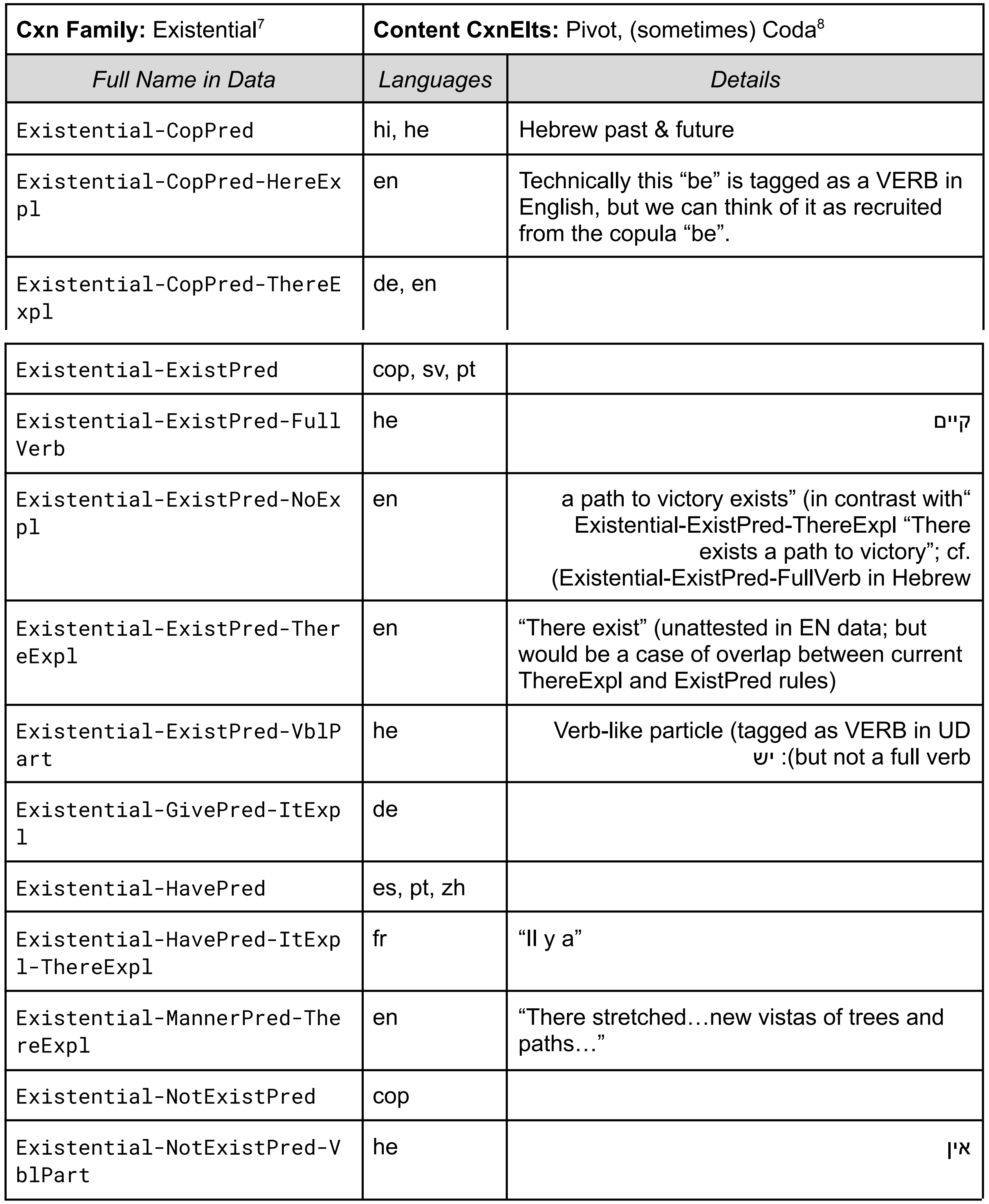
Hebrew	 nsubj
	advmod advmod amod
	ADV VERB# ADV NOUN# ADJ#
	כלומר יש כאן דבר פרדוקסלי
	that_is there_is here thing paradoxical
		pattern
		PRED [lemma="יש"];
		PRED -[nsubj]-> PIV ;
		without
		LE[lemma="ל"];
		PRED -[obl]->N; N-[case]->LE;

Mandarin	 obj obl:lmod 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	 obj
	advmod 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;

UCxn: Weissweiler et al. (2024)

