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]

[1]LMU Munich & MCML, [2]HHU Düsseldorf, [10]University of New Mexico, [8]Carnegie Mellon University
[3]Université Paris Nanterre, CNRS, [4]Georgetown University, [5]Federal University of Juiz de Fora
[6]The Open University of Israel, [7]Institute for Human and Machine Cognition, [9]Uppsala Univ. and RISE
weissweiler@cis.lmu.de, nathan.schneider@georgetown.edu

Abstract

The Universal Dependencies (UD) project has created an invaluable collection of treebanks with contributions in over 140 languages. However, the UD annotations do not tell the full story. Grammatical constructions that convey meaning through a particular combination of several morphosyntactic elements—for example, interrogative sentences with special markers and/or word orders—are not labeled holistically. We argue for (i) augmenting UD annotations with a "UCxn" annotation layer for such meaning-bearing grammatical constructions, and (ii) approaching this in a typologically informed way so that morphosyntactic strategies can be compared across languages. As a case study, we consider five construction families in ten languages, identifying instances of each construction in UD treebanks through the use of morphosyntactic patterns. In addition to findings regarding these particular constructions, our study yields important insights on methodology for describing and identifying constructions in language-general and language-particular ways, and lays the foundation for future constructional enrichment of UD treebanks.

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

1. Introduction

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]

(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?' (ⲉ-ⲓ-ⲛⲁ-ϫⲉ-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 (§3). Our goal is to annotate them on a large scale across many languages in UD treebanks as automatically and accurately as possible. In this paper, we demonstrate 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. We present a case study of five construction families and ten languages to illustrate the challenges and opportunities of this approach.

Our 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 (like the object relation and the interrogative pronoun in (1)) 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*.

Continuing with the example of interrogative constructions highlights some of the challenges, even within English. (2) illustrates ambiguity with exclamatives, as well as noncanonical kinds of interroga-

---

[1]In many cases, prosody or punctuation can also indicate a clause is interrogative. Coptic texts, however, do not use question marks, and e.g., web data contains nonstandard punctuation use (Sanguinetti et al., 2022).

tives involving ellipsis, idioms, and echo questions.

(2) a. WOW what stunning views. [en-EWT]
        (Inferred interpretation: 'What stunning views!', not 'What stunning views?')
    b. Can you tell us where. [en-EWT]
    c. WELL GUESS WHAT!!! [en-EWT]
    d. She didn't have what? [en-GUM]

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, in our view, 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. On the more practical side, construction annotation could be used for downstream tasks like inducing frame-semantic representations, information extraction, or predicting grammatical difficulty for L2 learners depending on strategies found in the L1 language, or for heritage learners depending on strategies found in the dominant language (Bhatia and Montrul, 2020).

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). Our proposed framework, **UCxn**, is grounded in ideas from Construction Grammar and linguistic typology (§2). Our empirical methodology (§3) is to annotate treebanks in each of 10 languages—English, German, Swedish, French, Spanish, Portuguese, Hindi, Mandarin, Hebrew, and Coptic — for selected constructions by constructing graph pattern queries and matching them against UD trees. The constructions are interrogatives (§4), existentials (§5), conditionals (§6), resultatives (§7), and noun-adposition-noun combinations where the noun is repeated (NPN; §8). Highlights from our corpus investigations corresponding to each construction are discussed in each section, with a quantitative and qualitative discussion in §9.[2]
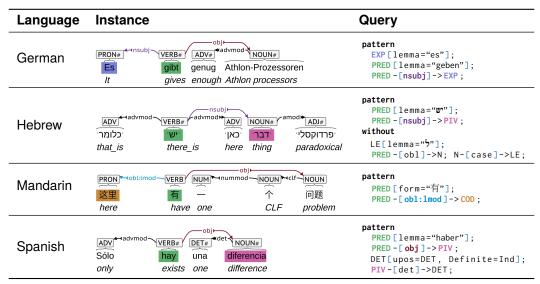
---

2. Background

Universal Dependencies (UD) is a framework for crosslinguistically consistent morphosyntactic annotation, which to date has been applied to over 140 languages (Nivre et al., 2016, 2020; de Marneffe et al., 2021). UD annotation consists of two layers: a morphological layer, where each word is assigned a lemma, a universal part-of-speech tag, and a set of morphological features; and a syntactic layer, where all words are connected into a dependency tree labeled with universal syntactic relations. The syntactic representations are defined to prioritize direct relations between content words, which are most likely to be parallel across different languages; function words are treated as grammatical markers on content words. While the inventories of part-of-speech tags and syntactic relations are fixed to support crosslinguistic comparison, the framework allows elaboration through the use of language-specific morphological features and subtypes of syntactic relations. CoNLL-U is the standard file format for UD treebanks; trees are encoded in 10 tab-separated columns. The last of these, MISC, is open-ended to support annotations beyond the UD standard itself.

