Reference and Modification in Universal Dependencies

Joakim Nivre
Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

William Croft
University of New Mexico
Department of Linguistics
wacroft@icloud.com

Abstract

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology? To address this question, we need to systematically review how UD represents linguistic constructions in the world's languages, and how it handles the range of morphosyntactic variation attested in linguistic typology. In this paper, we start this review by discussing reference and modification constructions. The review shows that, although UD can represent all major constructions in this area, there are a number of cases where UD categories do not align systematically with a typological classification of constructions, and where constructional similarity is therefore not transparent across languages. We also identify limitations in the representation of certain morphosyntactic strategies, notably indexation and linkers. To overcome these limitations, we propose a number of revisions that may be considered for future versions of UD.

1 Introduction

Universal Dependencies (UD) is a framework for morphosyntactic annotation, which is designed to be applicable to all human languages in a way that enables meaningful cross-linguistic comparisons (Nivre et al., 2016, 2020; de Marneffe et al., 2021). To find out whether UD meets these requirements, Nivre (2025) proposes to build a constructicon for UD based on the survey of universal constructions and morphosyntactic realization strategies in Croft (2022) and the MoCCA database of comparative concepts derived from it (Lorenzi et al., 2024). In this framework, *constructions* are form-function pairings defined solely in terms of their function (hence universal), while *strategies* are defined by the pairing of a function with some cross-linguistically identifiable morphosyntactic form.

If we can provide a UD analysis for every combination of a construction and a strategy, then we can assess to what extent UD systematically captures cross-linguistic similarities and differences. In the same process, we can also gather evidence of gaps or inconsistencies in the current UD guidelines, and propose improvements for future versions.[1] In this paper, we present a first contribution to this project by discussing one of the most central constructions in the world's languages, that of nominal phrases, or referring expressions, including modification constructions that provide additional information about referents via their properties, their quantities, or their relations to other objects. Reference and modification constructions are discussed in Chapters 3–5 of Croft (2022).

2 Reference Constructions

Referring phrases are used to pick out and identify a referent, and they can be classified semantically into three broad categories, illustrated with a Russian example in (1) (Croft, 2022, p. 66):

(1)

ja	dal	knig-u	Ver-e
I	gave	book-ACC	Vera-DAT

'I gave Vera the book'

While *ja* identifies its referent *contextually* as the individual fulfilling the speaker role, *knigu* identifies the referent as belonging to a certain *type* (the book type), and *Vere* refers to a unique *individual*. Given these three semantic categories, we can define three basic reference constructions: *pronouns*, with contextual reference; *nouns*, with type reference; and *proper nouns*, with individual reference.

The basic reference constructions can be further subdivided both with respect to semantic content and information packaging. On the semantic side, nouns are often subdivided according to (degrees of) animacy, and pronoun systems often reflect ontological categories like person, thing, place, time

[1] An early review of UD from a typological perspective can be found in Croft et al. (2017).

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 1–10
August 27, 2025 ©2025 Association for Computational Linguistics

Construction	Strategy	UD Annotation
Pronoun	Zero	–
	Indexation	Predicate[Features]
	Word	PRON/DET
Noun	Word	NOUN
Noun + Determiner	Affix	NOUN[Features]
	Word	NOUN $\xrightarrow{\text{det}}$ DET
Proper Noun	Word	PROPN

Table 1: UD annotation of reference constructions and strategies. Features = indexation features.

and manner (Haspelmath, 1997). On the information packaging side, the most important notion is information status, which concerns the identifiability and accessibility of a referent. Information status is really a continuum, ranging from referents already mentioned in the discourse to purely hypothetical ones, but a broad distinction can be made between definite and indefinite referring expressions, where the defining criterion of the former is that the referent is already known to both speaker and hearer (Croft, 2022, p. 72–99).

The basic constructions are commonly realized on their own, as in (1), but especially (common) nouns are often combined with other expressions to form complex referring phrases. This often involves modification constructions, which will be discussed in Section 3, but it is also common to combine basic reference constructions. For example, contextual reference and type reference are often combined, as in the phrase *this book*, where the demonstrative *this* is used to constrain the type reference of the noun *book*. This combined use of the demonstrative – as opposed to its use as a pronoun on its own – is a subtype of the *determiner* construction, which also includes articles. Determiners are generally used to indicate the information status of the referent. However, while articles only express information status, demonstratives in addition encodes location with respect to the speaker and hearer (Croft, 2022, p. 73–74). Besides determiner constructions, it is common to combine several proper nouns, as in *Susan Smith*, and to combine common and proper nouns, as in *Aunt Susan*. More complex nominal phrases can be formed via modification, as discussed below in Section 3.

