Universal Dependencies

Marie-Catherine de Marneffe
The Ohio State University
Department of Linguistics
demarneffe.1@osu.edu

Christopher D. Manning
Stanford University
Department of Linguistics
manning@cs.stanford.edu

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

Daniel Zeman
Charles University
Faculty of Mathematics and Physics
zeman@ufal.mff.cuni.cz

Universal dependencies (UD) is a framework for morphosyntactic annotation of human language, which to date has been used to create treebanks for more than 100 languages. In this article, we outline the linguistic theory of the UD framework, which draws on a long tradition of typologically oriented grammatical theories. Grammatical relations between words are centrally used to explain how predicate–argument structures are encoded morphosyntactically in different languages while morphological features and part-of-speech classes give the properties of words. We argue that this theory is a good basis for crosslinguistically consistent annotation of typologically diverse languages in a way that supports computational natural language understanding as well as broader linguistic studies.

1. Introduction

Universal dependencies (UD) is at the same time a framework for crosslinguistically consistent morphosyntactic annotation, an open community effort to create morphosyntactically annotated corpora for many languages, and a steadily growing collection of such corpora. In all these respects, UD has undeniably been very successful, growing in only six years from ten treebanks and a dozen researchers to 183 treebanks for 104

Submission received: 2 July 2020; revised version received: 9 February 2021; accepted for publication: 25 February 2021.

https://doi.org/10.1162/COLI_a_00402

languages with contributions from 416 researchers around the world.[1] UD treebanks are now widely used in natural language processing research, including but not limited to research on syntactic and semantic parsing, and increasingly also in linguistic research, particularly on psycholinguistics and word order typology.

Some people think that UD is only a tool for annotation, and as such a rather eclectic approach building on existing de facto standards with many practical compromises. Although UD borrows terminology and concepts from many earlier grammatical theories, it is nevertheless a coherent theory resulting from a large amount of careful community work aiming at a principled but broadly applicable view of morphology and syntax. We believe that a clearer description of the underlying theory will help people to fully understand UD, its merits, and its limitations, and we attempt to articulate that theory, in particular, for version 2 of UD, in this article.[2]

The article is organized as follows. Section 2 introduces the basic theoretical assumptions of UD, including a commitment to words and grammatical relations as fundamental building blocks of grammatical structure. Section 3 is a survey of linguistic constructions and their analysis in UD, with examples from a broad range of languages. In Section 4, we zoom in on core arguments, which play a central role in UD, and discuss how they can be analyzed across typologically different languages. Section 5 discusses the design principles of UD against the backdrop of previous sections and Section 6 concludes with a brief outlook.

2. Basic Tenets of UD

The goal of UD is to offer a linguistic representation that is useful for morphosyntactic research, semantic interpretation, and for practical natural language processing across different human languages. It therefore puts an emphasis on simple surface representations that allow parallelism between similar constructions across different languages, despite differences of word order, morphology, and the presence or absence of function words.

2.1 Linguistic Representation and Information Packaging

When humans observe the world, they see entities (or objects) that participate in events (actions and states). The organization of all human languages reflects this basic world view. Therefore, in UD, we organize description around the two fundamental linguistic units of a **nominal**, canonically used for representing an entity, and a **clause**, canonically used for representing an event. Both nominals and clauses are often refined by describing an attribute of the entity or event, which can be done by the third fundamental linguistic unit of a **modifier**.

2.1.1 Heads and Dependents. A clause has a main predicate that expresses the state or action, and in most cases, states and actions involve participants expressed as nominals. In such a way, language has a hierarchical structure: Clauses can contain nominals, modifiers, and other clauses; nominals can also contain all three phrasal units; and modifiers

1 Release v2.7, November 15, 2020. For more information, see `https://universaldependencies.org`.

2 Because our focus in this article is theoretical, we do not go into practical matters concerning annotation, treebanks, and parsing. We also do not discuss the historical development of UD. For these aspects we refer to the papers on UD v1 (Nivre et al. 2016) and v2 (Nivre et al. 2020) and to the UD Web site (`https://universaldependencies.org`).

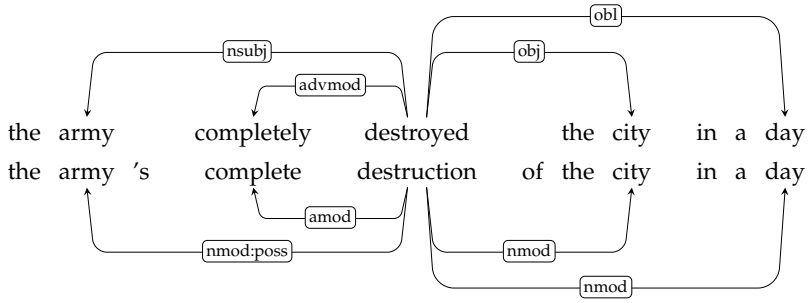


Figure 1
Partial UD analysis for a clause (top) and a nominal (bottom).

can contain modifiers. To express these ideas in UD, we adopt a **dependency grammar** perspective: A phrase has a **head** and other things that it contains are **dependents** of that head.

Dependency is a binary asymmetrical relation, which we represent in diagrams by an arrow from the head to the dependent (or, more precisely, to the head word of the dependent, when the dependent is itself a multiword unit), as in Figure 1. Through these dependencies, the words of a sentence are organized into a tree structure with the main predicate as the root.[3] Dependencies are typed with grammatical relation labels, as further discussed in Section 2.3. The head in a dependency is informally the main word of a phrase. The head of a nominal is canonically a noun. The head of a clause, commonly referred to as the **predicate**, is most commonly a verb but may also be an adjective or adverb, or even a nominal. The most common modifier heads are adjectives and adverbs. Sometimes linguistic head functions are divided between a structural center (an auxiliary or function word) and a semantic center (a lexical or content word), such as for periphrastic verb tenses like *has arrived*. This is what Tesnière (2015 [1959], ch. 23) refers to as a **dissociated nucleus**. In such cases, UD chooses the lexical or content word as the head, and makes function words dependents of the head in the dependency tree structure, while recognizing that they do form a nucleus together with the content word. A consequence of this decision, further discussed in Section 2.3.3, is that a UD tree represents a sentence's observed surface predicate–argument structure rather than necessarily accurately capturing phrase-internal syntactic constituency.

2.1.2 Nominals, Clauses, and Modifiers. In more detail, UD assumes a simple typology of three kinds of phrasal units (which might minimally be just a single word):

1. Nominals: the primary means for referring to entities
2. Clauses: the primary means for referring to events
3. Modifiers: the canonical attributive modifiers of nominals, clauses, and other modifiers

3 The tree constraint holds for the *basic* UD representation, which is the focus of this article. UD also defines an *enhanced* representation, which makes explicit additional implicit relations between words (such as propagating relations between conjuncts and adding subject relations for control and raising constructions). For more information about the enhanced representation, which is a rooted directed graph, see Nivre et al. (2020).

Nominals are similar to the notion of a noun phrase or determiner phrase in many theories, but encompass the entire nominal extended projection (Grimshaw 1991 [2005]), also covering prepositional phrases. While the basic use of nominals is always to refer to entities, they may be used in other functions. For example, most languages allow the nominalization of an event: *The continuation of hostilities* describes an event, but has the syntactic form of a nominal. Clauses can be either the root sentence or an embedded clause, typically express events and states, and have a main predicate, which is canonically a verb but can be other parts of speech used predicatively.

Both nominals and clauses can have their meaning added to by the presence of modifying phrases. Sometimes these phrases are themselves nominals or clauses. For example, in Example (1a), there is a nominal modifying a clause; in Example (1c), there is a nominal modifying a larger nominal; and in Example (1d), there is a clause modifying a nominal. However, sometimes modifying phrases are single words or smaller modifying phrases that do not expand into the same rich structures as nominals and clauses. We describe this third class of linguistic units as **modifiers**. In Example (1b), there is a modifier modifying a clause, and in Example (1e), there is a modifier modifying a nominal. Modifiers can themselves be modified: The modifier *somewhat* modifies *rusty* in Example (1e). It is generally true in languages that there is not an infinite regress: The modifiers of modifiers are limited and normally of the form of basic modifiers, and so we continue to call them all modifiers.

(1) a. [He opened the can [with a screwdriver]]
b. [He opened the can [carefully]]
c. [the screwdriver [on the table]]
d. [the screwdriver [which my mother bought me]]
e. [the [[somewhat] rusty] screwdriver]

This taxonomy is not unique to UD. As it reflects the basic structure of human language, similar taxonomies can be found in many other frameworks, especially those starting from a functional or typological perspective on language. For example, Croft (1991, forthcoming) distinguishes **reference**, **predication**, and **modification** as three basic information packaging functions, or propositional act functions, underlying syntactic constructions. These correspond straightforwardly to the canonical usages of our nominals, clauses, and modifiers, respectively.

The distinction between nominals and clauses is fundamental to UD, which systematically uses different dependency relations in the two types of structures, as illustrated in Figure 1. The clause *the army completely destroyed the city in a day* is headed by the verbal predicate *destroyed*, while the nominal *the army's complete destruction of the city in a day* is headed by the noun *destruction*. The predicate has two core arguments (*the army*, *the city*), while the noun has two genitive modifiers accompanied by different kinds of case markers (*the army's*, *of the city*). The adverbial modifier (`advmod`, *completely*) of the predicate corresponds to an adjectival modifier (`amod`, *complete*) of the noun. Even the temporal modifier *in a day*, which has the form of a prepositional phrase in both cases, is classified as an oblique modifier (`obl`) of the predicate but as a nominal modifier (`nmod`) of the noun. Similarly, the typology of dependency relations also captures whether the dependent is a nominal, a clause, or a modifier. For example, a modifier of a nominal will be respectively a nominal modifier (`nmod`), an adjectival modifier (`amod`), or an adnominal clause (`acl`) depending on the type of the dependent. Hence phrasal types are recoverable without being explicitly represented.

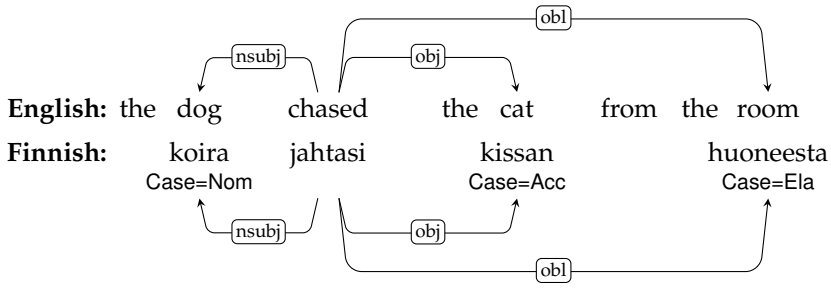


Figure 2
Simplified UD annotation for equivalent sentences from English (top) and Finnish (bottom).

2.2 Words as Basic Units

UD follows traditional grammar in giving primary status to **words**. Words are the basic elements connected by dependency relations; they have morphological properties and enter into syntactic relations. The primacy of words can be understood as a commitment to the lexical integrity principle (Chomsky 1970; Bresnan and Mchombo 1995; Aronoff 2007), which states that words are built out of different structural elements and by different principles of composition than syntactic constructions. Despite the challenges in defining words in a crosslinguistically consistent manner—faced with phenomena like clitics, compounding, and incorporation, to mention only a few[4]—we believe that this approach is more interpretable and useful for most potential users of UD and generalizes better across languages than trying to segment words into smaller units like morphemes. This view is further supported by developments in morphological theory, which favor word-based abstractive models over morpheme-based constructive models (Stump 2001; Blevins 2006; Blevins, Ackerman, and Malouf 2017).

It is important to note, however, that the morphosyntactic notion of word does not always coincide with orthographical or phonological units. For instance, clitics (Spencer and Luís 2012) often need to be separated from their hosts and treated as independent words even if they are not recognized as such in conventional orthography (for instance, the English *'s* genitive, as in *the army's* in Figure 1, acts as a phrasal clitic, as can be seen by expansions such as *the army of the undead's*). Similarly, compound words need a special treatment, because in some languages their written form may contain boundary markers such as whitespace (as in *night school* in English) whereas in other languages they do not (as in *Abendschule* 'night school' in German).

2.2.1 Content Words, Function Words, and Grammaticalization. We expect the words that enter into the main syntactic relations to be autosemantic, that is, content words with an independent meaning—typically verbs, nouns, adjectives, or adverbs, as well as corresponding pro-forms with a contextually determined referential meaning. For instance, the backbone of the UD morphosyntactic representations for the English and Finnish sentences in Figure 2 consists of the three relations that are common to these sentences: argument and modifier relations involving predicates and nominals.

These content words often occur together with grammatical markers, synsemantic elements that further specify their meaning or syntactic role. Typical examples are

4 See also Haspelmath (2011a).

markers of tense, mood, and aspect for (verbal) predicates and of number, definiteness, and case for nominals. As explained in Section 2.1.1, UD attaches such elements as dependents of the content word, analyzing them as parts of dissociated nuclei.

The distinctions between content words and function words, and between function words, clitics, and inflectional morphemes, are not always clear-cut. We know from the literature on grammaticalization that grammatical markers normally develop out of content words and first appear as separate function words but often later become clitics and eventually inflectional affixes, a process sometimes referred to as the cline of grammaticalization (Hopper and Traugott 2003). At any given historical stage, a language will contain constructions that are at intermediate stages of this development, and where it is therefore not straightforward to classify the components of the construction. Consider, for example, the Swedish sentences in Example (2).