Existentials/presentationals



Cxn Family: Existential[7]	**Content CxnElts:** Pivot, (sometimes) Coda[8]	
Full Name in Data	*Languages*	*Details*
Existential-CopPred	hi, he	Hebrew past & future
Existential-CopPred-HereExpl	en	Technically this “be” is tagged as a VERB in English, but we can think of it as recruited from the copula “be”.
Existential-CopPred-ThereExpl	de, en	
Existential-ExistPred	cop, sv, pt	
Existential-ExistPred-FullVerb	he	קיים
Existential-ExistPred-NoExpl	en	a path to victory exists” (in contrast with“ Existential-ExistPred-ThereExpl “There exists a path to victory”; cf. (Existential-ExistPred-FullVerb in Hebrew
Existential-ExistPred-ThereExpl	en	“There exist” (unattested in EN data; but would be a case of overlap between current ThereExpl and ExistPred rules)
Existential-ExistPred-VblPart	he	Verb-like particle (tagged as VERB in UD יש :(but not a full verb
Existential-GivePred-ItExpl	de	
Existential-HavePred	es, pt, zh	
Existential-HavePred-ItExpl-ThereExpl	fr	“Il y a”
Existential-MannerPred-ThereExpl	en	“There stretched…new vistas of trees and paths…”
Existential-NotExistPred	cop	
Existential-NotExistPred-VblPart	he	אין

UCxn v1 Guidelines

Questions

1. How can we apply **general** syntactic categories to instances of **idiosyncratic** constructions?

2. How can we **annotate** instances of constructions in a **crosslinguistic** way?

   💡 Develop functional definitions of constructions as comparative concepts

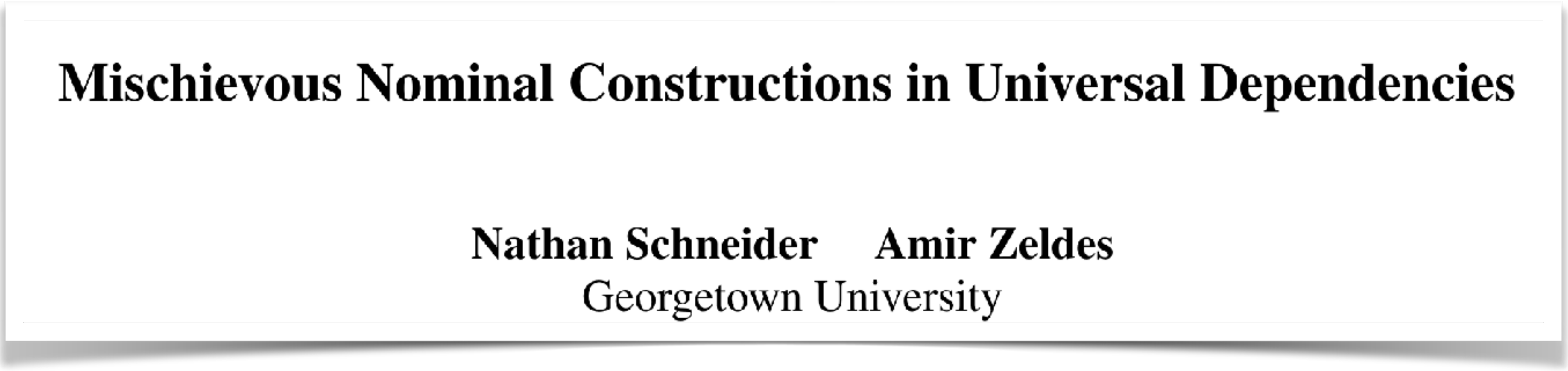
   💡 Query syntactic treebanks to identify form pattern

      ✦ There may be several such patterns, even within the same language

   💡 Manually disambiguate senses

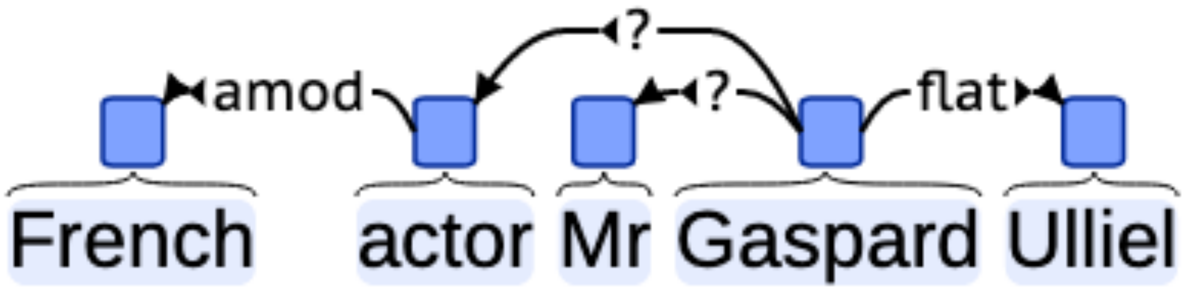
3. How well do LMs implicitly capture constructions' **form and meaning**?