Construction Grammar (CxG) is an approach to linguistic analysis in which the basic unit is a pairing of form and meaning and in which meaning-bearing units can have multiple parts (Construction Elements) (Fillmore et al., 2012; Croft, 2001; Fillmore et al., 1988; Goldberg, 1995, 2006; Hoffmann and Trousdale, 2013). A construction grammar is generally represented as a network indicating taxonomic, partonomic and other relations between constructions, called a Constructicon (Diessel, 2019; Fillmore et al., 2012; Lyngfelt et al., 2018).

Typology Much work in CxG is done in a single language, where, for example, a specific construction like the English Interrogative Construction is defined by a specific form in English and its functions. This is a *semasiological* approach: it starts from a form and examines its possible functions. However, form varies greatly across languages, and many features of morphosyntactic form are defined in language-specific terms such as specific morphemes (English **do**) or word classes (English Auxiliary). In linguistic typology, one must find a basis for crosslinguistic comparison of a construction, that is, a comparative concept (Haspelmath, 2010).

This basis is primarily the function of a construction. For example, a typology of interrogative constructions compares sentence forms across languages that express the function of a speech act requesting information from an interlocutor. This is an *onomasiological* approach: it starts from a function and examines its possible forms. In typol-

Language	Instance	Query

German	PRON# Es *It* — VERB# gibt *gives* — ADV# genug *enough* — NOUN# Athlon-Prozessoren *Athlon processors* (nsubj, obj, advmod)	`pattern`
		`EXP[lemma="es"];`
		`PRED[lemma="geben"];`
		`PRED-[nsubj]->EXP;`

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

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

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

Table 1: Existential/presentential 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.

ogy, a construction such as the interrogative construction is the set of form-function pairings across languages expressing a particular function.[3,4]

Morphemes and word order can also be described in a crosslinguistically valid fashion. For example, many languages use a special morpheme in interrogative constructions, as with Chinese 吗 *ma*. We can describe this as an "interrogative marker". Other languages change word order in comparison to the declarative construction, albeit in different ways. An interrogative marker or a word order change are two different *strategies* for expressing the interrogative function. We can then describe languages as using the same strategy, or different strategies, for the interrogative construction.

Another important concept is *morphosyntactic recruitment*. If two different constructions such as an existential construction and a possessive construction are morphosyntactically similar, we may say that one construction has *recruited* a strategy from the diachronically or conceptually prior construction, although the directionality or their etymological source may not always be clear.

Related Work Prior work on creating datasets annotated with constructions has been in the form of various Constructicon projects, repositories describing the constructions of a language. One of the first and best known is the Berkeley FrameNet Constructicon for English (Fillmore et al., 2012).

Some Constructicons incorporate UD annotations and corpora (for German, Brazilian Portuguese, and Russian; Ziem et al., 2019; Torrent et al., 2018; Bast et al., 2021). While those Constructicons may select individual attestations from corpora to exemplify a construction, in this paper we are concerned with labeling as many instances of the construction in the corpus as possible. Here we take a fundamentally crosslinguistic view of constructions, though the annotation layer could just as well include language-specific constructions. Ultimately we foresee a healthy feedback loop between Constructicon development and corpus enrichment of the kind pursued in this paper.

Construction Grammar has also recently gained popularity in NLP. There have been practical studies using CxG to probe the inner workings of large language models (Weissweiler et al., 2022; Mahowald, 2023), as well as general observations about the compatibility of usage-based constructionist theories with the recent successes of language models (Goldberg, To appear; Weissweiler et al., 2023). Earlier work (Dunn, 2017; Dunietz et al., 2017, 2018; Hwang and Palmer, 2015) in construction-based NLP focused on the annotation, automatic detection and induction of constructions.

3. Methodology

Selection of Constructions For the purpose of crosslinguistic comparison, we define constructions in terms of function (e.g., a speech act requesting information), rather than form (e.g., subject-auxiliary inversion). We take a modified onomasiological approach: start from a function, and identify the most conventional forms that express the function. In

---

[3]In order to distinguish language-specific constructions from constructions as comparative concepts, we follow typological practice and capitalize the names of language-specific constructions.

[4]Some work in CxG such as Hasegawa et al. (2010) is onomasiological and crosslinguistic, using frame semantics as the meaning.

many cases, a language conventionally uses more than one strategy for a construction's function. We annotated a few, but not all, of the more conventionalized strategies for each construction in each language. Our aim is to see if morphosyntactic queries can detect each strategy in each language, starting only with the information available in UD.

We chose our constructions to be as diverse as possible. We have selected a speech act construction (interrogative), an information structure construction (existential), a complex sentence construction (conditional), an argument structure construction (resultative), and a phrasal construction (NPN). These constructions cover a broad range of specificity, probably annotation complexity, and size. With the NPN construction (Jackendoff, 2008), we examine a strategy, to compare the functions it expresses in the languages in our sample whereas with the other constructions, we examine the functions to compare the strategies recruited in the languages in our sample.