When it comes to morphosyntactic strategies, pronouns, nouns and proper nouns are most commonly realized as single words, but pronouns can also appear in reduced forms in many languages. They may be realized as clitics, that is, as morphemes that have the syntactic characteristics of a word but depend phonologically on another word or phrase; they may appear as affixes on the predicate, a strategy known as *indexation*; or they may not be phonetically realized at all, a *zero* strategy. Determiners can be realized as affixes on the noun or as separate words (sometimes with indexation of noun features such as number and gender). In general, strategies for referring expressions can be ranked on an *accessibility scale* (Givón, 1983; Ariel, 1988, 1990), where shorter expressions are preferred for higher accessibility referents.

How are the basic reference constructions represented in UD? Table 1 gives an overview of the constructions and strategies described in this section and their annotation in UD. In the following subsections, we discuss each case in detail and also make some observations about issues with the current guidelines and annotation practice in UD.

2.1 Pronouns in UD

The first thing to note here is that zero pronouns are not represented at all in UD. This is a consequence of the data-driven approach of UD (and most corpus annotation efforts), where the goal is to assign an interpretation to overtly observable forms, rather than to account for the realization of a certain content. This principle, which has been summarized in the slogan “Don’t annotate things that are not there!” (Nivre, 2015), does have the drawback that core arguments of a predicate are sometimes not represented at all, but changing it would be a major reorientation of the UD approach to morphosyntactic annotation.

Moving on to pronouns that are realized by indexation on a predicate, also known as agreement, this is captured in the UD annotation through morphological features on the predicate. However,

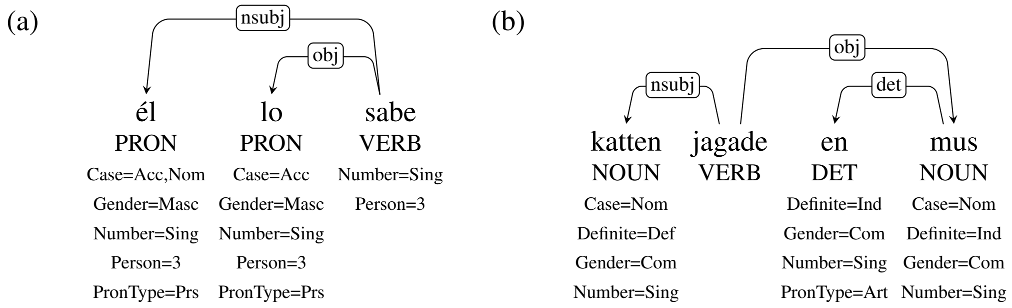

Figure 1: Simplified UD annotation of reference constructions: (a) pronouns, (b) nouns (with determiners).

when the indexation occurs together with an overt pronoun, there is nothing in the representation that links one to the other; and when there is no overt pronoun, there is no explicit information about which argument is being indexed. Thus, in the Spanish example in Figure 1(a),[2] there is nothing to indicate that the feature bundle [Number=Sing, Person=3] on the verb *sabe* (knows) corresponds to the specification of the subject pronoun *él* (he) rather than the object pronoun *lo* (it) (which happens to have the same values for these features). So if the subject pronoun is omitted, which is normal in Spanish if the subject is accessible in the context and not emphasized, it may accidentally look like Spanish has object-verb agreement. A possible improvement in future versions of UD could be to represent not only the features but also the target of the indexation, which would make the annotation more informative regardless of whether there is an overt realization of the target or not.

When a pronoun is realized as a word, finally, it is assigned the part-of-speech tag PRON and further subclassified using features, as illustrated in the Spanish example in Figure 1(a).[3] The feature PronType is used to distinguish major types of pronouns (and determiners) such as demonstrative, personal, interrogative, indefinite, and so on, while features such as Case, Gender, Number, and Person capture contrasting features within a given pronoun paradigm.

A special case of pronouns realized as words are clitics, which have the syntactic characteristics of a word but depend phonologically on another word or phrase, and which may therefore not be clearly recognizable as separate words in the standard orthography of a language. A typical case is Spanish *véase*, which consists of the imperative verb form *vea* (see) and the reflexive pronoun *se*. Cases like these are accommodated in UD through the mechanism of *multiword tokens*, which allows a single orthographic token (*véase*) to be treated as two separate syntactic words in the morphosyntactic annotation. However, multiword tokens are always optional, which means that there may be languages in UD that do not recognize all clitic pronouns as independent words. In such cases, the clitic will essentially be treated as an inflectional morpheme and analyzed by means of features (as the case of indexation discussed earlier).