What if the basic syntax is in doubt?



Mischievous Nominal Constructions in Universal Dependencies

Nathan Schneider **Amir Zeldes**
Georgetown University



What is the syntactic head?
Lake Michigan
Chapter 1

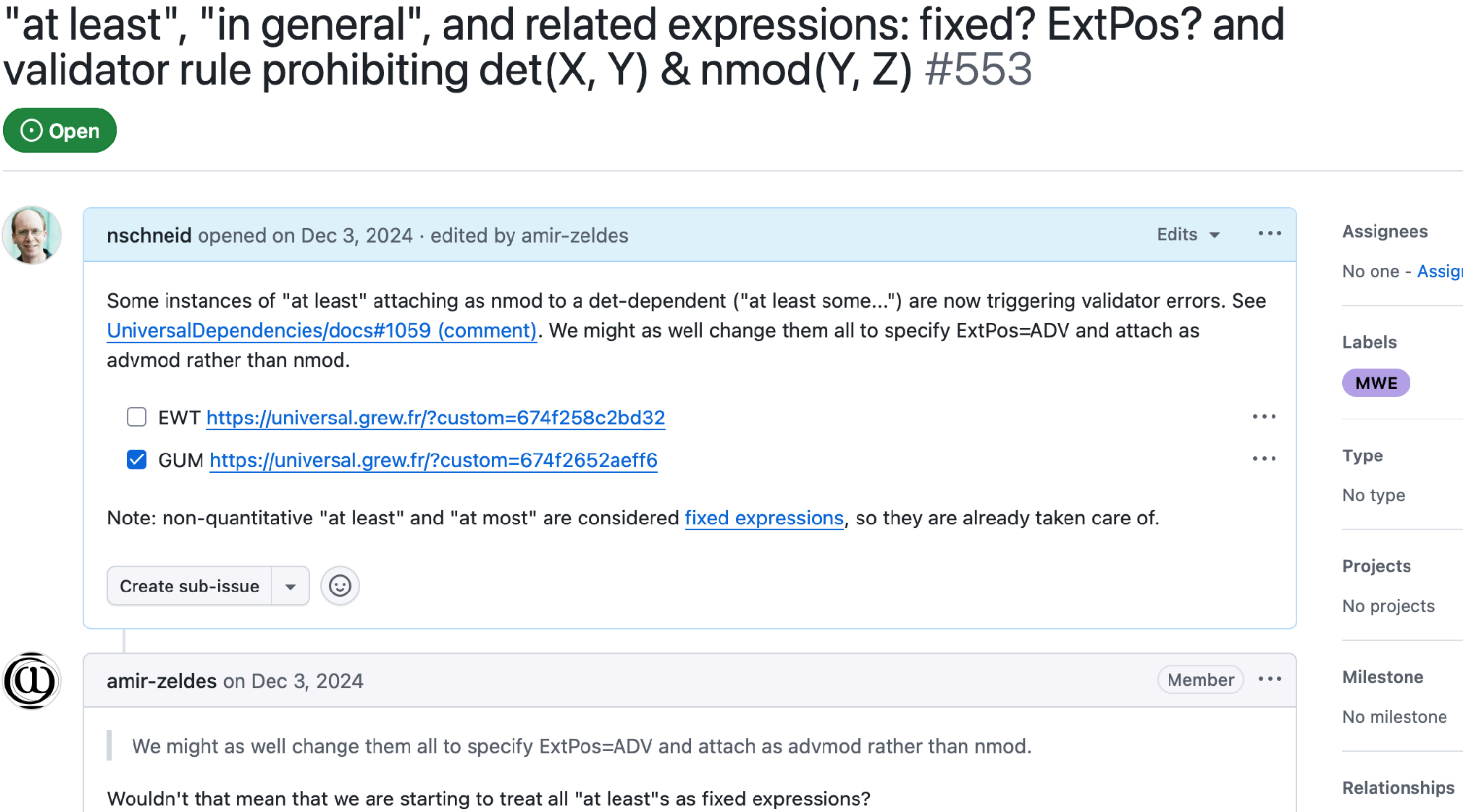
A category-inventory like UD relations needs elasticity

- Revised some of the universal guidelines to be more flexible and prototype-based:
  - The **flat** relation is used to combine the elements of an expression where none of the immediate components can be identified as the sole head using standard substitution tests.…The prototypes for **flat** are: (i) personal names, (ii) foreign expressions, (iii) iconic sequences, and (iv) items separated for readability.

- Language-specific subtypes
  - [en] **nmod:desc**: descriptor modifier in nominal

    This relation subtype applies to nominal modifiers that we term **descriptors**. These are bare nominals that occur in or with a name, and are not prepositional/possessive or part of the English compound construction. For personal names, titles and role descriptions are a prime example.…

What counts as a fixed grammatical expression in UD?