Previous work has explored the relationship between UD annotation and the annotation of (semantically idiosyncratic) multiword expressions (Savary et al., 2023). Here, by contrast, we focus on constructions that are not fully lexically specified—but we share the goal of identifying structures with more to them semantically than meets the eye.

Selection of Languages We select 1 or 2 treebanks for each of a set of languages, ensuring diversity with respect to treebank size and language family. Each language is worked on by at least one linguist who is also a native or proficient-level speaker. Our languages and the treebanks we use can be seen in Table 5 in Appendix A. We use UD v2.13.

Although our sample of languages is not representative of global language diversity, covering several languages from several regions ensures that we will cover some variation in strategies.

Identifying Constructions Constructions are defined crosslinguistically in terms of their *function*, but UD annotates morphosyntactic *form*. For some languages and datasets, we do have functional annotations in addition to syntax trees: e.g., the UD English GUM corpus is also annotated with Rhetorical Structure Theory (RST, Mann and Thompson 1988), which identifies pragmatic functions for clauses (such as conditional ones), regardless of how they are expressed. Although we can use this type of information to help identify the scope of ways of expressing a certain meaning or class of meanings in a language, we assume that such annotations are either unavailable for most languages, or do not cover the full breadth of functions whose corresponding constructions we are interested in. 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, as shown in Table 1. 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 We propose a new annotation layer, "UCxn", to represent construction instances in UD treebanks. In our data release, UCxn information is incorporated directly into CoNLL-U files, which support arbitrary key-value annotations via the MISC field (10th column). We introduce 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	...	_

A technical specification[2] offers full details on the format and naming conventions in our data. It also offers the option of annotating *construction elements* in a CxnElt field. At present, we annotate only content elements (such as the protasis and the apodosis clauses for conditionals; §6), but not functional elements like subordinators that may be strategy-specific. Next, we proceed construction by construction, first describing a construction in general terms, then highlighting findings from querying treebanks.

4. 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.

		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.

Automatic Annotation Efforts In this section we compare information questions in which the interrogative phrase is placed either in the same position as its non-interrogative counterpart as in (3) or in a different, often fronted position as in (4).

(3) You went where?

(4) Where did you go?

To identify interrogatives, we relied on either the presence of WH items (**what**, **who**, etc.), word order (in languages using it for marking), as well as the presence of question marks or sentence type annotations where available. In some languages, WH items are identical to indefinite pronouns or free-relative heads (e.g., *I ate what you cooked* is not interrogative, despite containing **what**), but the UD morphological feature `PronType=Int` helps to disambiguate. We did not see the special verb form strategy in our treebanks.

Table 2 shows pre- and post-posed realization frequencies for different grammatical functions for WH pronouns in interrogatives (i.e., excluding uses such as 'I know who!'), compared to overall usage excluding such pronouns. The table shows the strong preference to front WH objects in English (28:3 in favor of pre-posed; for other objects the ratio is 265:8889). For other functions, the picture is more complex: interrogative adverbials such as 'when' and 'where' appear almost exclusively pre-posed, while non-interrogative phrases strongly prefer fronting, but only at a rate of 8258/2196 (79%).

Turning to Coptic for comparison, Table 2 shows a rather different picture. The tendency for placing subjects before their heads and objects after them is much weaker (5:2, but based on only 7 cases); for adverbial interrogatives, fronting occurs proportionally less than in non-interrogatives, though there is very little data. The frequent presence of the Coptic focalizing marker **ere**, which indicates a contrast

with a previously uttered or implied phrase, plays a role in promoting late realization of arguments, above and beyond the tendency for each grammatical function (Green and Reintges, 2001).

Takeaways Although typological literature often classifies languages in terms of basic word order or the possibility of word order changes, the actual picture in individual language data is much more complex. We have shown that quantitative analyses with construction-annotated data give a more nuanced picture of how languages realize such word order dependencies in interrogatives.

5. Existentials

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.

Languages vary with respect to the predicate that they use in the existential. One class of languages employ a construction-specific lexeme, such as Swedish **finnas**. In Coptic there are lexicalized negative and positive existence predicates, ⲟⲩⲛ **oun** and ⲙⲙⲛ **mmn**. Historically, predicative possession used the same items, but through lexicalization, the possessive versions are now lexically distinct from the existentials.

The relationship between existence and possession also has synchronic manifestations. Our sample includes languages that use a possession verb as the predicate in an existential, one predicate to express both existence and possession, such as **ter** 'to have' in Brazilian Portuguese, French **avoir** in the phrase *il y a*, or the Mandarin predicate 有 **yǒu**. This duality is also found in Hebrew, where possession is expressed by adding a possessive dative argument to the existential construction (5).