2.2 Nouns and Determiners in UD

Nouns are almost always realized as independent words. In UD they are tagged with the universal part-of-speech tag NOUN and subclassified using features, as illustrated in the Swedish example in Figure 1(b). Common features for nouns include intrinsic features such as Gender, NounClass and Animacy, as well as inflectional features like Case, Number and Definite.[4]

The Swedish example also illustrates two different strategies for realizing determiners. In the subject noun phrase *katt-en* (cat-DEF), the definite article is realized as an inflectional suffix, captured in the annotation by the feature Definite=Def on the head noun. In the object noun phrase, the indefinite article is realized as an independent word *en*, which is linked to the head noun with the syntactic relation *det* (for determiner). The standalone article is tagged DET and has its own set of fea-

[2] In this and all following examples, we simplify the UD representations by omitting (a) lemmas and (b) morphological features that are not relevant for discussion (such as Tense and Mood features on the verb in this example).

[3] Some treebanks keep the tag DET also in pronominal uses for demonstratives regularly used as determiners, such as *all* and *some* in English, which is currently a source of cross-linguitic inconsistency in UD.

[4] The occurrence of nominative case features (Case=Nom) on both the subject and object is due to the Swedish case system having been reduced to just two cases: nominative and genitive.

tures, including PronType=Art and Definite=Ind, which together distinguish indefinite articles from other types of determiners.

The tag DET and the relation *det* are used not only for articles but also for other determiners, including demonstratives (*this book*), quantifiers (*all books*), and sometimes possessives (*my book*), which can be problematic from a typological perspective. We will return to this issue in Section 3.

2.3 Proper Nouns in UD

UD has a special part-of-speech tag PROPN for words that are primarily used as proper nouns, such as *Mary*, *Smith*, *London*, and *Sweden*. When several proper nouns are combined into a referring phrase, such as *Mary Smith*, they are tagged PROPN and combined with the syntactic relation *flat* (unless one of them is clearly distinguishable as the syntactic head).

It should be noted that not all phrases that are used with individual reference are analyzed in this way in UD. Phrases like *the North Sea* and *Gone with the Wind*, which are compositional phrases where the head word may not even be nominal, are annotated according to their internal syntactic structure even when they are used as phrases with individual reference. Hence, the proper noun construction, defined as the coupling of any linguistic form with the function of individual reference, is only partially captured in UD by the part-of-speech tag PROPN. One way of improving the correspondence would be to add the feature ExtPos=PROPN to names with internal syntactic structure.[5]

3 Modification Constructions

The information packaging function of modifiers, or attributive phrases, is to add information to help identify the referent of a referring phrase. Croft (2022) distinguishes six basic modification constructions, exemplified in (2):

(2) a. the *black* dog
  b. *five* books
  c. the *third* day
  d. *a pound of* sugar
  e. *Peter's* mother
  f. the man *who got away*

In (2a), *black* is an *adjectival modifier*, or simply *adjective* in Croft's terminology, which helps identify the referent of the head noun *dog* by adding information about a *property*, in this case its color. Besides color, adjectival modifiers commonly denote properties such as shape (*round*), age (*old*), value (*good*), and dimension (*big*) (Dixon, 1977). From an information packaging point of view, adjectival modifiers are said to be *subcategorizing*. Adjectival modifiers can be combined with *admodifiers*, like *very* in *a very big house*, which describe semantic operations on the scale denoted by the modifier. Besides intensifiers like *very*, admodifiers can be downtoners (*a rather big house*), comparatives (*a bigger house*), or superlatives (*the biggest house*).

The following three examples (2b–2d) instead belong to the class of *selecting* modifiers, whose information packaging function is to select an instance or set of instances using information about quantity or set membership. This is a diverse class, where three main constructions are distinguished: *numeral quantifiers*, *set-member terms*, and *mensural terms* (Croft, 2022, p. 109–111). Numeral quantifiers include cardinal numerals like *five* in (2b) together with a wide range of different quantifiers like *several* (a vague numeral), *most* (a proportional quantifier), and *each* (a distributive quantifier). Set-member terms include ordinal numerals like *third* in (2c) as well as non-numerical terms like *next*, *last*, and *other*. Numeral quantifiers and set-member terms both presuppose that the head noun denotes an individuated entity. This is not the case for mensural terms, exemplified by the measure term *a pound of* in (2d) and including a diverse set of constructions referring to containers (*cup*), groups (*flock*) or pieces (*slice*), among others (Koptjevskaja-Tamm, 2001).