(2) a. Hon kunde (*att) sjunga
she could to sing
'She could sing'

b. Hon började (att) sjunga
she began to sing
'She began to sing'

c. Hon gillade *(att) sjunga
she liked to sing
'She liked to sing'

In Example (2a), it is impossible to insert the infinitive marker *att* before the verb *sjunga* 'sing', which shows that *kunde* 'could' is an auxiliary verb. In Example (2c), it is equally impossible to *omit* the infinitive marker, which shows that *gillade* 'liked' is a main verb taking an infinitive complement. In Example (2b), however, the infinitive marker is optional, which makes the status of *började* 'began' unclear, all the more as its meaning is mainly aspectual and of a kind that could undergo grammaticalization. In annotation, we are forced to make a somewhat arbitrary choice and make Example (2b) parallel to either Example (2a) or Example (2c)—but not both. Note that in Example (2a), the verb *sjunga* is the head of the sentence, while in Example (2c), *gillade* is the head, so changing the analysis of Example (2b), as grammaticalization proceeds, requires not just a part-of-speech change but a fundamental syntactic reanalysis of the sentence, or what Gerdes and Kahane (2016) refer to as a "catastrophe." Making particular categorical decisions in such intermediate cases will inevitably add some distortion to our representation of the linguistic reality, but we can only do our best to maintain consistency in these decisions and carefully document the criteria.

Similar issues arise in word segmentation, where it is sometimes difficult to decide whether a grammatical marker should be treated as an inflectional affix, clitic, or function word, despite extensive discussion of discriminative criteria, such as in Zwicky and Pullum (1983).

2.2.2 Part-of-Speech Categories. All linguistic theories assume that words can be classified by a **word class** or **part of speech** (POS) according to their behavior within the language system. Partly for broad comprehensibility, UD stays fairly close to traditional parts of speech, such as the eight parts of speech commonly recognized for English, but it makes a few finer distinctions, better reflecting modern linguistic typology, and adds some classes for punctuation and other symbols. As a result, UD distinguishes 17 coarse-grained classes of words and other elements of text, and assigns them the

Table 1
Universal part-of-speech tags (UPOS). Typos and abbreviations are given the category of the unabbreviated or correct word.

Traditional POS	UPOS	Category
noun	NOUN	common noun
	PROPN	proper noun
verb	VERB	main verb
	AUX	auxiliary verb or other tense, aspect, or mood particle
adjective	ADJ	adjective
	DET	determiner (including article)
	NUM	numeral (cardinal)
adverb	ADV	adverb
pronoun	PRON	pronoun
preposition	ADP	adposition (preposition/postposition)
conjunction	CCONJ	coordinating conjunction
	SCONJ	subordinating conjunction
interjection	INTJ	interjection
–	PART	particle (special single word markers in some languages)
–	X	other (e.g., words in foreign language expressions)
–	SYM	non-punctuation symbol (e.g., a hash (#) or emoji)
–	PUNCT	punctuation

labels ("universal part-of-speech tags," UPOS) shown in Table 1. These categories are widely attested in the world's languages. We do not claim that all languages must use all of these categories, but we do assume that every word in every language can be assigned one of them. Some word-class distinctions are particularly important in UD: For example, the dividing line between nouns and verbs plays a significant role in specifying whether a constituent is nominal or clausal (Section 2.1.2).

It is not easy in all cases to define word classes in a crosslinguistically consistent manner. Grammatical criteria used in word classification have to be specific for individual languages, although we do expect similar criteria in languages that are closely related. Because morphological criteria are not sufficient and available for all categories in all languages, in many cases we have to rely primarily on syntactic criteria. For instance, Czech adjectives inflect for three grammatical genders, two numbers, seven cases, and three degrees of comparison. They typically specify properties of nouns and are found right before the nouns they modify. In contrast, only a subset of English adjectives can inflect for degree of comparison, and none inflect for gender, number, or case. Yet their prototypical function and distribution is similar to Czech: If used attributively, they occur right before the nouns whose attributes they specify.

While the definition of word categories is not universal, their names are portable across languages so that same-labeled categories show partially similar syntactic behavior and overlapping semantic content (Schachter and Shopen 2007; Haspelmath 2001; Croft 1991). It is possible to have one category that will contain most words referring to entities, such as *mother, dog*, or *house*; words in this category will be called nouns. Similarly, the label "verb" is used for the class of prototypical action words (such as *go, buy, eat*), and the class of adjectives will likely contain equivalents of *small, good*, or *white.* In addition, each of these categories may contain words with less prototypical semantics, if they follow the language-particular rules that define the category. Hence the English nouns include words like *destruction* and *weakness* because their morphological and distributional behavior is noun-like, although their meaning is derived from the verb *destroy* and the adjective *weak*, respectively.

A common difficulty is that words of one category are sometimes used in positions and functions normally associated with a different category, without changing their morphology (if morphological criteria are available at all). For example, an English adjective may appear in the subject position with a definite article but without the modified noun (*the healthy*, *the sick*). We could treat such examples as instances of ellipsis (where the underlying noun phrase could be *the healthy/sick people*) or we could say that the adjective has been converted into a noun in the given sentence. However, the part-of-speech classification is most useful if it captures regular, prevailing syntactic behavior and does not reflect sentence-specific exceptional behavior. If the POS category were completely predictable from the syntactic function (which is an independent part of UD annotation), then the POS tag would be uninformative. It would also be harder to find interesting crosslinguistic differences, for example, that language X allows words of category A to have syntactic function b, but language Y does not. Therefore in English we assign the ADJ tag to *healthy* even if it heads a nominal phrase.

Sometimes a functional shift is better explained by grammaticalization (see Section 2.2.1) rather than by exceptional usage in a specific sentence. The English adverb *so* is used as an adverb in Example (3a) and Example (3b), but as a discourse connective similar to a coordinating conjunction in Example (3c). However, we keep the word *so* in the adverb category in these three examples.

(3) a. People work so hard

b. If you have not done so already

c. We are aiming to have it next week, so I need to know if you can ship it quickly

Nevertheless, there are situations where we consider the two competing functions too distant and mutually incompatible, and we treat the word as a homonym whose category has to be disambiguated by context. Consider the Spanish examples in Example (4).

(4) a. los siete candidatos que compiten mañana
the seven candidates that compete tomorrow
'the seven candidates that will compete tomorrow'

b. Descubrimos que los tres reyes estaban aquí
discovered.1PL that the three kings were here
'We discovered that the three kings were here'

In Example (4a), the word *que* 'that' is a relative pronoun that represents the subject of the subordinate clause; we tag it PRON. On the other hand, the same word in Example (4b) has no argument role in the subordinate clause; it is merely a marker of subordination. We tag it SCONJ. The same holds for the English word *that* in the corresponding English sentences. Sometimes morphology in a paradigm makes the analysis clear: When English nouns are used as verbs like in *You should butter your bread*, we regard the word as a verb because it participates in a paradigm with usual verb morphology in the past tense or with third singular subject agreement.

2.2.3 Morphological Features. Many classes of words in many languages participate in paradigms of forms that express extra features, such as number or tense. We can further divide the appropriate POS classes into subclasses according to features that express

Table 2
Universal morphological features.

	Feature	Values
pronominal type	PronType	Art Dem Emp Exc Ind Int Neg Prs Rcp Rel Tot
numeral type	NumType	Card Dist Frac Mult Ord Range Sets
possessive	Poss	Yes
reflexive	Reflex	Yes
foreign word	Foreign	Yes
abbreviation	Abbr	Yes
wrong spelling	Typo	Yes
gender	Gender	Com Fem Masc Neut
animacy	Animacy	Anim Hum Inan Nhum
noun class	NounClass	Bantu1-23 Wol1-12 ...
number	Number	Coll Count Dual Grpa Grpl Inv Pauc Plur Ptan Sing Tri
case	Case	Abs Acc Erg Nom Abe Ben Cau Cmp Cns Com Dat Dis Equ Gen Ins Par Tem Tra Voc Abl Add Ade All Del Ela Ess Ill Ine Lat Loc Per Sub Sup Ter
definiteness	Definite	Com Cons Def Ind Spec
comparison	Degree	Abs Cmp Equ Pos Sup
verbal form	VerbForm	Conv Fin Gdv Ger Inf Part Sup Vnoun
mood	Mood	Adm Cnd Des Imp Ind Irr Jus Nec Opt Pot Prp Qot Sub
tense	Tense	Fut Imp Nfut Past Pqp Pres
aspect	Aspect	Hab Imp Iter Perf Prog Prosp
voice	Voice	Act Antip Bfoc Cau Dir Inv Lfoc Mid Pass Rcp
evidentiality	Evident	Fh Nfh
polarity	Polarity	Neg Pos
person	Person	0 1 2 3 4
politeness	Polite	Elev Form Humb Infm
clusivity	Clusivity	In Ex

paradigmatic position. For example, the VerbForm feature distinguishes the finite verb from various nonfinite forms, which often show a mix of verbal properties and properties of other classes (nouns, adjectives, or adverbs). Depending on the language, it is then possible to distinguish between a finite verb (VERB VerbForm=Fin), a verbal participle (VERB VerbForm=Part), a deverbal participial adjective (ADJ VerbForm=Part), and a common adjective (ADJ). It seems quite possible to define a universal set of features, covering what is described by morphology in the world's languages, and our system in UD is in line with other attempts to do this, such as UniMorph (Sylak-Glassman et al. 2015) and the GOLD Ontology (Farrar and Langendoen 2003).

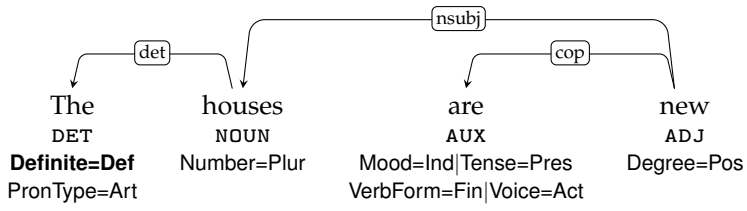
UD defines a set of feature-value pairs that are attested in multiple languages (Table 2).[5] Additional features may be defined in language-specific documentation for use in individual languages. Some features are lexical, meaning that the same value of the feature applies to the entire paradigm, that is, to all forms that share the same lemma. Such features serve to further partition the space of word categories by providing distinctions that are more fine-grained, or that cut across the boundaries of the

5 In the examples throughout this article, we show only selected Feature=Value pairs that we think are useful to understand the example. The actual UD annotation may contain other features that we omit in the interest of space.

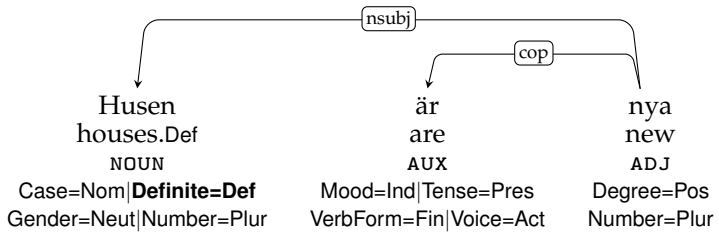
main UPOS categories. Prominent examples are PronType and NumType. For example, the interrogative and indefinite pronominal types are recognized with pronouns (*who* vs. *somebody*), determiners (*which* vs. *some*), as well as with adverbs (*where* vs. *somewhere*). Other features are inflectional, namely, different forms in a word's paradigm may have different values of the feature. A typical example, attested in many languages, is Number: A noun may have a special form when referring to more than one entity, a verb may cross-reference an argument and signal that it is a group of entities, and in many languages the nominal and the verbal inflection coexist, possibly accompanied by number inflection of other categories, such as adjectives. Finally, there are features whose nature differs depending on the part-of-speech category. For instance, Gender is a lexical feature of nouns but it is also an inflectional feature of words that show morphological agreement with nouns, such as verbs or adjectives.

A feature may also be marked on a function word that contributes the feature to a dissociated nucleus. For example, definiteness of nouns is expressed morphologically in Swedish (*husen* 'the houses' vs. *hus* 'house(s)'), hence Definite is an inflectional feature of Swedish nouns, but English nominals derive their definiteness from definite or indefinite articles (as shown in Example (5) and Example (6)). These are function words and the definite article is assigned a different lemma than the indefinite article, hence Definite is a lexical feature of articles (determiners) in English.

(5)



(6)



'The houses are new'

UD does not impose any universal constraints on compatibility between features and part-of-speech categories. Any feature value can potentially occur with any UPOS tag. However, constraints of this sort typically exist at the language-particular level. Hence, for instance, the Number feature is defined for English nouns and verbs but not adjectives, while in Czech it is defined for adjectives, too. The Case feature in English appears marginally with certain pronouns and uses only two values, while in Czech it has seven values and is defined for nouns, adjectives, pronouns, determiners, and numerals.

In some languages, some features are marked more than once on the same word. We say that there are several **layers** of the feature. The exact meaning of individual layers is language-dependent. For example, possessive adjectives, determiners, and pronouns may have two different values of both Gender and Number. One of the values is determined by agreement with the modified (possessed) noun. This is parallel to other (non-possessive) adjectives and determiners that agree in gender and number with the nouns they modify. The other value is determined lexically because it is a property of

the possessor. Layers are indicated by their identifier in square brackets after the feature name. For example, the Czech possessive pronoun *náš* 'our' is tagged Number[psor]=Plur to indicate that the possessor's number is plural. At the same time, the word form refers to a singular possessee; this layer of the Number feature is considered default for Czech possessives and needs no layer identifier: Number=Sing.

Where necessary, UD allows language-specific features, for example, FocusType has been used in the Niger–Congo language Wolof but it has not yet been established as applicable in other languages. It is used with Wolof auxiliaries, which, among other things, indicate whether the focus is on the subject of the clause, the verb, or the verb's complement (Dione 2019).