(5) hayu (la-nu) kama taxanot ba-derex
were.3P (to-us) few stops.PF in.the-way
'There were/We had a few stops on the way.'
(היו (לנו) כמה תחנות בדרך) [he-HTB]

An additional existential strategy shares a copula with the predicational locative construction. In Hebrew, the copula היה **haya** is used in past and future tense existentials (*hayu* in (5) is the inflected form). For Mandarin, the use of the copula 是 **shì** is an alternative to the lexicalized existential predicate 有 **yǒu**. The link between locative and existential is also found in locatives that grammaticalized into

unique existential forms such as English **There('s)** or French **y** (in *il y a*).

Finally, the existential predicates **haber** and **haver** in Spanish and Portuguese, respectively, also function as auxiliaries and modals, similarly to the English **have**, modulo possession.

The argument structure of existential predicates is not uniform crosslinguistically, with pivots exhibiting different degrees of subjecthood properties (Keenan, 1976). This diversity is manifested in the UD annotation. In one class of languages, no argument is identified as nsubj and the pivot is attached as obj in UD. This is the case in Spanish and Mandarin (see Table 1).

Other languages identify the pivot as nsubj. This is the case in Hebrew, where the copula standardly exhibits agreement with the pivot, as in (5). However, unlike typical subjects, the Hebrew pivot appears post-verbally, does not always trigger agreement, and in informal speech may receive accusative marking, if definite. Likewise, in Coptic the pivot is nsubj in postverbal position, though the adverb **there** is added in around 5% of cases in the UD data with no clear antecedent.

A different strategy involves employing an expletive as a co-argument to the pivot. This is found in our language sample in French and English (6) and in German (Table 1).

(6) il y   a   une salle à l'étage
     it there has a   room upstairs
     'There is a room upstairs.'   [fr-GSD]

Here, too, UD annotations vary across languages. In the English treebank the pivot is attached as nsubj and **there** as expl. In French, the expletive **y** is expl:comp and the pivot is obj. In German, the expletive **es** is nsubj and the pivot is obj.

Automatic Annotation Efforts   Our languages vary in the difficulty of identifying existential constructions. The easiest cases were those in which a construction-specific lexical item is employed (e.g., the lexicalized existential predicates in Coptic and Swedish). In French, instances of the existentials are identified by queries which target the construction-specific cooccurrence of the clitic **y** and the verb **avoir** in a comp:expl relation.

The more challenging cases are those in which the elements which encode existence are multifunctional. In some treebanks, this challenge is overcome by construction-specific annotations. Thus, for example, in the Hebrew HTB the predicate היה **haya** is annotated as HebExistential=Yes where it is used in its existential function.

When disambiguating annotations are not available, the queries rely on other distributional properties of the construction to avoid false positives. In French, the queries only target indefinite pivots, excluding definite determiners and numerals.

In Hebrew, the queries exclude instances where the predicate has a obl dependent with a dative case marker, i.e., a possessor (see query in Table 1). Furthermore, to distinguish between the predicational and existential functions of היה **haya** in UD_Hebrew-IAHLTwiki (Zeldes et al., 2022), where this information is not annotated, the query targets only post-verbal nsubj dependents.

Takeaways   Lexical items that are associated with the existential construction are often shared with other constructions. For this reason, in order to maximize accuracy the queries cannot only rely on these lexical items but also target morphosyntactic properties and dependency relations.

6.   Conditionals

Typological Overview   A conditional construction is a complex sentence construction describing a broadly "causal" link between the two states of affairs, the protasis (condition) and the apodosis (consequence) (Comrie, 1986, pp. 81–82). The strategies for conditional constructions are largely the typical ones for complex sentences in general (Croft, 2022, pp. 532–34). The construction may be an adverbial subordinate construction or a coordinate construction. The clauses may be balanced (identical in form to a declarative main clause) or deranked (one clause, usually the protasis, is in a distinct form, with a special verb form and other differences). There may be a subordinating conjunction such as **if**, or rarely a change in word order, as in English *Had he stayed, he would have seen it.* The nonfactual nature of conditionals may manifest in irrealis or subjunctive verb forms.

Automatic Annotation Efforts   Common strategies for conditionals are the use of a subordinating conjunction as in German *Wenn die Möglichkeit da ist* (lit. 'if the opportunity there is') (3291 instances in German-HDT, 240 in Swedish-Talbanken, 243 in Hindi-HDT, 495 in English-GUM), or word order inversion as in Swedish *Har du god kondition* (lit. have you good condition; if you are in good shape) (1182 instances in German-HDT, 68 in Swedish-Talbanken, 7 in English-GUM). A very different strategy involves conditional circumfixes. In Coptic **e- -šan** is a circumfix that conveys conditionality (CD) and applies to the pronominal subject of the conditional clause so that *e-f-šan-eibe* (CD-he-CD-thirst) means *If he is thirsty.*

Our investigation of conditionals has shown that it may not be possible, using the information available in UD, to create queries that accurately retrieve conditional sentences. There are three sources of difficulty: (1) the need for information that is not yet encoded in UD, (2) subordinating conjunctions and clause types that are not exclusively used in

conditional constructions, and (3) the variety of subordinating conjunctions and other strategies that are used to express the conditional construction.