"at least", "in general", and related expressions: fixed? ExtPos? and validator rule prohibiting det(X, Y) & nmod(Y, Z) #553

Open

nschneid opened on Dec 3, 2024 · edited by amir-zeldes

Edits ···

Some instances of "at least" attaching as nmod to a det-dependent ("at least some...") are now triggering validator errors. See UniversalDependencies/docs#1059 (comment). We might as well change them all to specify ExtPos=ADV and attach as advmod rather than nmod.

- ☐ EWT https://universal.grew.fr/?custom=674f258c2bd32 ···
- ☑ GUM https://universal.grew.fr/?custom=674f2652aeff6 ···

Note: non-quantitative "at least" and "at most" are considered fixed expressions, so they are already taken care of.

Create sub-issue ▾

amir-zeldes on Dec 3, 2024 Member ···

> We might as well change them all to specify ExtPos=ADV and attach as advmod rather than nmod.

Wouldn't that mean that we are starting to treat all "at least"s as fixed expressions?

Assignees

No one - Assig

Labels

MWE

Type

No type

Projects

No projects

Milestone

No milestone

Relationships

What counts as a fixed grammatical expression in UD?

- A few expressions like “as well as” clearly have transcended their historical/compositional syntactic behavior, warranting **fixed**

- For expressions like “at least” or “in order to”, where some internal structure is visible, categorizing as *syntactically* regular vs. idiosyncratic is really hard :(
  - Need more input from PARSEME/UniDive! Cf. Savary et al. (NEJLT 2023)

- Discussions reveal tension between those who favor conservatism (keeping current lexical list to avoid churn) and those who look for a more principled approach

Questions

1. How can we apply **general** syntactic categories to instances of **idiosyncratic** constructions?

   💡 Define general categories with prototypes for elasticity (+ consider subtyping)

2. How can we **annotate** instances of constructions in a **crosslinguistic** way?

   💡 Develop functional definitions of constructions as comparative concepts

   💡 Query syntactic treebanks to identify form pattern

   💡 Manually disambiguate senses

3. How well do LMs implicitly capture constructions' **form and meaning**?

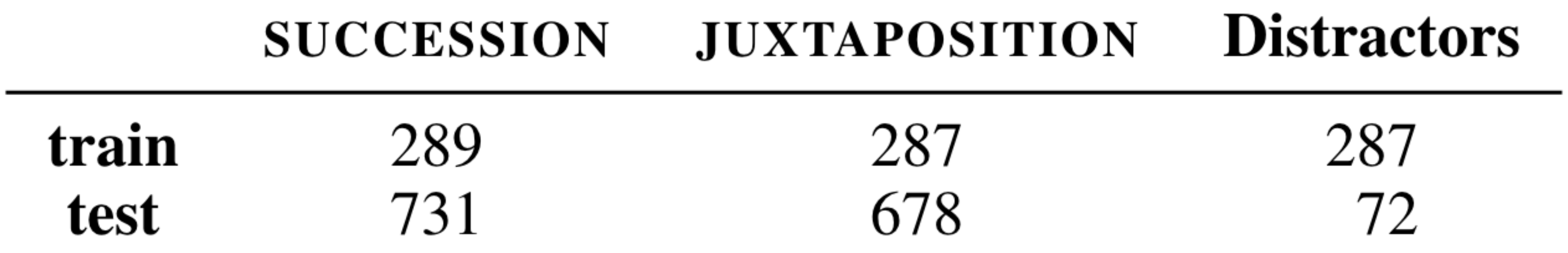
Do LMs “know” constructions?

- A growing body of work on how various language models represent/process various constructions, esp. in English:
    - Weissweiler et al. (2022): comparative correlative (*the X-er, the Y-er*)
    - Zhou et al. (2024): causal excess (*so X that Y*)
    - Misra & Mahowald (2024): AANN (*a ADJ NUM NOUN*)
    - Scivetti et al. (2025): way-manner, let-alone, and others

☞ Construction Grammar + NLP Bibliography

What do LMs “know” about NPN?