2.3 Grammatical Relations between Words

Perhaps the most distinctive feature of UD is its taxonomy of grammatical relations between words.[6] Each dependent of a head, and also any function words that belong with a head, are connected to the head via a grammatical relation drawn from a universal typology of 37 grammatical relations, listed in Table 3. As discussed earlier, the grammatical relations are organized around whether the head is the head of a clause or nominal, and whether the dependent is a clause, nominal, or modifier, although a number of other distinctions and special cases, prominent in the world's languages, are also represented. Table 4 illustrates the organization of the grammatical relations. The `root` relation is used for the root of the sentence, with a dummy head that does not need to be explicit. The `dep` relation is used when no other relations are deemed appropriate. The relations are illustrated throughout Section 3. The set of allowed relations is closed, but UD allows relation subtypes separated from the main relation by a colon to provide further distinctions or to capture language-specific constructions. For example, a number of languages mark relative clauses as `acl:relcl` and predeterminers as `det:predet`.

2.3.1 Usefulness of Grammatical Relations for Linguistic Typology. One of the basic tenets of UD is that *grammatical relations* like *subject* and *object* provide a useful level of abstraction to account for the complex mapping from overt coding properties like case-marking, agreement, and word order to the underlying semantic predicate–argument structure of sentences. In particular, they provide a happy middle ground of usually being easily surface-form identifiable while being useful for crosslinguistic description.

In this respect, UD follows in the tradition of theories as diverse as relational grammar (Perlmutter 1983), lexical-functional grammar (LFG) (Kaplan and Bresnan 1982; Dalrymple 2001; Bresnan et al. 2016), word grammar (Hudson 1984, 1990), functional generative description (FGD) (Sgall, Hajičová, and Panevová 1986), meaning-text theory (MTT) (Mel'čuk 1988; Milicevic 2006), role and reference grammar (Van Valin, Jr. 1993), and head-driven phrase structure grammar (Pollard and Sag 1994). Moreover, grammatical relations have always played a prominent role in linguistic typology, starting with the pioneering works of Greenberg (1963) and then Comrie (1981), and continuing in contemporary work like that of Croft (2001, 2002), Andrews (2007), Dixon (2009), and Haspelmath (2011b). Although the universality of grammatical relations is sometimes debated, their status as useful theoretical constructs for crosslinguistic studies is rarely questioned.

6 We generally use the term *grammatical relation* rather than *grammatical function* or *dependency label*, but we regard the terms as essentially synonymous—unlike, for example, Andrews (2007).

Table 3
The 37 syntactic relations in UD, with a brief explanation of the relation and a reference to an example.

Relation	Definition	Ex.
acl	adnominal clause; finite or non-finite clause modifying a nominal	(28)
advcl	adverbial clause modifying a predicate or modifier word	(27)
advmod	adverb or adverbial phrase modifying a predicate or modifier word	(20a)
amod	adjectival modifier of a nominal	(12)
appos	appositional modifier; a nominal used to define, name, or describe the referent of a preceding nominal	(15)
aux	auxiliary; links a function word expressing tense, mood, aspect, voice, or evidentiality to a predicate	(16c)
case	links a case-marking element (preposition, postposition, or clitic) to a nominal	(9)
cc	links a coordinating conjunction to the following conjunct	(23)
ccomp	clausal complement of a verb or adjective without an obligatorily controlled subject	(26b)
clf	(numeral) classifier; a word reflecting a conceptual classification of nouns linked to a numeric modifier or determiner	(11)
compound	any kind of word-level compounding (noun compound, serial verb, phrasal verb)	(37)
conj	conjunct; links two elements which are conjoined	(23)
cop	copula; links a function word used to connect a subject and a nonverbal predicate to the nonverbal predicate	(17a)
csubj	clausal syntactic subject of a predicate	(25)
dep	unspecified dependency, used when a more precise relation cannot be determined	
det	determiner (article, demonstrative, etc.) in a nominal	(10)
discourse	discourse element (interjection, filler, or non-adverbial discourse marker)	(20b)
dislocated	a peripheral (initial or final) nominal in a clause that does not fill a regular role of the predicate but has roles such as topic or afterthought	(22b)
expl	expletive; links a pronominal form in a core argument position but not assigned any semantic role to a predicate	(22c)
fixed	fixed multiword expression; links elements of grammaticalized expressions that behave as function words or short adverbials	(39)
flat	flat multiword expression; links elements of headless semi-fixed multiword expressions like names	(40)
goeswith	links parts of a word that are separated but should go together according to standard orthography or linguistic wordhood	(44)
iobj	indirect object; nominal core argument of a verb that is not its subject or (direct) object	(16c)
list	links elements of comparable items interpreted as a list	(46)
mark	marker; links a function word marking a clause as subordinate to the predicate of the clause	(27a)
nmod	nominal modifier; a nominal modifying another nominal	(13)
nummod	numeric modifier; numeral in a nominal	(10)
nsubj	nominal subject; nominal core argument which is the syntactic subject (or pivot) of a predicate	(16)
obj	object; the core argument nominal which is the most basic core argument that is not the subject, typically the most directly affected participant	(16)
obl	oblique; a nominal functioning as a non-core (oblique) modifier of a predicate	(21)
orphan	links orphaned dependents of an elided predicate	(43)
parataxis	links constituents placed side by side with no explicit coordination or subordination	(32)
punct	punctuation attached to the head of its clause or phrase	(23b)
reparandum	repair of a (normally spoken language) disfluency	(45)
root	root of the sentence	(16)
vocative	nominal directed to an addressee	(22a)
xcomp	clausal complement of a verb or adjective with an obligatorily controlled subject	(26a)

2.3.2 Core Arguments and Oblique Modifiers. In classifying grammatical relations, UD distinguishes the **core arguments** of a predicate, essentially subjects and objects, from all other dependents at the clause level, collectively referred to as **oblique modifiers**. The core–oblique distinction is commonly assumed in typological linguistics (see, e.g., Thompson 1997; Andrews 2007) and is ultimately an information packaging distinction.

Table 4
Typology of the syntactic relations.

Head \ Dependent	Nominals	Clauses	Modifier words	Function words
Clausal core arguments	nsubj obj iobj	csubj ccomp xcomp		
Clausal non-core arguments	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

All or nearly all languages have a standard way of encoding the one or two arguments of most verbs, and this unmarked form of argument expression defines core arguments for that language. The specific criteria used to identify core arguments are ultimately language-specific, but the following criteria recur in many languages:

- Verbs usually only agree with core arguments.
- Core arguments often appear as bare nominals while obliques are marked by adpositions or other grammatical markers.
- Core arguments often appear in certain cases, traditionally called nominative, accusative, and absolutive.[7]
- Core arguments in many languages occupy special positions in the clause, often adjacent to the verb.
- Properties such as being the controller of a subordinate clause argument are often limited to core arguments.
- Valency-changing operations such as passive, causative, and applicative are often restricted to the promotion or demotion of core arguments.

UD assumes that all languages have a way of identifying usually two core arguments, and reserves the relations of **subject** and **object** for these. If additional dependents that are treated similarly to the basic core arguments appear in a clause, with or without valency-changing operations targeting them, these are also regarded as core arguments. For example, some languages allow indirect (or secondary) objects, while other languages do not.

It is important to note that status as a core argument is decoupled from the semantic role of a participant. Depending on the meaning of a verb, many different semantic

7 See Section 4.4 for a discussion of ergative case.

roles can be expressed by the same means of encoding core arguments. Nevertheless, there is a correlation: Agent and patient or theme roles of predicates in their unmarked valence are normally realized as core arguments. It is also important to note that the core–oblique distinction has to do with the morphosyntactic encoding of dependents, not with their status as obligatory or selected by the predicate. Thus, UD does not assume the traditional argument–adjunct distinction found in many linguistic theories, which we take to be sufficiently subtle and hard to apply consistently both within and across languages that the best solution is to avoid it. This position has been defended on theoretical grounds by Haspelmath (2014) and Przepiórkowski (2016), and is also adopted for practical reasons in many treebanks, notably the Penn Treebank for English (Marcus, Santorini, and Marcinkiewicz 1993). The distinction between core arguments and oblique modifiers is only applied at the clausal level; all dependents of nominals are treated as oblique.

2.3.3 UD as Tectogrammar. The emphasis on grammatical relations makes UD representations similar to syntactic representations that are midway between surface constituency and argument structure in multistratal theories, such as the f-structures in LFG (Bresnan et al. 2016), the deep syntactic or tectogrammatical representations in multistratal versions of dependency grammar (Sgall, Hajičová, and Panevová 1986; Mel'čuk 1988), or final relations in relational grammar. In particular, UD captures the observed surface predicate–argument structure rather than any sort of abstracted or underlying deeper structure. However, being a monostratal theory, UD also needs to incorporate aspects of surface realization, such as word order, function words, and morphological inflections, which typically belong to a separate surface-oriented representation in multistratal theories. As a result, UD representations end up looking like a hybrid of deep and surface-oriented representations, but where the tree structure is primarily determined by predicate–argument structure. We believe the failure to appreciate this to be one of the primary causes for misunderstandings about the theoretical foundations of UD. More specifically, this means that UD represents classic surface constituency only to the level of demarcating clauses, nominals, and modifiers. The internal structure of each of these phrases represents predicates and grammatical relations, somewhat similarly to an LFG f-structure, an MTT SyntR, or an FGD tectogrammatical representation, and commonly does not capture fine details of surface constituency as regards auxiliary verbs, adpositions, and so on.[8]

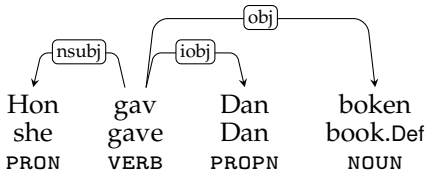
3. Analyzing Linguistic Constructions in UD

Having explained the basic principles of UD as a linguistic theory, we now illustrate how this theory can be applied to a range of linguistic phenomena. We start with nominals and (simple) clauses, as the most fundamental constructions, and gradually move on to more complex phenomena, including some that are ubiquitous in language use but not often discussed in grammars.

8 In contrast, Osborne and Gerdes (2019) and Gerdes et al. (2018) argue for and present a dependency annotation model that does respect surface constituency, while other annotation schemes are closer to (semantic) argument structure (Baker, Fillmore, and Lowe 1998; Palmer, Gildea, and Kingsbury 2005).

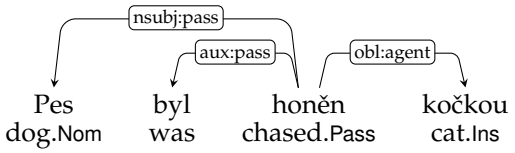
3.1 Nominals

Nominals[9] are a fundamental linguistic unit in all languages, and typically refer to entities (in a wide sense). Nominals occur as core arguments of predicates and in a range of other functions, including predicative uses. In the simplest case, a nominal consists of a single head word, which is typically a noun (NOUN), proper noun (PROPN), or pronoun (PRON). Depending on the language, nominal head words may carry a number of morphological features, of which the most common are gender (Gender), number (Number), case (Case), and definiteness (Definite). In the Swedish Example (7), the subject nominal is the pronoun *hon* 'she', the indirect object nominal is the proper noun *Dan*, and the direct object nominal is the noun *boken* 'the book'.

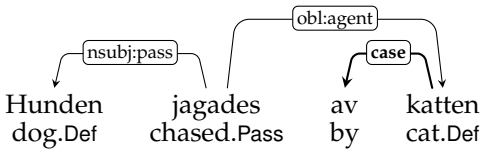


(7) 'She gave Dan the book'

3.1.1 Case Markers. Case marking is one of the strategies that languages use to encode the grammatical function of a nominal. Case marking can be realized through morphological inflection (captured in UD by the Case feature) or by clitics or adpositions (prepositions and postpositions). In the interest of crosslinguistic parallelism, UD takes a radical approach and treats all adpositions as case markers, attaching them to the nominal head with the special case relation.[10] This allows us to analyze the following examples as both having a direct dependency relation from the predicate to the nominal filling the (oblique) agent role of a passive, despite the fact that Czech Example (8) uses a noun in the instrumental case (*kočkou*) while Swedish Example (9) adds a preposition (*av*):



(8) 'A/the dog was chased by a/the cat'



(9) 'The dog was chased by the cat'

9 The term *nominal* is roughly equivalent to the more commonly used term *noun phrase*. However, we prefer the term *nominal* both because phrases are not primitive notions in UD and because we include among nominals some constructions that would not normally be classified as noun phrases, notably, prepositional phrases.

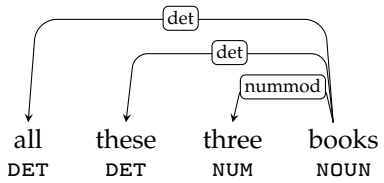
10 In the typological linguistics literature, Haspelmath (2019) also argues for a unified treatment of case markers and adpositions, suggesting it is "very unclear how they could be distinguished consistently as comparative concepts."

This means that prepositional (and postpositional) phrases are treated in UD as nominals, where the nominal head is the referential core while the adposition is a functional marker. This can be seen as an instantiation of Tesnière's notion of a dissociated nucleus and does not entail that the adposition is seen as a syntactic dependent of the noun in the narrow sense.

3.1.2 Determiners, Numerals, and Classifiers. Nominals headed by nouns often contain determiners, which can be roughly divided into four classes: articles, demonstratives, interrogatives, and quantifiers. Articles, like English *a*(*n*) and *the*, specify definiteness or related properties. They are obligatory in some languages (at least with some types of nouns), and completely absent in others. Demonstratives, like Latin *hic* 'this', *iste* 'that (of yours)' and *ille* 'that', anchor the noun phrase deictically and seem to be available in all languages. Interrogatives, like English *which*, are used to form noun phrases that can be used in interrogative (and sometimes relative) clauses. Quantifiers, like French *tout* 'all', *quelque* 'some', and *aucun* 'any', specify quantity or existence of the referent. In many languages, different determiners are in complementary distribution or have special constraints on their cooccurrence and possible order. Regardless of whether a noun phrase contains one or more determiners, UD uses the `det` relation to connect them all directly to the nominal head, as illustrated in Example (10) below.