Conditional subordinating conjunctions can be divided into: simple subordinating conjunctions like **agar**/**yadi** (Hindi) ('if'); complex subordinating conjunctions like **förutsatt att** (Swedish) ('provided that'); and V2 sentence embedders like **angenommen** (German) ('presumed') ([Breindl et al.], [2014]). Complex subordinating conjunctions and V2 sentence embedders are problematic in German HDT because the part of speech is not *conjunction* and the dependency label is not *mark* (or *fixed expression* as in Swedish). The query needs to specify the connector lemma, giving many false positives.

In Germanic languages, conditional constructions without subordinating conjunctions usually express the protasis as verb-initial clauses that precede the main clause. In Swedish and German, any verb can be used in a verb-initial protasis clause. In English, however, it is restricted to certain auxiliaries (e.g. **Had** *I gone, I would have seen you*).

While the subtypes of English conditionals (e.g., neutral or negative epistemic stance) require many search queries but are in principle findable with UD, this is not the case in German. A major problem for German conditionals—especially with regard to semantic and syntactic subcategorization—was posed by the inadequate mood annotations. German HDT does not annotate conditional or potential verb forms and marks most verb forms as indicative, even when there is a clear conditional or subjunctive structure. It is therefore not possible to search for semantic subcategories based on different mood annotations in HDT, although verb mood is the most common indication of grouping conditionals in German ([Schierholz and Uzonyi], [2022]).

Takeaways   The conditional strategies are in principle searchable, although writing these rules requires an exhaustive study of the phenomenon in each language. Search requirements vary in complexity depending on the depth of the underlying linguistic analyses of the phenomenon. Annotation practices may complicate the search process and even make some distinctions impossible.

7. Resultatives

Typological Overview   From a functional perspective the resultative construction expresses an event with two subevents: a *dynamic* subevent such as **paint** and a *resulting state* subevent such as **red** in (7).

(7) They painted the door **red**.

The English resultative construction is a prime example of an argument structure construction ([Hovav and Levin], [2001]; [Goldberg and Jackendoff],

[2004]). A basic transitive clause describing an event is augmented with a secondary predicate describing the result state of a participant, but there are many strategies to express this function. English, for example, also uses adverbial subordination of the dynamic event: *The door was red as a result of their painting it* or *I flattened the metal by hammering*. In our study of the resultative construction, we are only annotating cases where the language provides a conventionalized strategy for expressing the resultative event as a complex predicate composed of a dynamic action and a result state.

Automatic Annotation Efforts   In the sample languages, we encountered several challenges. First, in some languages a resultative conceptualization is lacking: they do not combine a dynamic event with a stative result event into a complex predicate. In Hebrew, the most natural way of expressing the painting event literally translates as 'They painted the door in red' (the result expressed with an oblique marked by the prepositional prefix **be-** 'in'). In Hindi, complex predicates expressing a cause-result relation have a dynamic event as a result (8). We consider these languages as lacking the resultative construction as defined above.

(8) ... ki   veh duSman ko **maar**
     ... that it   enemy   ACC **hit**
     bhagaa-ye   [hi-HUTB]
     run.CAUS-SUBJ
     '... that it beat and chase away the enemy.'
     (कि वह दुश्मन को मार भगाए ।)

Second, in several languages many of the complex events with a dynamic subevent and resulting state subevent were of the form ['make/do' X STATE], where the dynamic event is the causative verb 'make/do', e.g., Hindi **kar** 'do', Swedish **göra** 'make/do', or German **machen** as in *Nvidia machts möglich* 'Nvidia makes it possible'. This construction is generally analyzed as the causative of a stative event, and is excluded from the resultative category. In the German and Swedish treebanks, removal of the causative left few or no examples of genuine resultatives.

Third, in some languages, the UD annotation of the resultative construction is indistinguishable from another construction such as depictive secondary predication. For example, *I hammered the metal flat* (resultative) has the same structure as *I left the door open* (depictive). It would be necessary in English for queries to incorporate lexical lists of predicates licensing the construction, in order to disambiguate from other sentences with similar UD structures, at the expense of generalizability to predicates that have not been seen in the resultative construction.

Finally, Chinese has a very productive resultative construction (9), which is already annotated in the

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

Table 3: Semantic categories of NPN and their crosslinguistic attestation in UD treebanks. – means that the target meaning is not possible in the language. (+) signals that the meaning is possible but not attested in the UD treebanks. (?) means that the existence of this meaning is unclear, see footnote 6. Succession: SU, Comparison: CO, Opposition: OP, Proximity: PR, Quantification: QU

treebank Chinese-HK (Wong et al., 2017) with a label specifically designed for resultative complements: `compound:vv`. They are trivially extracted by querying for that dependency relation.