- Using the aforementioned dataset of N-*to*-N construction, we train probing classifiers to understand BERT’s contextualized representations of the token “to”
  - Subset of annotated COCA data used for BERT experiments:



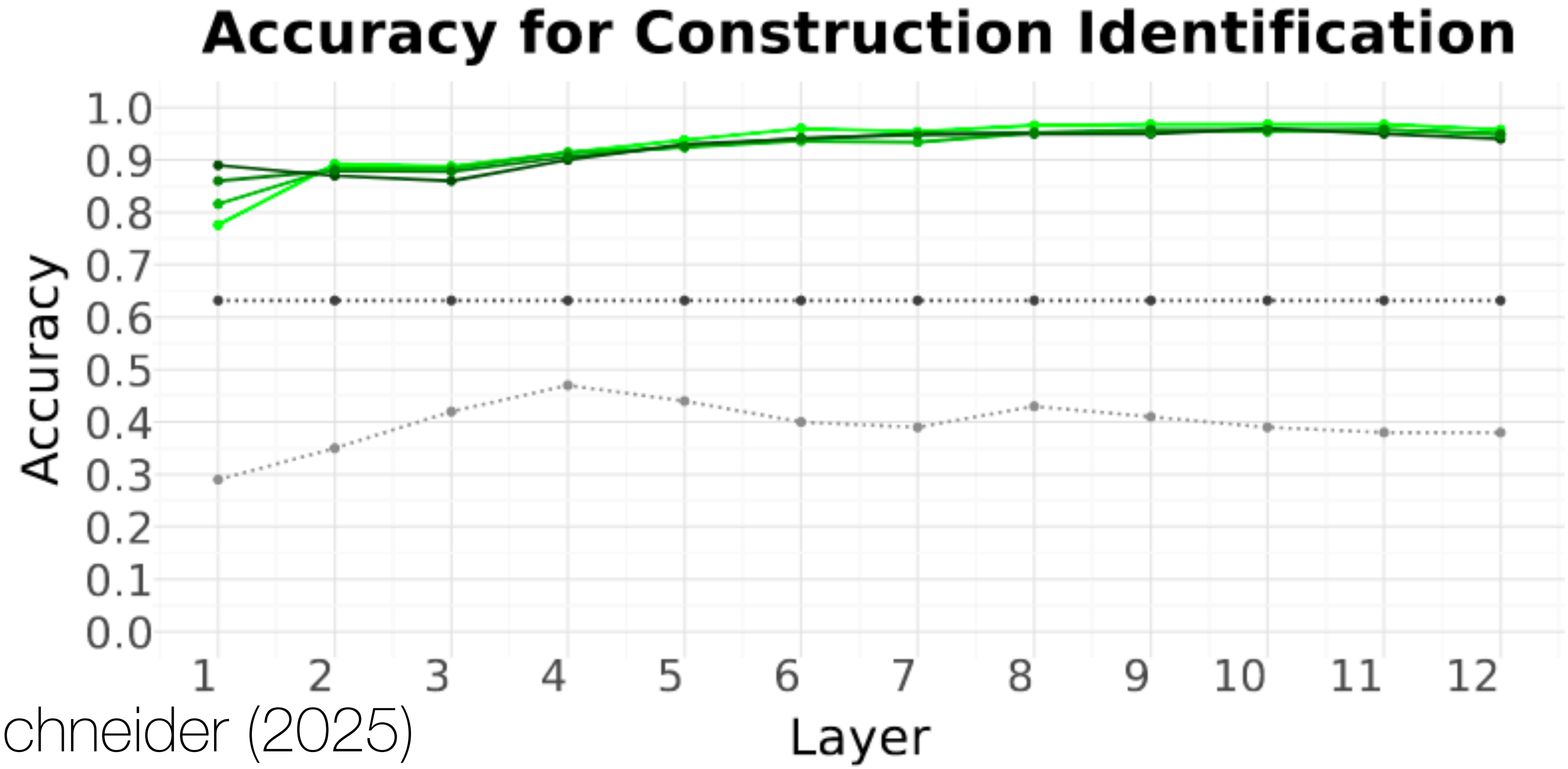
	SUCCESSION	JUXTAPOSITION	Distractors
train	289	287	287
test	731	678	72

  - No lexical overlap of nouns between train and test

Scivetti & Schneider (2025)

What do LMs “know” about NPN?