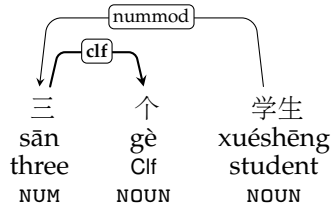
Nominals headed by nouns may also contain numerals, which express exact numerical quantities (1, 2, 3, ...). Numerals resemble determiners and can often replace them (*one book* vs. *a book* or *this book*) but have special properties in many languages, in particular in relation to classifiers (see below), and UD therefore uses the special `nummod` relation to connect a numeral to the head noun, as in Example (10). Note that the `nummod` relation is only used for cardinal numerals (*one*, *two*, *three*). Ordinal numerals (*first*, *second*, *third*) are instead treated as adjectives both morphologically and syntactically.

(10)

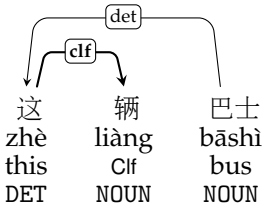


A classifier is a word that accompanies a noun in certain grammatical contexts. The prototypical case is that of numeral classifiers, where the word is used with a numeral for counting objects and where the numeral normally cannot occur without the classifier. A classifier generally reflects some kind of conceptual classification of nouns, based principally on features of their referents. UD uses the `clf` relation to connect the classifier to the numeral (or determiner) together with which it modifies the noun, as in Example (11) from Chinese.

(11) a.



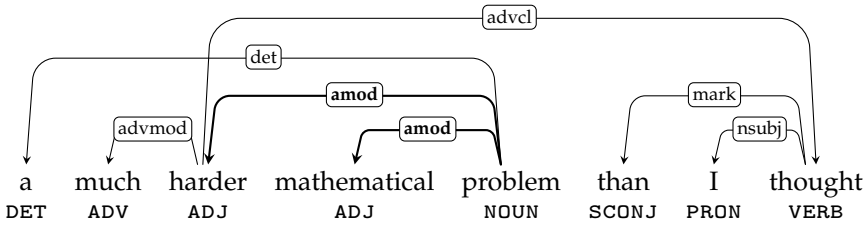
'three students'



b. 'this bus'

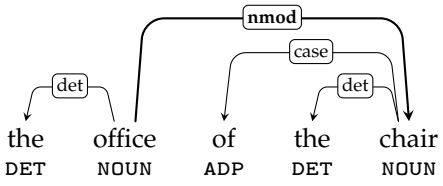
The morphological analysis of classifiers is debated. Etymologically, classifiers are normally nouns, and UD generally recommends using the NOUN tag. It has been suggested that a special feature should be added to distinguish the classifier use, since the words can normally also be used as regular nouns, but there is currently no such feature.

3.1.3 Adjectival and Nominal Modifiers. Adjectives modifying the head of a nominal are linked to the head noun with the amod relation. Unlike case markers, determiners, numerals, and classifiers, adjectives can be freely multiplied and can themselves be the head of complex constructions involving modifiers of various kinds, as illustrated in Example (12). A special case of adjectival modifiers are ordinal numerals (as in *the second Harry Potter book*), which are analyzed as adjectives in UD.

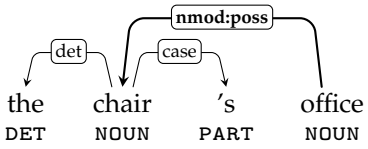


(12)

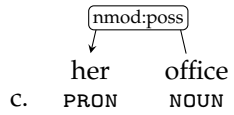
The head of a nominal may also be modified by another nominal, whose head is then linked to the higher noun with the nmod relation as in Example (13a). A special case is the genitive construction as in Example (13b), which may occur with or without overt case markers. Possessive pronouns, when used to modify nouns, are treated as a special case of nominal modifiers. In many treebanks, the subtype nmod:poss is used both for possessive pronouns Example (13c) and full genitive noun phrases Example (13b). However, if the grammatical rules of the language treat the possessive word analogously to determiners (i.e., the possessive is not a nominal), det(:poss) is used as in the Croatian Example (14).



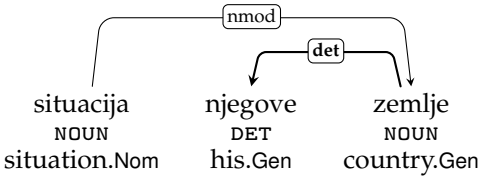
(13) a.



b.

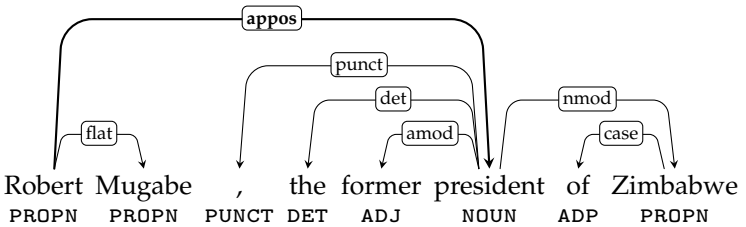


c.



(14) 'the situation of his country'

A special type of nominal modification, recognizable in some languages, is apposition, for which UD has a special appos relation. It connects two nominals that have the same (or overlapping) referents, as exemplified in Example (15). According to the UD criteria, the two nominals involved in an appositive construction are syntactically independent, can often be reordered, and are usually separated by a comma in writing. The appos relation is also strictly left-to-right, meaning that the first nominal is always treated as the head. This is a more narrow-scoped definition than the notion of apposition found in some grammars, which may also include modifiers that precede the head or that are not themselves syntactically independent nominals.



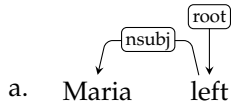
(15)

There are a number of additional structures that may appear in nominals, including compounding and flat structures (see Section 3.4.1 and 3.4.3, respectively), and clausal modifiers, particularly in relative clauses (see Section 3.3.2).

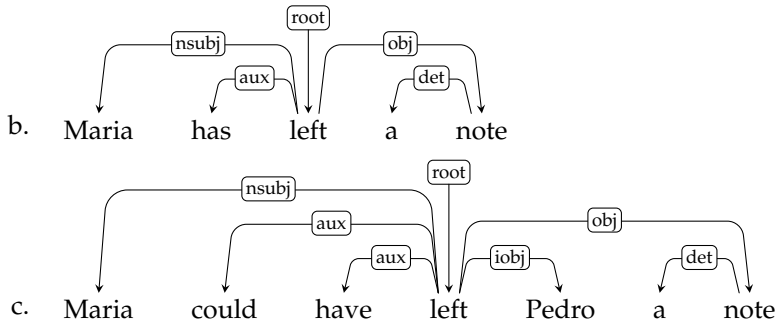
3.2 Clauses

A clause consists of a predicate together with its core arguments and oblique modifiers. In this section, we focus on *simple* clauses where dependents of the predicate are nominals, adverbs, or function words. Complex clauses, where a subordinate clause acts as a core or oblique dependent, are discussed in Section 3.3.

3.2.1 Predicates and Core Arguments. In most clauses, the main predicate is a verb, which can be intransitive, transitive, or (in some languages) ditransitive, as illustrated in Example (16a–16c).



(16) a.

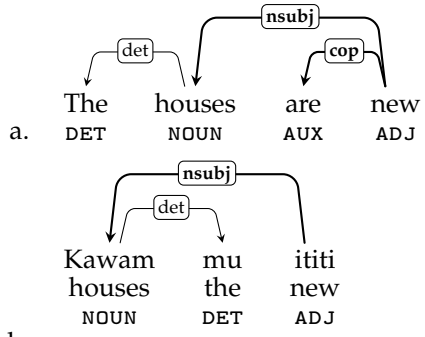


In Example (16a), the intransitive verb *left* has a single core argument, a nominal subject (`nsubj`). In Example (16b), the verb takes an additional core argument, a direct object (`obj`). In Example (16c), finally, there is a third core argument analyzed as an indirect object (`iobj`). In English, nominal core arguments are never introduced by prepositions. Therefore, in a sentence like *Maria could have left a note for Pedro*, the prepositional phrase *for Pedro* is analyzed as an oblique nominal dependent (`obl`) despite its near semantic equivalence to *Pedro* in Example (16c). We introduced the problem of identifying core arguments in Section 2.3.2 and will return to its crosslinguistic application in Section 4, after we have completed the overview of grammatical constructions in UD.

Examples (16b) and (16c) also illustrate that auxiliary verbs are treated as dependents of main verbs in UD with the `aux` relation. Auxiliary verbs help specify verbal features such as tense, aspect, modality, evidentiality, or voice. They may also carry agreement features that cross-reference the subject or other core arguments. The criteria for distinguishing auxiliaries are again language-specific, but auxiliaries are always a closed class and usually a small one. If there are multiple auxiliaries in one clause, a flat structure is created where all auxiliaries are attached directly to the main verb.

Another basic clause construction is that of nonverbal predication, where the main predicate is not a verb but a noun or adjective which usually takes a single core argument analyzed as a nominal subject (`nsubj`). This is a common construction in most languages, but languages differ in the strategies they use to realize the construction morphosyntactically. This is illustrated in the examples below, which show equivalent sentences in English Example (17a) and Waskia Example (17b).

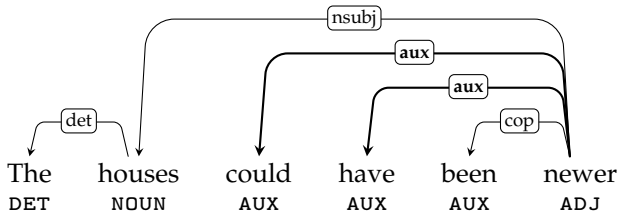
(17)



b. 'The houses are new'

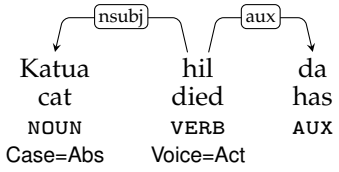
English uses a copula strategy, with a special verb linking the predicate to its subject, while Waskia uses a zero strategy, with no overt linker. By attaching the subject to the nonverbal predicate in both cases, UD highlights the similarity of the construction across languages with different realization strategies. The copula verb is attached to

the nonverbal predicate with the cop relation. Other auxiliaries are attached to the nonverbal predicate with the aux relation, as in Example (18).

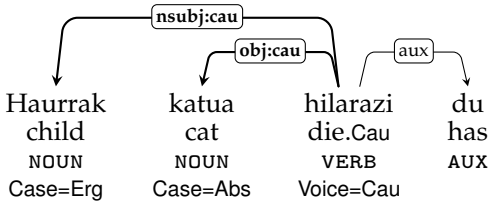
(18)


Many languages have ways of altering the mapping between the grammatical relations and the semantic roles of a verb. Such transformations involve changing the form of the verb (using morphology, auxiliaries, or both) as well as the encoding of its dependents. For example, in the **passive** construction the original object is promoted to subject, while the original subject either disappears or is demoted to an oblique modifier. The UD analysis acknowledges the new grammatical relations of dependents, and in this case labels them as subject and oblique, respectively. Nevertheless, to signal that the mapping from grammatical relations to semantic roles has changed, UD provides the subtype nsubj:pass for the passive subject (and the subtype obl:agent for the oblique modifier). In addition, a passive auxiliary will be labeled aux:pass and a morphologically inflected verb will carry the feature Voice=Pass. Examples of passive constructions can be found in Example (8) and Example (9).

While the passive removes a core argument, the **causative** construction instead adds a new core argument. In the Basque examples below (Oyharcabal 2003), the intransitive sentence Example (19a) is converted to the transitive Example (19b). Here the subtypes nsubj:cau and obj:cau are used to signal the extended valency frame.

(19) a.


'The cat died'

b.


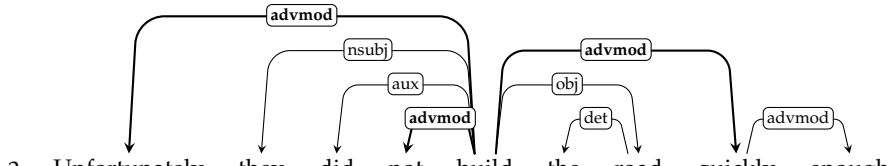
'The child caused the cat to die'

Other valency-changing operations are antipassive, applicative, and the symmetric voice in Western Austronesian languages. For broader typological considerations of voice, see Section 4.

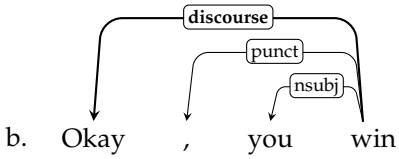
3.2.2 Oblique Modifiers. While predicates and their core arguments form the backbone of a clause, predicates can also be modified in a number of different ways. A large and relatively heterogeneous class of modifiers consists of adverbs, which modify either the predicate or the entire clause with respect to categories such as manner (*quickly*

in Example (20a)), polarity (*not* in Example (20a)), and speaker attitude (*unfortunately* in Example (20a)). All of these modifiers are attached to the main predicate with the `advmod` relation. For discourse particles and interjections, the `discourse` relation is used, as illustrated in Example (20b).

(20) a.