(9) wǒ **qiāo píng** le   dīngzi   [zh-HK]
     1SG **hit flat** PERF nail
     'I hammered the nail flat.' (我敲平了钉子.)

Takeaways   In summary, the attempt to annotate the resultative construction has shown us several difficulties: annotating a construction where boundaries are in dispute within the literature, which might not even exist in all languages depending on the definition, and where considerable linguistic expertise and manual effort is required to write a comprehensive set of rules, indicating the need for collaboration among theoretical linguists, corpus linguists, typologists and computational linguists. Efforts such as ours can reveal constructions that need further linguistic investigation, or can help solidify linguistic consensus on the definition of the construction.

8. NPN

Typological Overview   With the preceding four constructions, we took an onomasiological approach, examining them crosslinguistically on a functional basis. Most work in CxG, however, takes a semasiological (form-first) approach to characterizing a formal pattern and its function(s), usually within a single language. In our terms, this approach starts with a strategy and examines the range of functions using that strategy. The UD framework offers a common vocabulary for describing formal categories of morphology, parts of speech, and grammatical relations across languages. In this section, we consider how a se-

masiological or strategy-based inquiry can be conducted crosslinguistically using UD corpora. As a case study, we look at the "NPN" strategy, in which a meaning related to quantification or iteration is expressed with a repeated noun and an adposition or case marker on the second noun. Examples in English include *day after day, shoulder to shoulder, box upon box,* etc. While infrequent and often a source of idioms, this strategy recurs across many languages (Postma, 1995; Matsuyama, 2004; Jackendoff, 2008; König and Moyse-Faurie, 2009; Roch et al., 2010; Pskit, 2015, 2017; Kinn, 2022).[5]

Automatic Annotation Efforts   We find examples of NPN strategies across 8/10 languages.[6] In our queries, we limit ourselves to instances where the two Ns are the same lemma, though there are related NPN uses where the two Ns are not the same (Jackendoff, 2008). A few examples of NPN from our treebanks are presented in (10).

(10) PT: ***frente a frente*** 'face to face' (lit. 'front to front'), FR: ***jour pour jour*** 'to the day' (lit. 'day for day'), SV: ***steg för steg*** 'step by step' (lit. 'step for step'), HE: ***mila be-mila*** 'word for word' (lit. 'word in word')

In terms of morphosyntactic form, NPN strategies are well captured by our queries because of the strict precedence relationship between the constituent elements. We find that there is considerable variability in whether NPNs are analyzed as fixed expressions in UD (using the `fixed` relation type), or whether the second N is analyzed as an `nmod` of the first N.[7]

The semantics of NPN have been well investigated in previous literature (Jackendoff, 2008; Roch et al., 2010; Sommerer and Baumann, 2021; Kinn, 2022). We find that most of the previously proposed semantic subcategories emerge in our languages. Following the categorization and discussion in Jackendoff (2008) and later works, we find the following semantic subtypes of NPN: SUCCESSION (*hour after hour*), COMPARISON (*man for man*), OPPOSITION (*brother against brother*), PROXIMITY (*hand in hand*) and QUANTIFICATION (particularly of a large quantity, *snacks upon snacks*). Qualitatively, we noticed

---

[5] The studies cover Dutch, English, French, German, Norwegian, Japanese, Mandarin, Polish, and Spanish.

[6] We did not find any attestations of NPN in the Chinese or Hindi treebanks. It is unclear whether NPN is productive in these languages, but we are aware of expressions that might qualify: e.g., Mandarin **yī tiān bǐ yī tiān** and Hindi **din ba din** (both 'day by day').

[7] We restrict our queries to exclude cases where the first N is marked by another adposition, because we find that in many languages the PNPN strategy (*from time to time*) has a different range of meanings than the NPN strategy. We also exclude cases where nouns are modified with adjectives, as these are extremely rare.

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).

that the SUCCESSION submeaning was most prevalent, and OPPOSITION is typically restricted to body parts, as in (10). Table 3 summarizes our empirical findings by semantic subtype and language.

Takeaways Using a strategy, like NPN, as the basis of typological comparison is not without issue (Croft, 2022); however, we do find considerable functional overlap in terms of the meanings which are conveyed by the NPN strategy in our language sample. Notably, NPN is the only investigated construction/strategy for which the query is almost universal across languages, meaning that it is the most well-integrated with UD: if the promise is universality across languages, then ideally a query would also work across all languages. It makes perfect sense that this only works with strategies, which are defined by their form, and not for constructions, which are defined by their meaning, as UD itself focuses on form.