The next modification construction, exemplified in (2e), is *nominal modification*, where the referent of the head noun is identified by its relation to another referent, such as the kinship relation in *Peter's mother*. The number of relations that can be invoked in this way is in principle unlimited, but other common relations are ownership (*Peter's car*), body-part relations (*Peter's arm*), and figure-ground relations (*the book on the table*) (Kay and Zimmer, 1990). From an information packaging point of view, nominal modifiers have a *situating* function, which makes them semantically similar to the pronoun and determiner constructions discussed in Section 2.

[5] The ExtPos feature can be used in UD to specify the part-of-speech category that a multiword expression would get if it were analyzed as a single word.

Construction	**Strategy**	**UD Annotation**
Adjectival modifier	Simple/Indexation	RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Admodifier	Word	ADJ $\xrightarrow{\text{advmod}}$ ADV
	Affix	ADJ[Degree=X]
Numeral quantifier	Simple/Indexation	RE $\xrightarrow{\text{nummod}}$ NUM[Features]
		RE $\xrightarrow{\text{nummod}}$ NUM $\xrightarrow{\text{clf}}$ NOUN
		RE $\xrightarrow{\text{det}}$ DET[Features]
		RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Set-member	Simple/Indexation	RE $\xrightarrow{\text{amod}}$ ADJ[Features]
Mensural quantifier	Simple	RE $\xrightarrow{\text{nmod}}$ NOM
	Flag: Adposition	NOM $\xrightarrow{\text{nmod}}$ RE $\xrightarrow{\text{case}}$ ADP
Nominal modifier	Simple/Indexation	RE $\xrightarrow{\text{nmod}}$ NOM[Features]
		RE $\xrightarrow{\text{det}}$ NOM[Features]
	Flag: Affix	RE $\xrightarrow{\text{nmod}}$ NOM[Case=X]
	Flag: Adposition	RE $\xrightarrow{\text{nmod}}$ NOM $\xrightarrow{\text{case}}$ ADP
	Compounding	RE $\xrightarrow{\text{compound}}$ NOM

Table 2: UD annotation of modification constructions and strategies. RE = referring expression; NOM = nominal (noun, proper noun or pronoun); Features = indexation features (possibly empty).

The final example in (2f) shows *action modification*, where the referent is being identified in relation to an event, and its prototypical realization in the form of a relative clause. Since subordinate clauses will be the topic of a later paper, we will not discuss action modification further in this paper, except to note that the relative clause construction can sometimes be recruited for property modification as an alternative to adjectival modification (Croft, 2022, pp. 113–114).[6]

The morphosyntactic strategies of modification constructions can be divided into four main types Croft (2022, pp. 114–138):

- **Simple:** The modifier is combined with the referring expression without any additional element; the modifier may be realized as an independent word (*juxtaposition*), through *compounding* or *affixation*. Juxtaposition is exemplified in the English examples (2a–c), where the adjective *black*, the cardinal numeral *five*, and the ordinal numeral *third* are all placed next to the head noun without any additional element.

- **Relational:** The combination involves a third element – a *flag* – that encodes the semantic relation between referent and modifier (Malchukov et al., 2010); the flag may be realized as an independent word (*adposition*) or as an affix (*case marker*). The former is exemplified by the preposition *de* in French *la mère de Pierre* (Pierre's mother), the latter by the genitive case inflection in Latin *mater Petri* (Petrus's mother); cf. Figure 4(a–b).

- **Indexical:** The combination involves a third element – an *index* – encoding features of the referent (such as person, number or gender/class) (Croft, 2003). Indices are most commonly realized as affixes, as in French *chien noir* (black dog), where the inflectional form of the adjective *noir* indicates the gender (masculine) and number (singular) of the noun; cf. Figure 2(a). A special case of the indexical strategy is the *classifier* strategy, commonly found with numerals, as in Chrau *du tong aq* (one crossbow), where the classifier *tong* (for long objects) is required with the numeral *du* (one) and the noun *aq* (crossbow) (Thomas, 1971).