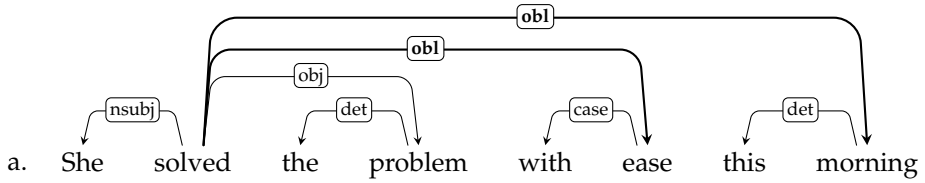


b.

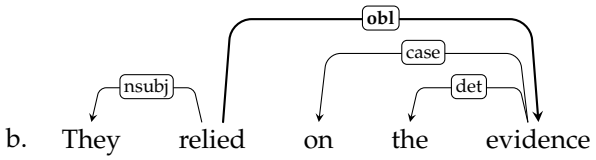


In addition to adverbs and discourse particles, oblique modifiers may also appear in the form of nominals. The `obl` relation is reserved for nominals that are dependents of clausal predicates[11] but do not satisfy the criteria for being core arguments. This includes not only nominals whose function is similar to adverbial modifiers, like *with ease* and *this morning* in Example (21a), but also nominals that are arguments semantically, like *on the evidence* in Example (21b). The criteria for distinguishing the latter type from core arguments is discussed in more detail in Section 4.

(21) a.



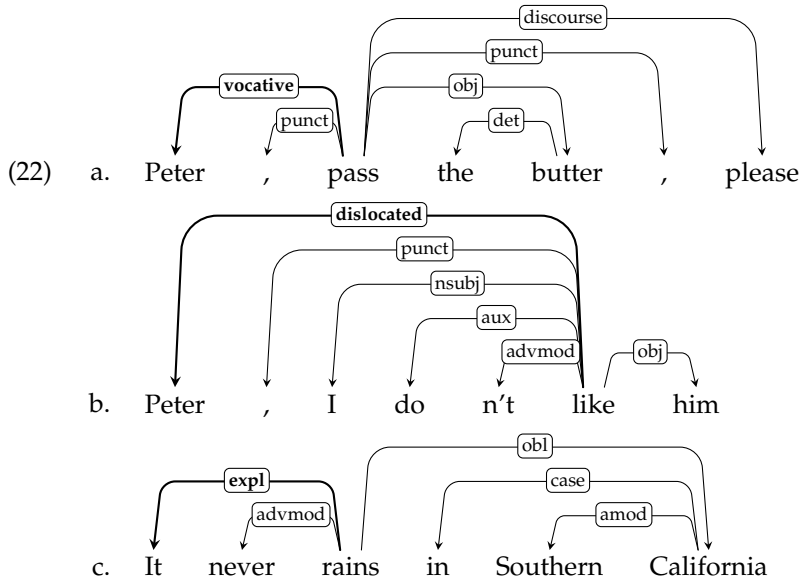
b.



The `obl` relation covers most non-core nominal dependents of predicates, but there are three special cases for which other relations are used, exemplified below. First, vocatives are nominals that are directed to an (imagined or real) addressee, as in Example (22a). They are attached to the main predicate with the `vocative` relation. Note that the vocative is not the subject of the imperative clause, even if it happens to refer to the actor of the event (and a vocative could equally well occur in a declarative sentence or a question). Second, dislocated nominals are nominals that occur peripherally (initially or finally) in a clause and that serve to contextualize or emphasize a participant of the clause. They do not fulfill a core argument role in the clause but often have discourse

11 To be precise, oblique nominals and adverbial modifiers are used for modification of non-nominals, including modification of adjectives and adverbs that are not clausal predicates. Because adjectives and adverbs normally act as modifiers and their own modification is possible but infrequent (Section 2.1.2), UD reuses the modifier phrase type and the relations `advmod` and `obl` rather than defining new relation types.

prominence, such as being a topic, and are usually anaphorically related with a core argument. The relationship is often coreference, such as in Example (22b), where the nominal *Peter* introduces a topical referent, which is then picked up anaphorically by the nominal object *him*, but there are also cases of bridging anaphora, such as the Japanese topic Example (62) in Section 4.3. Third, expletives are pronominal forms that occur in a core argument position but are not assigned any semantic role. A typical example is the dummy subject of a weather verb, which occurs in English and other languages that require the subject position to be filled in (non-imperative) clauses, as exemplified in Example (22c).[12]



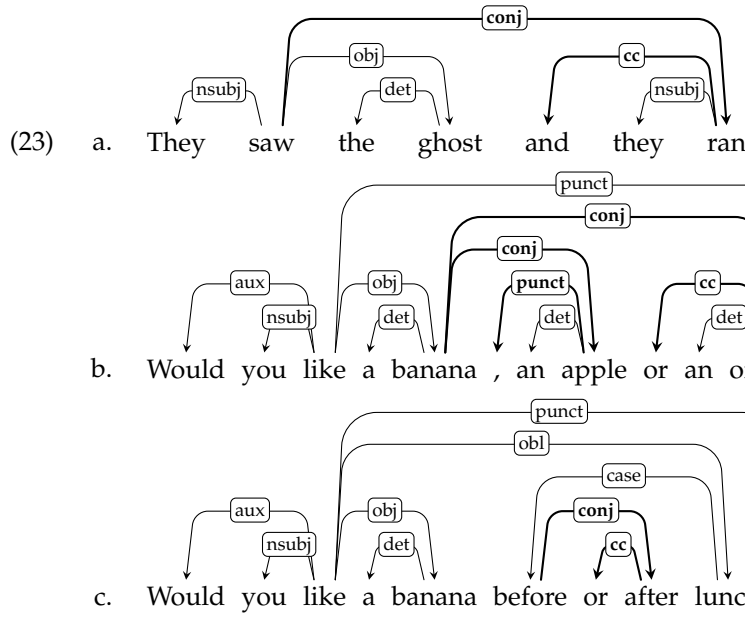
3.3 Complex Constructions

In this section, we describe a variety of linguistic structures, which have in common that they involve clauses embedded into larger structures through relations of coordination or subordination.[13] It will not be possible to survey this class of constructions exhaustively, so the emphasis is on illustrating the general principles underlying their treatment in UD.

3.3.1 Coordination. All cases of coordination, at the clause Example (23a), phrase Example (23b), or word Example (23c) level, receive the same analysis. UD in principle assumes a symmetric relation between conjuncts, which have equal status as syntactic heads of the coordinate structure. However, because the dependency tree format does not allow this analysis to be encoded directly, the first conjunct in the linear order is by convention always treated as the parent of all other conjuncts. Coordinating conjunctions and punctuation delimiting the conjuncts are attached to an adjacent conjunct using the `cc` and `punct` relations, respectively.

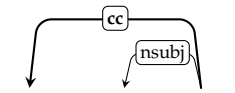
12 A detailed discussion of different expletives and their treatment in UD can be found in Bouma et al. (2018).

13 The only exception is phrase and word-level coordination, which is discussed together with clausal coordination in Section 3.3.1 for convenience.



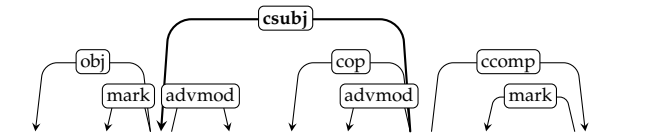
(23) a.
b.
c.

As pointed out by Gerdes and Kahane (2016), the attachment choice of the coordinating element to an adjacent conjunct is motivated by structural properties in many languages, because they together constitute a phrase. Furthermore, such an analysis can provide a parallel analysis for sentences introduced by a conjunct as in Example (24).

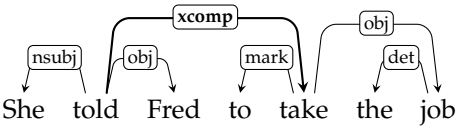


(24)

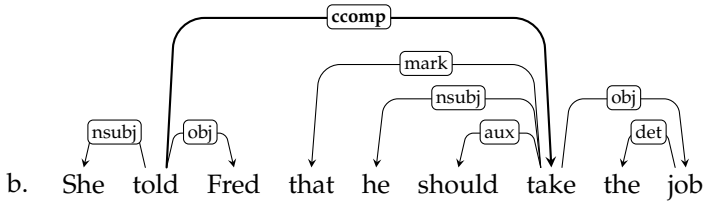
3.3.2 Subordination. UD distinguishes four types of subordinate clauses: clausal subjects (`csubj`) as in Example (25); clausal complements (objects), divided into those with obligatory subject control (`xcomp`) as in Example (26a) and those without (`ccomp`) as in Example (26b); adverbial clause modifiers (`advcl`) as in Example (27); and adnominal clause modifiers (`acl`), with relative clauses as an important subtype in many languages Example (28).



(25)



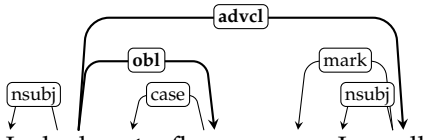
(26) a.



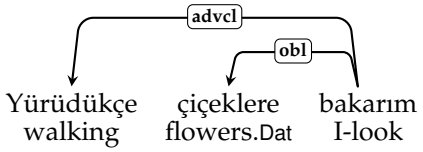
b.

Following the principle of prioritizing relations between content words, the head of a subordinate clause is its predicate, while markers of subordination (e.g., subordinating conjunctions), if any, are attached to the head of the clause they are in, with the relation `mark`. This leads to parallel analyses in English and in Turkish despite different strategies for expressing the subordinated clause: The adverbial clause in English Example (27a) is introduced by the subordinating marker *as* where Turkish uses the morphological marker *-çe* Example (27b).

(27) a.



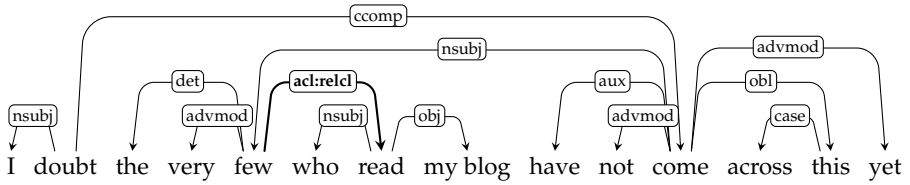
b.



'I look at flowers as I walk'

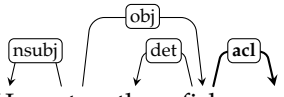
In the case of relative clauses as in Example (28), relative pronouns are attached to the head of the relative clause with the relation corresponding to their grammatical function in that clause (e.g., `nsubj`, `obj`, `obl`).

(28)

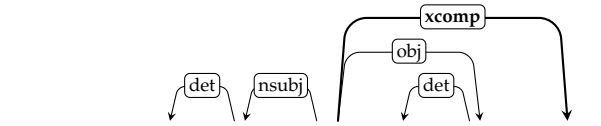


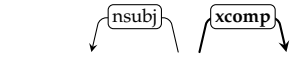
The `acl` relation is also used for optional depictives, such as Example (29), which are thus analyzed as reduced non-verbal clauses, modifying a nominal.

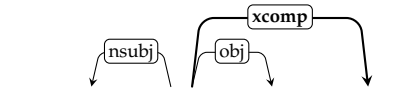
(29)

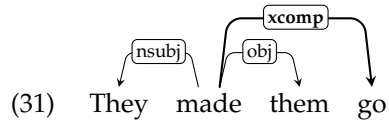


All other secondary predicates (see Huddleston and Pullum [2002] ch. 4), optional resultatives Example (30a), as well as obligatory depictives Example (30b) and obligatory resultatives Example (30c), are treated as core arguments, following Huddleston and Pullum (2002), and given an `xcomp` analysis. UD adopts the same analysis for small clauses, such as Example (31), which share properties of obligatory secondary predicates.

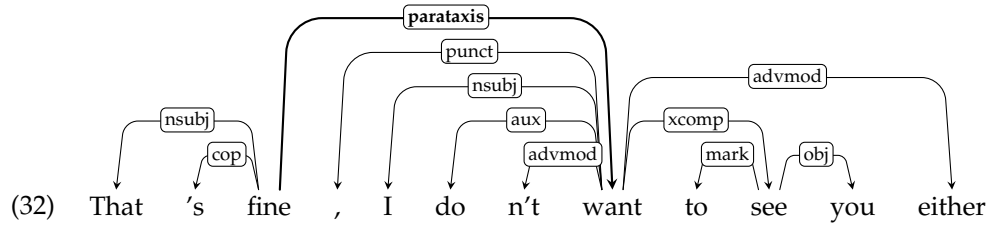




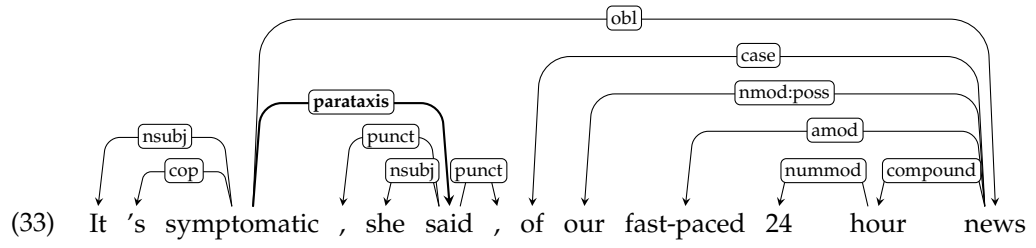


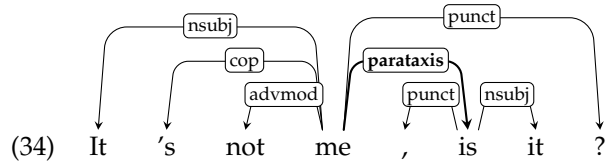


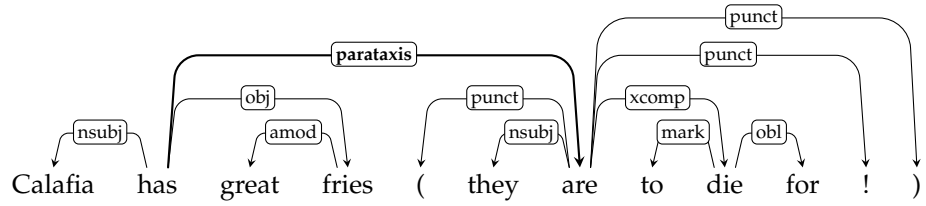
3.3.3 Parataxis. UD introduces the `parataxis` relation to capture clauses or other constituents placed side by side without any explicit coordination or subordination, as in Example (32). This subtype of parataxis can be viewed as a discourse-like equivalent of coordination—whether or not there is punctuation (comma, semi-colon, or colon)—and therefore we follow the same convention as coordination, with the first constituent being the parent.