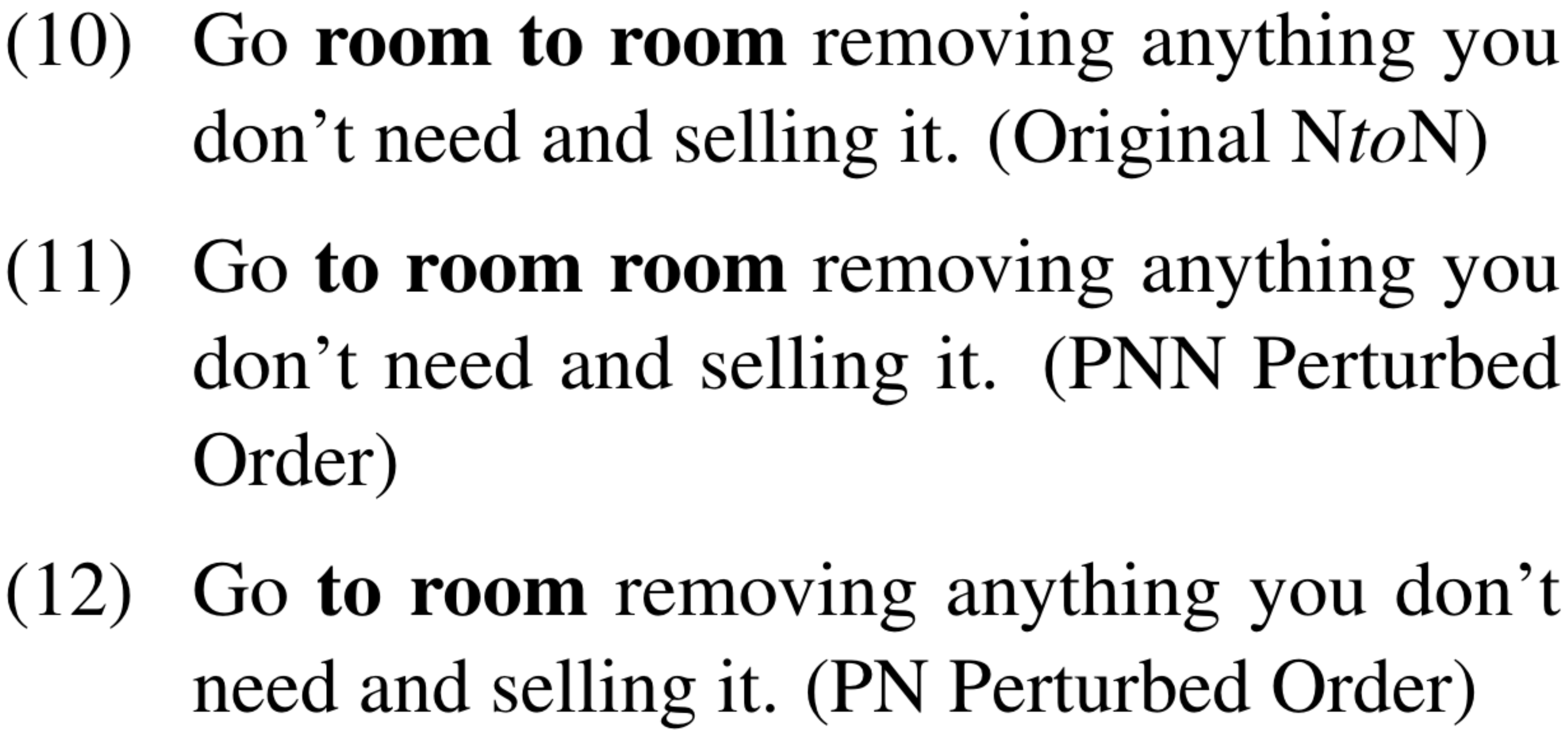
- Q1: Do the representations distinguish true construction instances vs. distractors?
  - ✴ Yes, with accuracy >90% at middle and higher layers, even in few-shot case (10 training examples for probe):



Scivetti & Schneider (2025)

What do LMs “know” about NPN?

- Q2: Are the representations sensitive to word order?



(10) Go **room to room** removing anything you don’t need and selling it. (Original N*to*N)

(11) Go **to room room** removing anything you don’t need and selling it. (PNN Perturbed Order)

(12) Go **to room** removing anything you don’t need and selling it. (PN Perturbed Order)

✴ Somewhat: In higher layers, a majority of perturbed versions are rejected as instances of the construction if there are ≥25 training examples

Scivetti & Schneider (2025)

What do LMs “know” about NPN?

- Q3: Do the representations disambiguate semantics?