- **Linker:** The combination involves a third element – a *linker* – invariant with respect to the features characteristic of flags and indices

[6] Recruitment occurs when one construction borrows the morphosyntactic form of another (usually more prototypical) construction (Croft, 2022, p. 53–58).

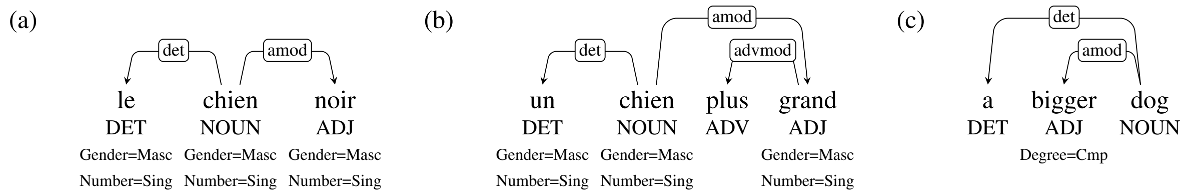

Figure 2: UD annotation of adjectival modifiers: (a–b) with indexation, (b–c) with admodifier.

(that is, it neither indicates a specific semantic relation nor encodes referent features). A typical example is Persian *âb-e garm* (hot water), where the linker *e* appears with the nominal *âb* (water) and the adjectival modifier *garm* (hot) but does not contrast with any other morpheme used to relate these construction elements to each other.

As discussed above, modifiers can furthermore be accompanied by admodifiers, which can be realized as independent words or affixes.

How does UD represent the different modification constructions and strategies? In Table 2, we give a schematic overview of the most important cases to be discussed in following subsections.

3.1 Adjectives and Admodifiers in UD

Adjectival modifiers are normally realized as independent words, which are analyzed in UD by attaching them to their nominal head with the *amod* relation and assigning them the part-of-speech tag ADJ. In an English example like *the black dog*, this is a simple juxtaposition strategy, as there is no additional word or morpheme mediating the relation. However, adjectives are also commonly found with the indexation strategy, where the adjective inflects to agree with the nominal head with respect to features such as gender and number, as illustrated for the corresponding French example *le chien noir* (the black dog) in Figure 2(a). In the latter case, the UD annotation combines the *amod* relation and the ADJ tag with morphological features on the adjective. As noted in Section 2.1, nothing in the annotation indicates that this is indexation, as opposed to the adjective and noun accidentally having identical features values for gender and number, but this is less problematic here because there is only one candidate controller.

Admodifiers realized as independent words are linked to their adjectival modifier heads with the *advmod* relation and are normally assigned the part-of-speech tag ADV, as shown for the French example *un chien plus grand* (a bigger dog) in Figure 2(b). Admodifiers realized as affixes are instead captured by morphological features on the adjective, as seen in the English example *a bigger dog* in Figure 2(c), where the adjective *bigger* is assigned the feature Degree=Cmp. It is worth noting that, while the features used for inflectional admodifiers are quite specific, the use of the *advmod* relation lumps the independent word admodifiers together with a very large and diverse group of expressions, including manner adverbials as well as temporal and locative expressions, among other things.

3.2 Selecting Modifiers in UD

While the annotation of adjectival modifiers in UD appears straightforward, the situation is more complex for the group of selecting modifier constructions. Starting with numeral quantifiers, UD clearly distinguishes *cardinal numerals*, which are annotated with the *nummod* relation and the part-of-speech tag NUM, categories that are not used for any other construction. This is illustrated for the Swedish example *fem böcker* (five books) in Figure 3(a). Other numeral quantifiers, like *many* and *all*, should according to the guidelines be annotated as determiners (with the *det* relation and the tag DET), like the Swedish example *alla böcker* (all books) in Figure 3(b). In practice, however, they are sometimes annotated as adjectival modifiers (with the *amod* relation and the tag ADJ), like the Swedish example *många böcker* (many books) in Figure 3(c).[7] Numeral quantifiers can also be realized using the *classifier* strategy, in which the numeral (or determiner) is combined with a special classifying element (often historically a noun and tagged as such in UD), as shown schematically in Table 2.

Set-member terms include ordinal numerals, which in UD are not analyzed using the *num-*

[7] A similar variation between DET and ADJ is found also in other languages, such as English (*all* = DET, *many* = ADJ) and French (*plusieurs* = DET, *tous* = ADJ).