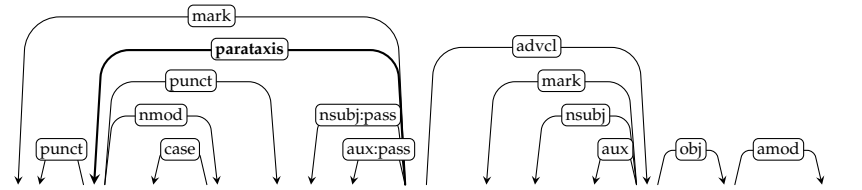


Some other constructions are also given a `parataxis` analysis: reported speech Example (33), tag questions Example (34), interjected clauses Example (35), or interjected constituents Example (36). In these cases, the added material is the `parataxis` dependent (and the parent does not necessarily occur before the child).





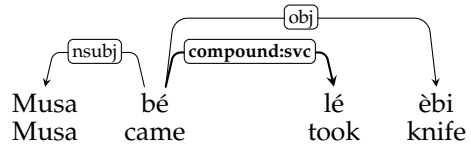




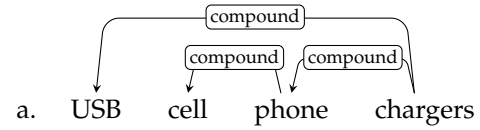
3.4 Multiword Expressions

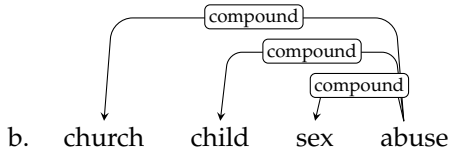
The most regular process of sentence construction in human languages is for a word to be able to take arguments and modifiers that themselves allow further expansion with their own modifiers. For example, *house* can take a modifier like *decrepit*, but that modifier can take its own modifiers and you can form an expression such as [*really rather decrepit*] *house*. However, languages also include constructions where multiple words form a compound or fixed expression. Under a lexicalist approach, such multi-lexeme units are fundamentally different from cases of phrasal modification. UD provides three relations to capture multiword expressions (MWEs), suggesting that these capture the main distinctive groups of MWEs.

3.4.1 Compound. The first, and best recognized, situation is compounding. The relation *compound* is used for any kind of word-level compounding: noun compounds (e.g., *phone book*), but also verb and adjective compounds, such as a Japanese light verb construction, such as *benkyō suru* 'to study', or the serial verbs that occur in many languages, such as this Nupe example:

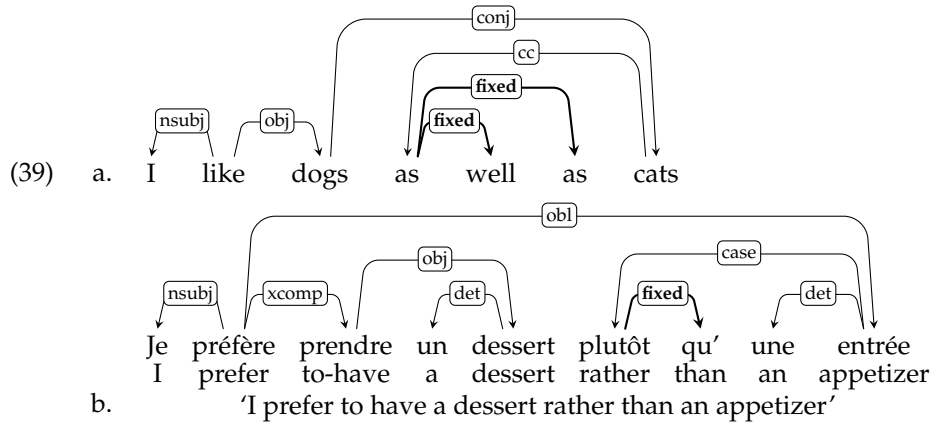


The compound relation is also used for phrasal verbs, such as *put up*: The adverb *up* is attached to *put* via `compound:prt`. Compounds are seen as regular headed constructions: The `compound` modification relationships indicate the structure of the compound, as shown in Example (38). This behavior distinguishes compounds from the other two types of MWEs.



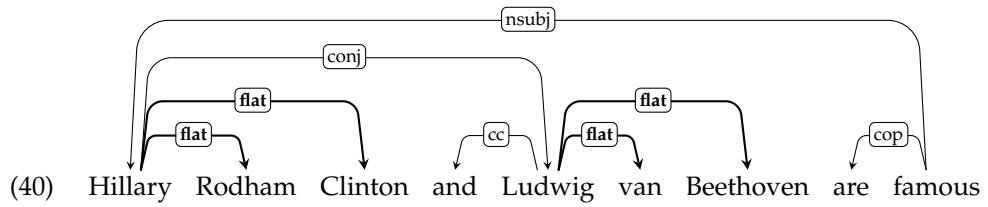


3.4.2 Fixed. The `fixed` relation is used for highly grammaticalized expressions that typically behave as function words or short adverbials. The name and rough scope of usage is borrowed from the fixed expressions category of Sag et al. (2002). Fixed MWEs are annotated with a flat structure. Because there is no clear basis for internal syntactic structure, we adopt the convention of always attaching subsequent words to the first one with the `fixed` label Example (39).



As with other clines of grammaticalization, it is not always clear where to draw the line between giving a regular syntactic analysis versus a fixed expression analysis of a conventionalized expression. In practice, the best solution is to be conservative and to prefer a regular syntactic analysis except when an expression is highly opaque and clearly does not have internal syntactic structure (except from a historical perspective).

3.4.3 Flat Multiword Expressions. The final class of MWEs is `flat`. This class is less clearly recognized in most grammars of human languages, but in practice there are many linguistic constructions with a sequence of words that do not have any clear synchronic grammatical structure but are not fixed expressions. These include names without internal syntactic structure, and calqued expressions from other languages. We again adopt the convention that in these cases subsequent words are attached to the first word with the `flat` relation, as in Example (40).

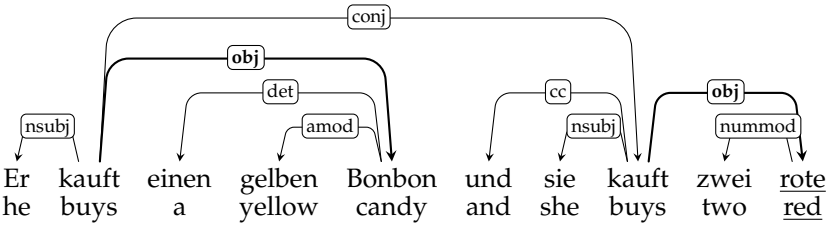


3.5 Ellipsis

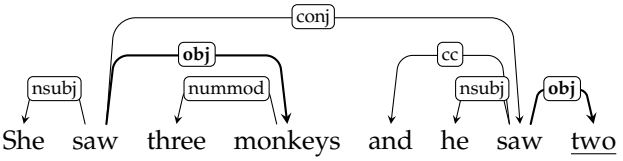
The analysis of ellipsis poses a challenge for all linguistic theories, especially those that do not make use of null nodes (or empty categories) to represent non-overt linguistic elements. UD adopts a compromise solution in this respect. The strategy for analyzing ellipsis is to preserve as many dependency relations as possible and resort to a special relation, which explicitly marks the ellipsis, only when absolutely necessary. The representation discussed here is restricted to overtly realized elements.[14] The strategy is realized as follows:

- If the elided element has no overt dependents, nothing is done.
- If the elided element has overt dependents, one of these is *promoted* to the role of the head.
- If the elided element is a predicate and the promoted element is one of its arguments or phrasal modifiers, the special `orphan` relation is used when attaching other non-functional dependents to the promoted head.

3.5.1 Ellipsis in Nominals. If a nominal head is elided, dependents are promoted as head in the following priority order: `amod` > `nummod` > `det` > `nmod` > `case`. In German Example (41a), the `amod` (*rote* 'red') of the elided noun (*Bonbon* 'candy') is promoted; in Example (41b), the `nummod` (*two*) is.

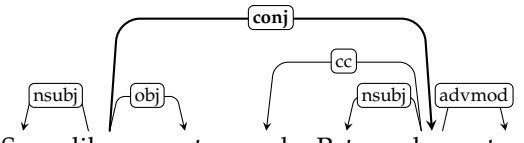


(41) a. 'He buys a yellow candy and she buys two red ones'



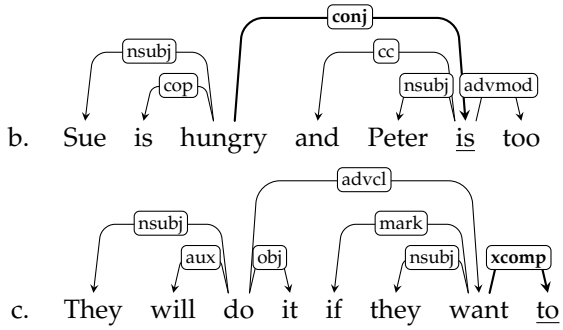
b.

3.5.2 Ellipsis in Clauses. If the main predicate of a clause is elided, the `aux`, `cop`, or a `mark` (in the case of an infinitival marker) dependents of the elided predicate are promoted, as illustrated in Example (42a), Example (42b), Example (42c), respectively.

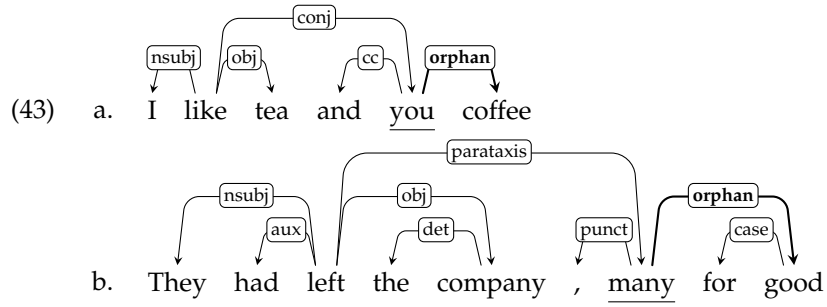


(42) a.

14 In some cases, null nodes are used in the *enhanced* representation to better capture the predicate–argument structure.



If there is no `aux` or `cop` to promote (or `mark` in the special case of infinitives), dependents are promoted in the following priority order: `nsubj` > `obj` > `iobj` > `obl` > `advmod` > `csubj` > `xcomp` > `ccomp` > `advcl` > `dislocated` > `vocative`. However, to avoid confusion and to signal that the dependency structure is incomplete, the special `orphan` relation is used to connect the non-promoted dependents to the promoted dependent, as exemplified in Example (43).



Note that the `orphan` relation is only used when an ordinary relation would be misleading (for example, when attaching an object to a subject). In particular, the ordinary `cc` relation should be used for the coordinating conjunction, which attaches to the pseudo-constituent formed through the orphan dependency, as shown in Example (43a) above, and similarly for the `punct` relation in Example (43b).

Using the `orphan` relation in cases of predicate ellipsis results in a severely underspecified predicate–argument representation but prevents the construction of a completely misleading dependency structure, where core argument and modifier relations are used to link words that are really co-dependents.

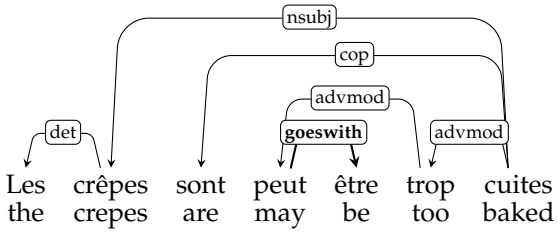
3.6 Miscellaneous Constructions Found in Corpora but Not Usually in Grammar Books

The application of the UD framework to naturally occurring data revealed the existence of several highly frequent constructions that are not discussed in comprehensive grammars. We give examples here, and the analysis proposed under the UD framework.

3.6.1 Special Relations for Informal Genres. Contrary to edited texts, text coming from informal genres, such as Web forums and social media data, and from speech transcripts often contain words wrongly broken into multiple tokens. Examples are given in Example (44a) where the French word for *maybe* is spelled over two tokens but should be one (*peut-être*), and in Example (44b) where the English word *nevertheless* is split into three

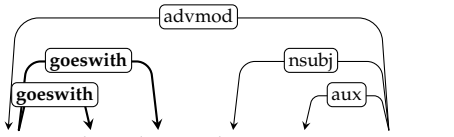
tokens. UD does not assume that a tokenization or normalization process can fix all these errors, and therefore provides a relation, `goeswith`, to indicate that these tokens should be seen as one word. Analogously to the `fixed` and `flat` relations, we adopt the convention of always attaching subsequent tokens to the first one.

(44) a.



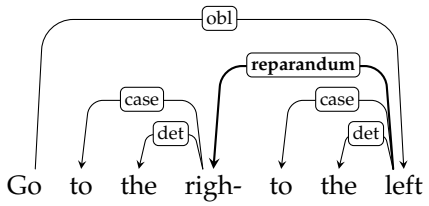
'The crepes may be overcooked'

b.