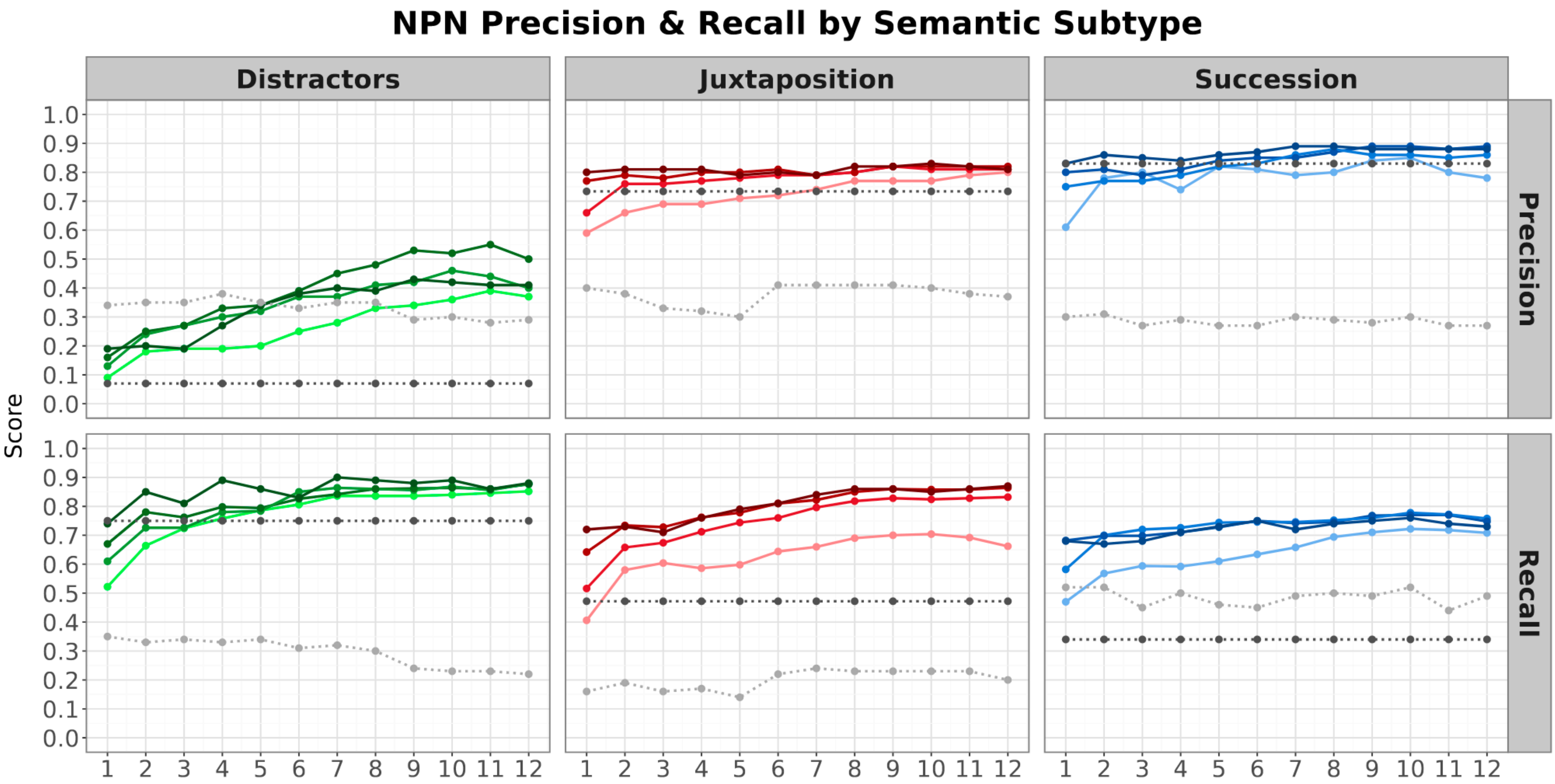
(3) I was living **moment to moment**. SUCCESSION

(4) You can preserve core warmth by huddling with a buddy, **chest to chest**. JUXTAPOSITION

✴ Largely: Especially in higher layers, and especially with ≥25 training examples, the contextualized representations exceed a static embedding baseline

Scivetti & Schneider (2025)

What do LMs “know” about NPN?



Scivetti & Schneider (2025)

Questions

1. How can we apply **general** syntactic categories to instances of **idiosyncratic** constructions?

   💡 Define general categories with prototypes for elasticity (+ consider subtyping)

2. How can we **annotate** instances of constructions in a **crosslinguistic** way?

   💡 Develop functional definitions of constructions as comparative concepts

   💡 Query syntactic treebanks to identify form pattern

   💡 Manually disambiguate senses

3. How well do LMs implicitly capture constructions' **form and meaning**?

   💡 BERT embeddings of *to* distinguish NPN from distractors, and at least partially capture word order and semantic sense

Take-home points

- Meaningful units in a language may be "packages" with lexical and/or grammatical constraints on form

- Such constructions may be frequent or rare

- Form <-> function mappings are often many-to-many

- Functional definitions can facilitate crosslinguistic comparison

- Annotation of constructions atop universal standards like UD facilitates empirical comparison

- Much to investigate about whether/how LMs "acquire" constructional forms and meanings

CxG/CL past and future

- Theme session on Computational Aspects of Frames and Constructions @ ICCG 2016 (Miriam R. L. Petruck, Nathan Schneider)

- LAW-MWE-CxG workshop @ COLING 2018 (Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, Miriam R. L. Petruck)

- CxGs+NLP @ GURT/SyntaxFest 2023 (Claire Bonial, Harish Tayyar Madabushi)