9. Survey Summary

Our 5 case studies have surveyed constructions and strategies in 10 languages. Table 4 provides a quantitative summary in terms of matched instances per treebank. Treebanks ranged in size from 9k to 3.5m tokens; in some cases, the scale was too small for a robust set of results. NPN was particularly sparse—this is simply a rare strategy.

Table 4 also provides a qualitative summary of some of the major kinds of issues encountered: definitional issues (D), annotation errors in the treebank (A), unavoidable occasional false positives (f), many false positives due to overlap with another construction (F), and unattested strategies for which at least one query returned 0 examples (H). For most of the languages, we abandoned attempts to quantify resultatives given the definitional challenges. Note that some of the larger treebanks had unattested strategies (H)—this is not necessarily because of a problem with the treebanks, but

reflects that more effort was put into writing queries for long-tail strategies in those languages.

We are pleased to see that UD annotation errors (A) were not a major source of difficulty for most of the treebanks examined. On the other hand, many constructions were fundamentally difficult to circumscribe (D) or distinguish from other constructions given the available UD annotations (F). These may necessitate human annotation and/or supplementary information from semantic analyzers.

For English and Hebrew, where we consulted two treebanks, we can see some differences in the construction counts that are not explained by the size of the treebank but rather by the domain. This underscores the importance of domain diversity in empirical studies of constructions.

10. Conclusion and Future Work

We have 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 our 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). Beyond the created resources, these efforts may prompt improvements to the UD annotation guidelines and to language-specific Constructicons. Crucially, this work has been a first attempt at bringing two important frameworks together. We aim to gather feedback and input from the community to further our goal of integrating constructions fully with UD.

Acknowledgments

Bibliographical References

Radovan Bast, Anna Endresen, Laura A. Janda, Marianne Lund, Olga Lyashevskaya, James McDonald, Daria Mordashova, Tore Nesset, Ekaterina Rakhilina, Francis M. Tyers, and Valentina Zhukova. 2021. The Russian Constructicon. An electronic database of the Russian grammatical constructions. Available at https://constructicon.github.io/russian/.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. 2017. The Hindi/Urdu treebank project. In *Handbook of Linguistic Annotation*, pages 659–697. Springer Press.

Archna Bhatia and Silvina Montrul. 2020. Comprehension of differential object marking by Hindi heritage speakers. In A. Mardale and S. Montrul, editors, *The Acquisition of Differential Object Marking*, pages 261–281. John Benjamins Publishing Company.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Eva Breindl, Anna Volodina, and Ulrich Hermann Waßner. 2014. *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*, volume Band 13 of *Schriften des Instituts für Deutsche Sprache*. De Gruyter, Berlin and München and Boston.

Bernard Comrie. 1986. Conditionals: a typology. In E. C. Traugott, A. ter Meulen, J. S. Reilly, and C. A. Ferguson, editors, *On Conditionals*, pages 77–99. Cambridge University Press.

William Croft. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford, UK.

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge University Press, Cambridge.

Jesse Dunietz, Jaime Carbonell, and Lori Levin. 2018. DeepCx: A transition-based approach for shallow semantic parsing with complex constructional triggers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1701, Brussels, Belgium. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292.

Jan Einarsson. 1976. *Talbankens skriftspråkskonkordans*. Institutionen för nordiska språk, Lunds universitet.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of 'let alone'. *Language*, 64(3):501–538.

Charles J. Fillmore, Russell R. Lee-Goldman, and Russell Rhodes. 2012. The FrameNet Constructicon. In Hans C. Boas and Ivan A. Sag, editors, *Sign-Based Construction Grammar*, pages 283–322. CSLI Publications, Stanford, CA.

Adele E. Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford.

Adele E. Goldberg. To appear. A chat about constructionist approaches and LLMs. *Constructions and Frames*.

Adele E Goldberg and Ray Jackendoff. 2004. The english resultative as a family of constructions. *Language*, 80(3):532–568.

Melanie Green and Chris H. Reintges. 2001. Syntactic anchoring in Hausa and Coptic wh-constructions. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 27(2).

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.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [conversion and improvement of Universal Dependencies French corpora]. *Traitement Automatique des Langues*, 60(2):71–95.

Yoko Hasegawa, Russell Lee-Goldman, Kyoko Hirose Ohara, Seiko Fujii, and Charles J. Fillmore. 2010. On expressing measurement and comparison in English and Japanese. In Hans C. Boas, editor, *Contrastive Studies in Construction Grammar*, pages 169–200. John Benjamins, Amsterdam.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford handbook of construction grammar*. Oxford University Press.

Malka Rappaport Hovav and Beth Levin. 2001. An event structure account of english resultatives. *Language*, 77(4):766–797.

Jena D. Hwang and Martha Palmer. 2015. Identification of caused motion construction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 51–60, Denver, Colorado. Association for Computational Linguistics.