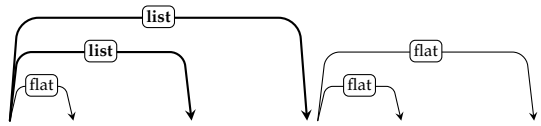
Similarly, transcripts contain speech repairs. UD uses the `reparandum` relation to indicate such disfluencies. The repair is chosen as the head because it constitutes the final utterance, with the disfluency being the dependent of the repair, as shown in Example (45).

(45)

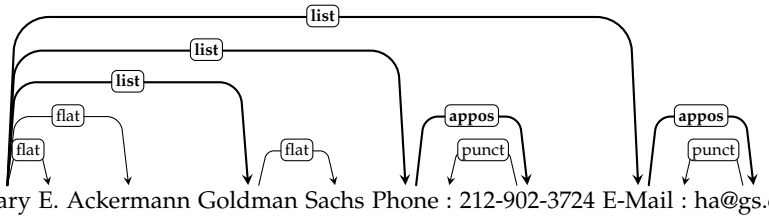


3.6.2 Lists. When dealing with Web data, we frequently encounter passages, parsed as single sentences, that are meant to be interpreted as lists. Email signatures are a typical example of such lists, as in Example (46a). UD uses the `list` relation to link the different elements, with the first one being the head. In some cases, the fields in the list are explicit, and take the form of a "key:value" structure. UD uses the `appos` relation to link a value to its key, as in Example (46b).

(46) a.



b.

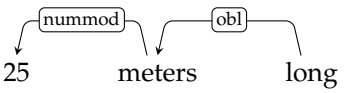


3.6.3 Noun + Number/Letter Constructions. Another frequent construction in all UD corpora is a noun followed by a number or a letter (or a combination of both), such as in the English examples in Example (47).

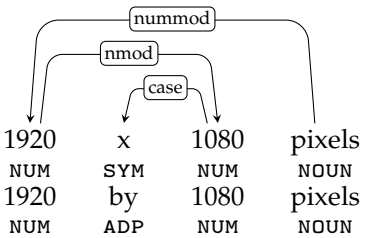
(47) a. This is the *number one* restaurant in town
b. He lives on *floor four*
c. *Bus 102L* takes you straight to the center
d. On *day 2* of our trip, we hiked to the bottom of the canyon
e. The meeting will be in *room A*

For a uniform treatment across such constructions, UD treats them as noun–noun constructions. While some of the examples above have an ordinal reading, such as Example (47b) or Example (47d), where the expressions can be paraphrased respectively as *on the fourth floor* and *on the second day*, UD analyzes the number as a noun to maximize the parallelism with constructions that use a letter or a combination of both number and letter; indeed, one can live *on floor C* where *C* acts as a noun. Therefore the number/letter expression attaches to the noun it modifies via a `nmod` relation, unless there is clear morphosyntactic evidence in the language for the opposite direction.

3.6.4 Measure Phrases. The analysis of simple measure phrases, such as *5 years old* or *25 meters long*, is relatively straightforward, illustrated in Example (48): The number serves to modify the meaning of the noun with a quantity and the measure noun is seen as functionally corresponding to an adverbial modifying the adjective.

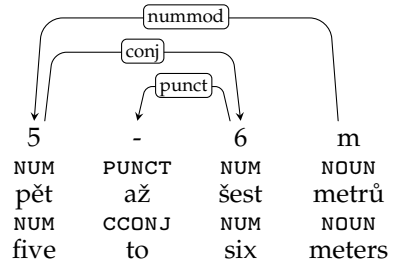
(48)


There are also complex measure phrases involving symbols, such as *1920 × 1080 pixels*, or ranges (*5 – 6 meters*). In such cases, the UD analysis follows the reading of the expression in the language. For instance, in the English Example (49), the symbol acts like a preposition *by* and is analyzed as such.

(49)


In some cases like in the Czech Example (50), the symbol is pronounced as a coordinating conjunction. It is thus analyzed as a punctuation PUNCT (see Section 3.3.1 on coordination) and the numerical constituent as a coordination.

(50)



'five to six meters'

4. Core Grammatical Relations: A Typological Perspective

One of the main challenges for a framework like UD is to ensure that universal categories are applied consistently across languages with sometimes radically different morphosyntactic encoding strategies. This can only be achieved through a complex interplay between abstract language-independent guidelines and concrete language-specific criteria. In this section, we will outline how this idea can be realized for core grammatical relations like subject and object, which play a central role in the UD theory. After stating general criteria derived from the typological literature, we will go through four groups of languages that illustrate different ways of instantiating the general criteria relative to language-specific evidence. The first group is what has been called Standard Average European (Whorf 1956; Haspelmath 2001), which is a homogeneous group but with some subtle differences, exemplified here by English, Czech, and Spanish. The second group is a selection of large non-Indo-European languages—Japanese, Arabic, and Swahili—which introduces more variety in the encoding of core grammatical relations. The third group comprises languages exhibiting different forms of ergativity, a phenomenon that is challenging for any theory based on the notions of subject and object. The fourth group includes languages with voice systems that are substantially different from the active–(middle)–passive that is found in the Indo-European family.

4.1 General Criteria

The starting point is that core arguments can be recognized relatively easily based on surface criteria such as word order, agreement, and case marking (both morphological and syntactic). However, for any given language, one has to first establish which of these criteria apply. For example, many languages have a morphological case called "dative," but dative nominals act as core arguments in some languages (or uses), and as oblique in other languages (or uses).

To determine which core arguments are available in a given language, and how they are morphosyntactically encoded, it is useful to start with so-called primary transitive clauses (Andrews 2007), that is, clauses with predicates that license the semantic roles of agent (actor) and patient (undergoer) in the prototypical sense. Clauses where the predicate is a verb describing a violent action are often good examples, such as *George killed the dragon*. In such a clause, the predicate has two core arguments: The more active argument (the agent) is said to have the grammatical function A; the other argument (the patient) is said to have the grammatical function P. By observing the coding strategies and grammatical rules that, within the language, are typical for arguments with the functions A and P, we can identify these functions also with other predicates,

regardless of their semantic roles. Such predicates will be called transitive and their arguments will also have the functions A and P, respectively. For instance, *John loves Mary* is a transitive clause, and *John* and *Mary* have the functions A and P, respectively, because the grammar treats them the same way as *George* and *the dragon* in the earlier example. The exact semantic roles are no longer important: John is an experiencer rather than actor, and Mary may not be affected by his love; she may not even be aware of it.

When we can recognize a predicate with two core arguments, we can also recognize predicates that have at most one (regardless of whether they also have additional non-core dependents). Clauses headed by such predicates are intransitive and their single core argument is said to have the grammatical function S. In general, nominals with functions S and A are subjects and labeled `nsubj`, while arguments with function P are objects and labeled `obj`. Finally, some verbs in some languages take three or more core arguments, more than one showing behavior that is characteristic of objects (Haspelmath 2015). Prototypically, such ditransitive constructions involve verbs of giving and transfer, and UD analyzes the theme (i.e., the entity that is transferred) as the direct object, and introduces the relation of indirect object (`iobj`) for the recipient. However, as noted earlier, the `iobj` relation should only be used if the nominal denoting the recipient is encoded as a core argument. In English, for example, this means that the nearly synonymous sentences *Mary gave John a book* and *Mary gave a book to John* differ in that the recipient is realized as an indirect object (*John*) in the former but as an oblique modifier (*to John*) in the latter.

We now discuss how these general principles can be applied to languages with different encoding strategies, starting with familiar Indo-European languages and gradually introducing more diversity.

4.2 Standard Average European

In Indo-European languages with case marking, nominative and accusative cases will usually map to subject and object core arguments, respectively. When there is no case marking, tests based on word order, pronominalization, and passivization can be used to identify core arguments.

English. In English, nominal core arguments are bare nominals (that is, without prepositions) and can be identified, to some extent, using word order. In an unmarked declarative sentence, the core argument preceding the verb is the subject. If there is another core argument following a transitive verb, it is the object, as in Example (16b). English has a remnant of morphological case for some of the personal pronouns: Subject pronouns are in the nominative form (*I, he, she, we, they*) whereas objects are in the accusative form (*me, him, her, us, them*).

The main complication when drawing the core–oblique distinction in English is that, while the presence of a preposition is a sufficient condition for obliqueness, it is not a necessary one. There are bare nominals that are used as oblique (temporal) modifiers, as in Example (51b).

(51) a. A baker works the dough
b. A baker works the whole week
c. John spends the whole week in Paris

The reasons why *the whole week* is not a core argument in Example (51b) (whereas *the dough* and *the whole week* are core arguments in Example (51a) and in Example (51c),

respectively) are complex, but we can use tests based on word order, pronominalization, and passivization to establish that *the whole week* does not behave as a core argument in Example (51b). An oblique modifier (*the whole week* in Example (52a) and Example (52b)) can swap positions with a locational modifier (e.g., *in Paris*), whereas this is not possible for a core argument Example (52c) vs. Example (52d):

(52) a. John works the whole week in Paris
b. John works in Paris the whole week
c. John spends the whole week in Paris
d. *John spends in Paris the whole week

Second, unlike a direct object, the temporal modifier cannot be pronominalized:

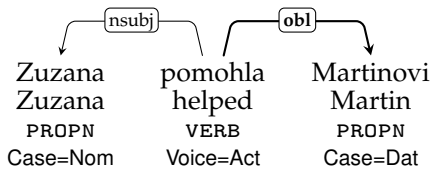
(53) a. *John worked it in Paris
b. John spent it in Paris

It is also not possible to promote a temporal modifier to subject by passivization: (**The whole week was worked by John*). This test is not decisive by itself in English, as there are transitive verbs that cannot passivize, and prepositional verbs that can. But taken all together, the tests indicate that *the whole week* in Example (51b) is not an object, and it will therefore attach to the verb with the `obl` relation.

Czech. Czech has substantive morphology that can be used to classify verbal arguments. Core arguments in Czech are bare noun phrases in the nominative for the subject and in the accusative for the object. Whereas SVO order is preferred by default, Czech word order is free: Other permutations are possible and may be required to distinguish topic and focus. Like in English, a bare accusative nominal is not necessarily a core argument. It can be an oblique (temporal) modifier, as *celý téden* 'whole week' in Example (54a) or *každou středu* 'every Wednesday' in Example (54b).

(54) a. Pracuje celý týden
works whole week
'He/she works the whole week'
b. Přichází každou středu
comes every Wednesday
'He/she comes every Wednesday'

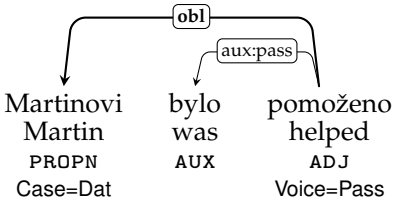
Many verbs in Czech take, in addition to a subject, a bare noun phrase in a case other than accusative (i.e., in the dative, genitive, or instrumental). UD invariably treats these as oblique (`obl`), as in Example (55).



(55) 'Zuzana helped Martin'

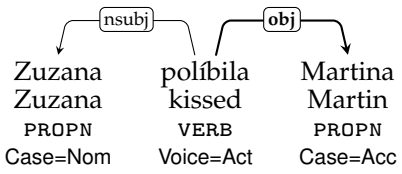
Whether these non-accusative second dependents should be seen as core arguments or not is debatable.[15] There are examples of verbs that take non-accusative second dependents and could be claimed to belong to transitive verbs (viz., *pomohla* 'helped' in Example (55)). However, such examples are rare, and non-nominative, non-accusative dependents tend to have semantic roles other than the proto-patient. Also, the treatment of these dependents by grammatical rules such as passivization is different from the treatment that accusatives receive. In Example (56a), which is the passive corresponding to Example (55), *Martinovi* is not promoted to subject: It stays in the dative case, and the passive predicate, instead of cross-referencing Martin's masculine gender, stays in the default neuter singular form. In contrast, the active sentence Example (56b) features an accusative argument *Martina*, and when passivized in Example (56c), this argument becomes the subject, taking the nominative form and triggering agreement both on the passive participle and on the auxiliary. Thus, UD treats only nominative and accusative dependents as core arguments in Czech.[16]

(56) a.



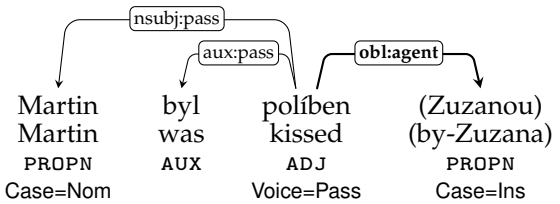
'Martin was helped'

b.



'Zuzana kissed Martin'

c.



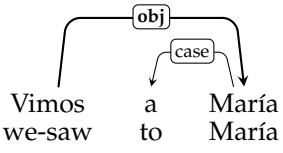
'Martin was kissed (by Zuzana)'

Spanish. Spanish is in many ways similar to English and Czech but does not adhere to the rule that the presence of a preposition is a sufficient condition for obliqueness. Spanish uses the preposition *a* with animate direct objects, as in Example (57a). Such objects, when pronominalized, use the accusative pronoun form Example (57b), and they can be promoted to subjects in passive constructions. Inanimate direct objects behave the same way except that they do not use the preposition. UD therefore treats a nominal with the preposition *a* as a core argument when it is an animate direct object.

15 For German, Andrews (2007, pp. 182–183) leaves the question open while Foley (2007, p. 377) has no doubt that the dative case is oblique.

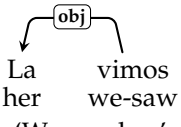
16 We ignore here certain anomalies in the Czech case system that involve quantified nominals. In the presence of a quantifier, the quantified noun may take the genitive form although the whole quantified phrase occupies a nominative or accusative position.