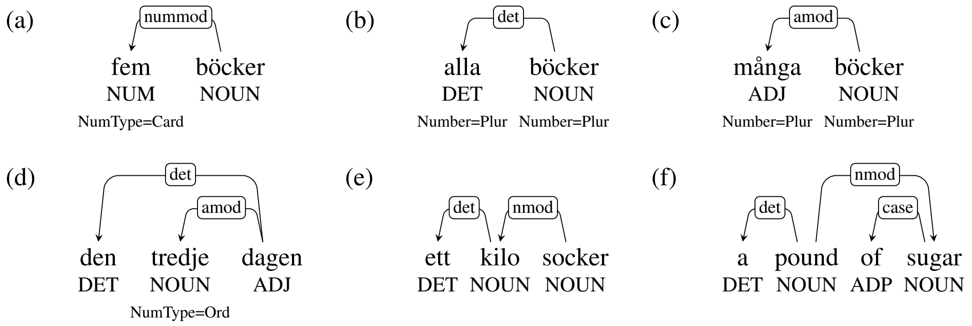

Figure 3: UD annotation of selecting modifiers: (a–c) numeral quantifier, (d) set member, (e–f) mensural quantifier.

mod relation and NUM tag reserved for cardinal numerals. Instead, they are analyzed as adjectival modifiers, as shown for the Swedish example *den tredje boken* (the third book) in Figure 3(d). The same analysis is used for non-numerical terms like *next*, *last*, and *other*, although we suspect that some of the more pronoun-like cases may be analyzed as determiners in some treebanks.

Mensural quantifiers are the construction type that causes the most problems for UD, because some of its strategies give rise to a mismatch between the syntactic and semantic head. The mensural part of this construction is typically realized as a nominal phrase, such as *a pound* or *a bottle*. When this is combined with the referring nominal using simple juxtaposition, as in the Swedish example *ett kilo socker* (a kilo of sugar) in Figure 3(e), the mensural part (*ett kilo*) can be analyzed as a modifier, linked to its head (*socker*) with the *nmod* relation, in which case the syntactic and semantic heads coincide. However, when the combination of the two parts involves an adposition, as in the English example *a pound of sugar* in Figure 3(f), the UD guidelines give priority to the grammatical form, which indicates that the mensural part is the syntactic head. As a result, we fail to capture the common construction in the two languages and end up with *nmod* relations going in opposite directions.

To arrive at a more transparent annotation of constructions and strategies, future versions of UD should consider some revisions of the annotation guidelines. Since the current distinction between *det*, *amod*, and *nummod* does not align well with comparative concepts from linguistic typology, these relations can be replaced by a general *mod* relation.[8] More specific constructions can be distinguished using part-of-speech tags, where we suggest restricting DET to demonstratives and articles, keeping ADJ for property words, and replacing NUM by a broader QNT category that includes quantifiers and set-member terms. If part-of-speech tags are not specific enough, syntactic subtypes may be used as well. In addition, we propose a new relation *admod* for the admodifier construction, which is distinct from other constructions currently covered by the *advmod* relation.

For mensural quantifiers, we advocate an analysis that consistently treats the measured noun as the head, as in the Swedish example in Figure 4(e). This can be seen as a *construction-oriented* analysis, where syntactic relations are aligned to capture constructional similarity whenever possible, as opposed to a *strategy-oriented* analysis, where priority is given to similarities in strategies. To resolve the apparent conflict between form and function in cases like the English example in Figure 4(f), we propose to analyze this as a linker strategy rather than a case marking strategy.

3.3 Nominal Modifiers in UD

Nominal modifiers are typically realized as nominal phrases, headed by a pronoun, noun or proper noun. With a few exceptions to be discussed shortly, they are analyzed in UD by attaching them to their nominal head with the *nmod* relation. Nominal modifiers are often realized using a relational strategy, with a flag that can take the form of an adposition, as in the French example *la mère de Pierre* (Pierre's mother) in Figure 4(a), or an affix, as in the Latin example *mater Petri* (Petrus's mother) in Figure 4(b). In the former case, the flag is attached to the nominal modifier with the *case* relation; in the latter, it is represented by a Case feature on the modifier.

There are two exceptions to the rule that nom-

[8] This change has been made in Surface Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018, 2019, 2021).