Ray Jackendoff. 2008. "Construction after Construction" and Its Theoretical Challenges. *Language*, 84(1):8–28.

Edward Keenan. 1976. Towards a universal definition of subject. In Charles N. Li, editor, *Subject and Topic*, pages 303–334. Academic Press New York, New York.

Torodd Kinn. 2022. Regular and compositional aspects of NPN constructions. *Journal of Linguistics*, 58(1):1–35.

Ekkehard König and Claire Moyse-Faurie. 2009. Spatial reciprocity: between grammar and lexis. In *Form and Function in Language Research: Papers in Honour of Christian Lehmann*, page 57–68. De Gruyter Mouton.

Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent. 2018. *Constructicography: Constructicon development across languages*, volume 22. John Benjamins Publishing Company.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Tetsuya Matsuyama. 2004. The N After N Construction. *English Linguistics*, 21(1):55–84.

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 Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings*

of the 10th International Conference on Language Resources and Evaluation (LREC), pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 4034–4043.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Gertjan Postma. 1995. Zero Semantics — The syntactic encoding of quantificational meaning. *Linguistics in the Netherlands*, 12:175–190.

Wiktor Pskit. 2015. The Categorial Status and Internal Structure of NPN Forms in English. In *Within Language, Beyond Theories (Volume I): Studies in Theoretical Linguistics*, page 27–42. Cambridge Scholars Publishing.

Wiktor Pskit. 2017. Linguistic and philosophical approaches to NPN structures. In *Topics in Syntax and Semantics. Linguistic and Philosophical Perspectives*, page 93–110. Wydawnictwo Uniwersytetu.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Claudia Roch, Katja Keßelmeier, and Antje Muller. 2010. Productivity of NPN sequences in German, English, French, and Spanish. In *Proceedings of the Conference on Natural Language Processing 2010*, page 158–163, Saarbrücken, Germany.

Shoval Sade, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes.

2022. Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 57:493–544.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME Meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Stefan J. Schierholz and Pál Uzonyi, editors. 2022. *Grammatik: Band 2: Syntax*, volume Bd. 1.2 of *Wörterbücher zur Sprach- und Kommunikationswissenschaft*. De Gruyter, Berlin and Boston.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lotte Sommerer and Andreas Baumann. 2021. Of absent mothers, strong sisters and peculiar daughters: The constructional network of english NPN constructions. *Cognitive Linguistics*, 32(1):97–131.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tiago Timponi Torrent, Ely Edison Matos, Ludmila Meireles Lage, Adrieli Laviola, Tatiane da Silva Tavares, Vânia Gomes de Almeida, and Natália Sathler Sigiliano. 2018. Towards continuity between the lexicon and the constructicon in FrameNet Brasil. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors, *Constructicography: Constructicon development across languages*, pages 107–140. John Benjamins, Amsterdam.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa, Italy. Linköping University Electronic Press.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proc. of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 192–201, Brussels, Belgium.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4331–4344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Ziem, Johanna Flick, and Phillip Sandkühler. 2019. The german constructicon project: Framework, methodology, resources. *Lexicographica*, 35(2019):61–86.

A. List of Treebanks

An overview of the treebanks used in this work along with their total number of sentences is provided in Table 5. The genres covered by each treebank are shown in Table 6.

Lang	Treebanks	Num. Sents
EN	EWT, GUM (Silveira et al., 2014; Zeldes, 2017)	16,662; 10,761
DE	HDT (Borges Völker et al., 2019)	189,928
SV	Talbanken (Einarsson, 1976; Nivre et al., 2006)	6,026
FR	GSD (Guillaume et al., 2019)	16,342
ES	AnCora (Taulé et al., 2008)	17,662
PT	Bosque (Rademaker et al., 2017)	9,357
HI	HUTB (Bhat et al., 2017)	15,649
ZH	Chinese-HK (Wong et al., 2017)	1,004
HE	HTB, IAHLTwiki (Sade et al., 2018; Zeldes et al., 2022)	6,143; 5,039
COP	Coptic Scriptorium (Zeldes and Abrams, 2018)	2,203

Table 5: UD treebanks used in our crosslinguistic study. Some cover specific varieties, e.g., AnCora represents European Spanish, whereas Bosque covers both European and Brazilian Portuguese. Chinese is limited to Mandarin. Coptic (Sahidic) is the only historical language.

	EN	DE	SV	FR	ES	PT	HI	ZH	HE	COP
academic	+									
bible										+
blog	+			+						
e-mail	+									
fiction	+									+
government	+									
grammar examples										
learner essays										
legal										
medical										
news	+	+	+	+	+	+	+		+	
nonfiction	+	+	+							+
poetry										
reviews	+			+						
social	+									
spoken	+							+		
web	+	+								
wiki	+			+					+	

Table 6: Genres covered by the UD treebanks used in the paper.