- **CxGs+NLP @ IWCS 2025** Dusseldorf: September 24 (Claire Bonial, Harish Tayyar Madabushi)
    - submission deadline June 6

- **UCxn** welcomes more languages and constructions! 😀

Natural Language Processing

Resources

Evaluation

Low-resource settings

Interpretability

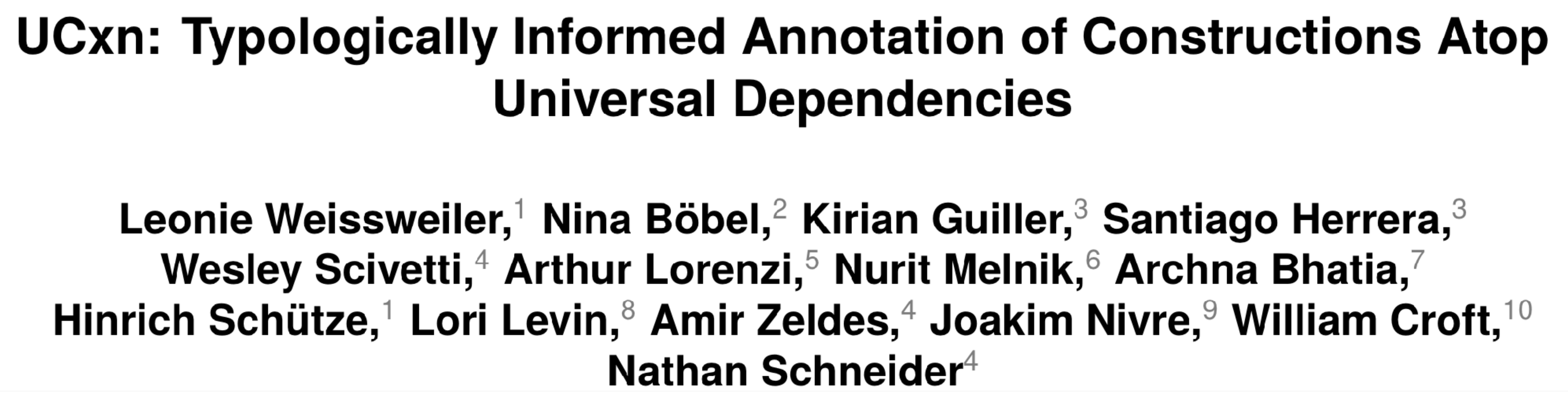
Explanation

Study of language

on Linguistics

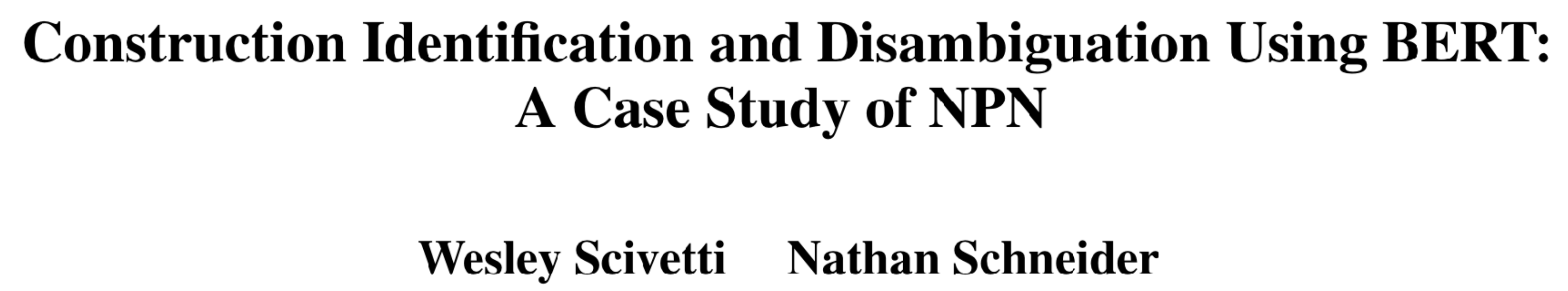
Juri Opitz*, Shira Wein*, Nathan Schneider* (2025)
Computational Linguistics

Thanks!



UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies

Leonie Weissweiler,[1] Nina Böbel,[2] Kirian Guiller,[3] Santiago Herrera,[3] Wesley Scivetti,[4] Arthur Lorenzi,[5] Nurit Melnik,[6] Archna Bhatia,[7] Hinrich Schütze,[1] Lori Levin,[8] Amir Zeldes,[4] Joakim Nivre,[9] William Croft,[10] Nathan Schneider[4]



Construction Identification and Disambiguation Using BERT: A Case Study of NPN

Wesley Scivetti   Nathan Schneider