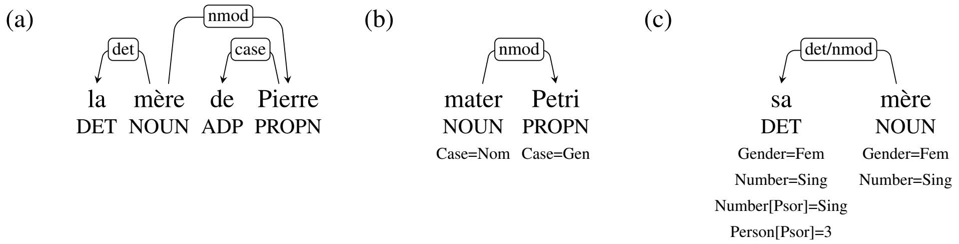

Figure 4: UD annotation of nominal modifiers.

inal modifiers are analyzed as an instance of the *nmod* relation. The first concerns possessive pronouns, as in the French example *sa mère* (his/her mother) in Figure 4(c). Possessive pronouns alternate with full nominal phrases to express nominal modification, so from a constructional point of view it makes sense to subsume them under the *nmod* relations. In many languages, however, they resemble determiners with respect to word order and/or indexation. The UD guidelines therefore currently allow possessives to be analyzed either as nominal modifiers (with the *nmod* relation) or as determiners (with the *det* relation).[9] We propose to use *nmod* consistently to capture constructional parallelism across languages. More generally, we think strategy-oriented relations should be restricted to strategies realized as independent function words, such as adposition flags, while relations between construction elements should be construction-oriented, as discussed above in connection with mensural classifiers.

The second exception is found in nominal modifiers that are realized with a compounding strategy, such as the English *iron pipe*, where the UD annotation prioritizes the information that the two components form a word, rather than a phrase, and attaches the modifier to the head with the *compound* relation. This highlights a more general issue in UD annotation, where construction-oriented relation types like *nmod* and strategy-oriented relation types like *compound* can come into conflict because the UD representation only allows one type for each syntactic relation. In line with the principle of prioritizing construction-oriented relations between construction elements, we propose to use *nmod* instead of *compound* for this kind of nominal modification. We are aware that this will blur the distinction between compounding and phrasal modification, but we think this can be remedied using features or subtyping.[10]

One final observation concerning nominal modification is that it is unclear how the linker strategy should be annotated in UD. If the linker is realized as an affix, a morphological feature can be added to its head, but if it is realized as an independent word, it seems the best we can do currently is to lump linkers and flags together under the *case* relation.[11] For future versions of UD, we therefore propose a new relation *lnk*, which can be used together with the tag PART to annotate linkers.

4 Discussion

While our review has shown that the UD annotation framework can represent all the major constructions and strategies for reference and modification discussed in Croft (2022), we have also seen that the correspondence between comparative concepts and elements of the UD annotation is quite complex. Of course, UD has developed without an explicit aim to represent constructions, let alone distinguish strategies for those constructions across languages. Nevertheless, our review reveals ways in which UD can capture properties of constructions and strategies, using a combination of syntactic relations and part-of-speech tags.

First, some, perhaps most, of the syntactic relations closely match the major information packaging constructions described in Croft (2022), albeit subdivided. For the constructions reviewed here, the main information packaging constructions are modification and admodification. Our suggestions

[9] As it happens, the French UD treebanks prefer the *det* analysis, while the English and Swedish UD treebanks use *nmod* instead, a discrepancy that does not appear to be motivated by the linguistic facts.

[10] It is worth noting in this context that the analysis of compounds in UD is not consistent across languages because of differences in orthography. Thus, the Swedish equivalent of *iron pipe*, *järnrör*, is currently not analyzed as a compound at all, because it is written as a single orthographic token.

[11] In UD v2.15, the Tagalog-Ugnayan and Cebuano-GJA treebanks use the *mark* relation for linkers that occur with nominal and adjectival modifiers, which is surprising given that the linkers do not mark relations between clauses.

for future revisions of the relations, specifically distinguishing *admod* from *advmod* and having a single *mod* relation replacing *det*, *amod* and *nummod*, increases this parallelism. Naturally, there will be some mismatches between construction function and morphosyntactic form, but we assume that these are relatively infrequent compared to the general matching of form and function. Another suggestion of ours that makes UD relations more closely match constructions is for a uniform representation of mensural constructions such that the measured noun is the head, which is facilitated by analyzing the etymological adposition in the pseudo-partitive strategy as a linker, not a flag.

Second, some of the part-of-speech categories are close to semantic categories. For example, NUM describes the semantic category of numerals; in fact, in many languages (most famously Russian), different numerals use different modification constructions. The categories NOUN, PROPN and PRON parallel the distinction between type reference, individual reference, and contextual reference. Modification is characterized by a similar information packaging function carried out by different semantic categories of modifiers. Our narrowing of DET to articles and demonstratives, and introducing the new tag QNT, again increases this parallelism.

Third, UD must represent independent words that form part of certain strategies (also known as "function words"), since they are part of morphosyntax. This is currently done using syntactic relations, part-of-speech tags, or both. For a better treatment of modification, we suggest a new *lnk* relation, distinct from *case*, to handle the grammaticalized function word in the pseudo-partitive strategy for mensural constructions, as well as in certain genitive and other modification constructions. The linker itself is assigned the tag PART, a tag that is used for function words in a variety of strategies.

5 Conclusion

In this paper, we have taken a first step towards a constructicon for UD, in the sense of Nivre (2025), by reviewing the way UD annotates constructions and strategies for reference and modification, following the taxonomy of Croft (2022). The constructicon is shown in Tables 1 and 2, where we outline how these constructions and strategies are currently annotated according to the UD guidelines. (In passing, we have also remarked on a few inconsistencies in the way that these guidelines are applied across languages.) On the positive side, we have found that UD can represent almost all constructions and strategies discussed in the survey. On the negative side, we have found that UD categories do not always align systematically with comparative concepts from typology, that there is sometimes a conflict between annotating elements of constructions and strategies, respectively, and that some strategies are not well captured in the UD framework.

In some cases, we have made concrete proposals for future revisions of UD, revisions that would improve the correspondence between UD categories and comparative concepts. It is worth clarifying that these revisions are incompatible with the current version of the UD guidelines (v2), since they involve changes to the set of syntactic relations and part-of-speech tags, which means that they could only be considered for v3 of the guidelines. Moreover, these proposals need to be evaluated also from other perspectives, since UD is designed as "a very subtle compromise between a number of competing criteria" (de Marneffe et al., 2021, p. 302) and should be suitable for language-specific analysis as well as typological language comparison, and should be accessible to non-experts and suitable for processing by computers. Finally, the discussion needs to be informed by a more comprehensive review of the UD framework, covering all major types of constructions and strategies. It is our goal to continue this review in a series of future publications.

Acknowledgments

We are grateful to members of the UniDive COST Action (CA21167) for useful discussions, and acknowledge the support of the Swedish Research Council (grant no. 2022-02909). We thank André Coneglian, Leonie Weissweiler and three anonymous reviewers for comments and suggestions that helped improve the article.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Mira Ariel. 1990. *Accessing Noun Phrase Antecedents*. Routledge.

William Croft. 2003. *Typology and Universals. Second Edition*. Cambridge University Press.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75.

Roger M. W. Dixon. 1977. Where have all the adjectives gone? *Studies in Language*, 1:19–80.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-Syntactic Universal Dependencies (SUD): Surface-syntactic relations and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories*, pages 126–132.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? Go SUD! Theoretical and practical benefits of the surface-syntactic distributional approach. In *Proceedings of the 6th International Conference on Dependency Linguistics*, pages 35–46.

Talmy Givón. 1983. Topic continuity in discourse: An introduction. In Talmy Givón, editor, *Topic Continuity in Discourse*, pages 1–41. John Benjamins.

Martin Haspelmath. 1997. *Indefinite Pronouns*. Oxford University Press.

Paul Kay and Karl Zimmer. 1990. On the semantics of compounds and genitives in English. In S. L. Tsohatzidis, editor, *Meanings and Prototypes: Studies in Linguistic Categorization*, pages 239–246. John Benjamins.

Maria Koptjevskaja-Tamm. 2001. 'a slice of the cake' and 'a cup of tea': partitive and pseudo-partitive constructions in the Circum-Baltic languages. In Östen Dahl and Maria Koptjevskaja-Tamm, editors, *Circum-Baltic Languages: Their Typology and Contacts*, pages 523–568. John Benjamins.

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig, and Ely A. Matos. 2024. MoCCA: A model of comparative concepts for aligning constructicons. In *Proceedings of the 20th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, pages 93–98.

Andrej Malchukov, Bernard Comrie, and Martin Haspelmath. 2010. Ditransitive constructions: a typological overview. In Andrej Malchukov, Bernard Comrie, and Martin Haspelmath, editors, *Studies in Ditransitive Constructions: A Comparative Handbook*, pages 1–64. Mouton de Gruyter.

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

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Joakim Nivre. 2025. Constructions and strategies in Universal Dependencies. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 419–423.

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.

David D. Thomas. 1971. *Chrau Grammar*. University of Hawaii Press.