(57) a.



'We saw María'

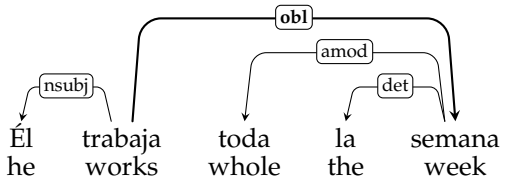
b.



'We saw her'

Similarly to English or Czech, a bare nominal is not necessarily a core argument, again with oblique temporal modifiers being a prime example, as in Example (58).

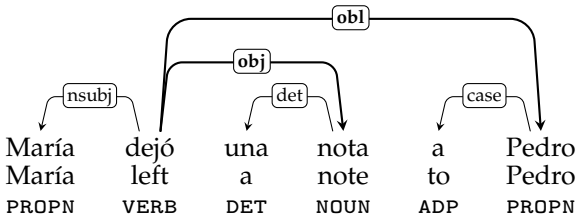
(58)



'He works the whole week'

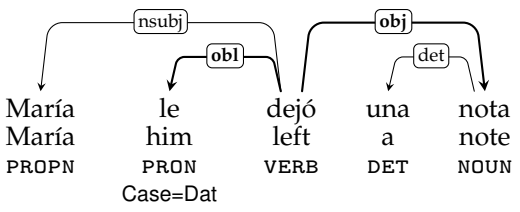
As mentioned in Section 3.2.1, some languages have two (or even more) object constructions, including Germanic and Bantu languages. For instance, in ditransitive constructions, the predicate has `obj` and `iobj` dependents (see Example (16c) for the English example *Maria could have left Pedro a note*). Traditionally, Romance languages have been viewed as lacking multiple object constructions, because there is always at most one bare object nominal, and other nominals are expressed with adpositions (as Example (59a) in Spanish).

(59) a.



'María left a note for Pedro'

b.



'María left him a note'

Still the dative seems to have something of a special status. Part of the evidence is the availability of dative clitics, as in Example (59b) (though French also has partitive and locative clitics); other evidence comes from relation-changing operations like causatives. Some people have argued for Romance datives being core arguments (Van Peteghem

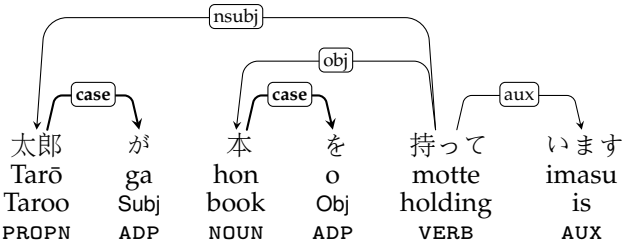
2006; Boneh and Nash 2012; Pineda 2013, inter alia) though others have argued against it (Kayne 1984, inter alia).

4.3 A Sample of Non-Indo-European Languages

In this section, we extend our discussion of core arguments in UD to three unrelated non-Indo-European languages, each with a large number of speakers: Japanese, Arabic, and Swahili.

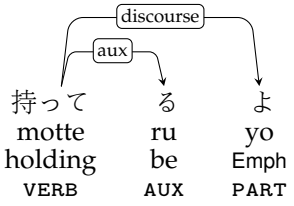
Japanese. In Japanese, while there is a predominant word order, there is considerable word order flexibility and nominal arguments can be freely omitted Example (60b). Grammatical relations are mainly expressed by case particles, which we regard as adpositions bearing the grammatical relation `case` Example (60a).

(60) a.



'Taroo has the book'

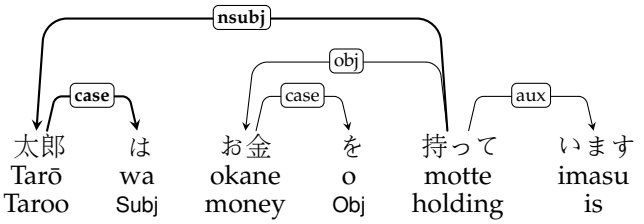
b.



'(I've) got (it)!'

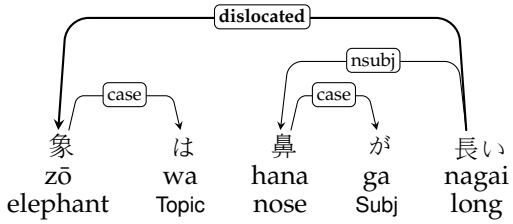
Japanese, like other East Asian languages, is a strongly topic-oriented language. Topics are marked with the case adposition は. Most commonly the topic-marked nominal will be the subject or another regular dependent of the clause, and は will then either replace (for `nsubj` or `obj`) or augment (for oblique dependents) the normal case adposition.

(61)



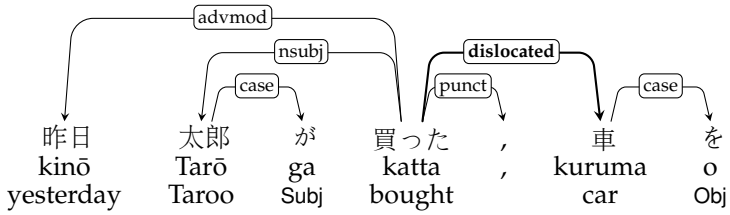
'Taroo has money'

However, a topic may also represent the context of the remainder of the sentence while not being part of the predicate-argument structure. A nominal that establishes a discourse context in this way takes the relation `dislocated`:



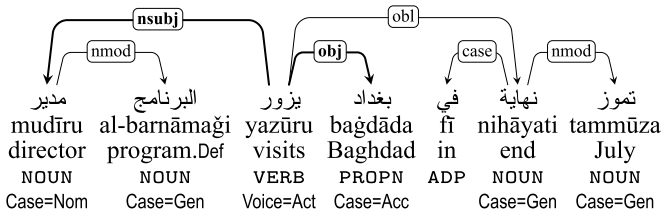
(62) 'The elephant's nose is long'

Although basically a head-final language, in spoken Japanese, nominal dependents and nominal dependents of dependents can also sometimes appear after the verb, as a kind of afterthought. These are also treated as dislocated elements:

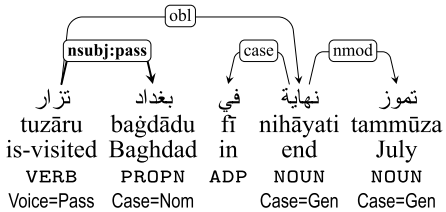


(63) 'Taroo bought it yesterday, a car'

Arabic. Arabic verbs cross-reference the person, number, and gender of their subjects. Nominals are case-marked: The subject is in the nominative, the object in the accusative (except in subordinate clauses with conjunction *ʾanna* 'that' Example (65a), where the subject is also in the accusative). Multiple word orders are possible, subject–verb–object and verb–subject–object being the most frequent. Passive clauses are agentless in Classical Arabic (Fischer 1997, p. 210) but oblique agent phrases are re-introduced in Modern Standard Arabic (Badawi, Carter, and Gully 2013, p. 385). The vowel pattern *a-ū* of the active verb in Example (64a) is replaced by the passive pattern *u-ā* in Example (64b). Furthermore, the masculine prefix *y-* is replaced by feminine *t-* to reflect the gender of the passive subject *baġdādu*.



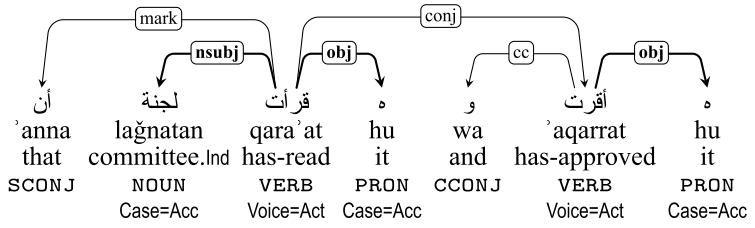
(64) a. 'The director of the program will visit Baghdad at the end of July'



b. 'Baghdad will be visited at the end of July'

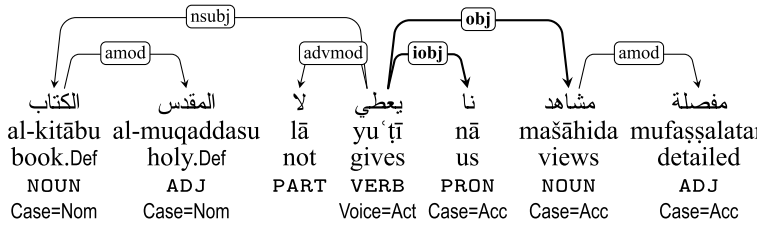
Subject pronouns can be dropped. Object pronouns are encliticized to the verb Example (65a) but treated as syntactic words in UD. In ditransitive clauses Example (65b), the verb governs two accusative objects; the recipient precedes the theme and, if pronominal, it is encliticized to the verb. Bare accusative nominals are not always core arguments; they can be adjuncts—for example *al-ʾusbūʿa al-māḍiya* 'last week' in Example (65c). Such 'adverbial accusatives' can denote time, location, direction, motivation, manner, and so forth (Fischer 1997, p. 216).

(65) a.



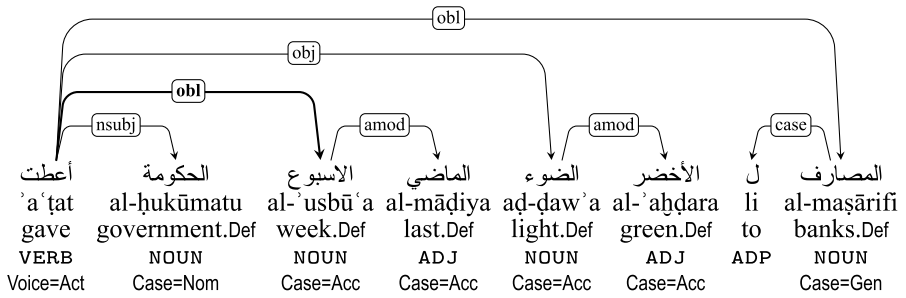
'… that a committee has read and approved it'

b.



'The Bible does not give us detailed views'

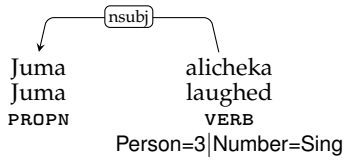
c.



'The government gave the green light to the banks last week'

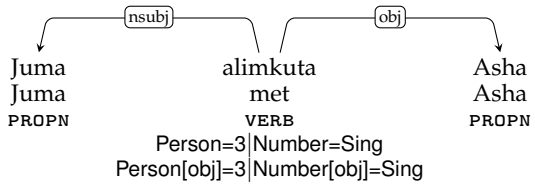
Swahili. In Swahili, core arguments are primarily marked by cross-referencing on the verb. There is no case marking and word order is relatively free, although subjects tend to precede and objects tend to follow the verb. Cross-referencing of the subject is obligatory, as illustrated in Example (66a–66e), where the prefix *a-* consistently marks the subject as third person singular. In transitive clauses, cross-referencing of the direct object is obligatory if it is animate, as in Example (66b) where the prefix *m-* marks the object as third person singular, but optional if it is inanimate, as in Example (66c). In ditransitive clauses, it is the object highest in animacy that is cross-referenced regardless of grammatical relation. Ditransitive clauses may be formed by an inherently ditransitive verb, as in Example (66d), or by an applicative transformation on a transitive verb, as in Example (66e), where the applicative suffix *-i* extends the valency frame of the verb *pik* 'cook' with an additional (indirect) object. The fact that the additional dependent is cross-referenced on the verb like any animate object supports its status as a third core argument.

(66) a.



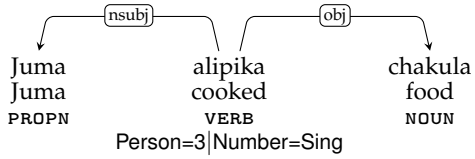
'Juma laughed'

b.



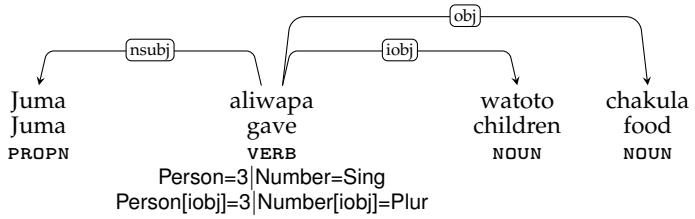
'Juma met Asha'

c.



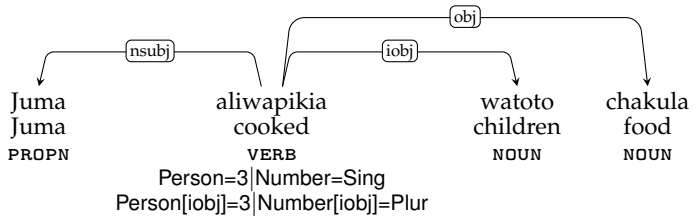
'Juma cooked food'

d.



'Juma gave children food'

e.



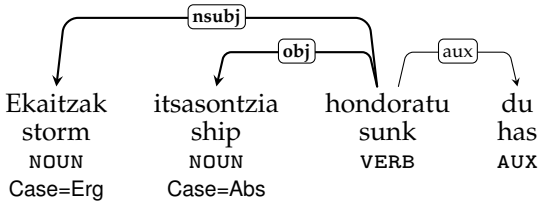
'Juma cooked food for children'

4.4 Ergativity

As discussed in Section 4.1, UD generally assumes that the `nsubj` relation covers the grammatical functions S and A, while `obj` is reserved for the grammatical function P. This fits well with the nominative–accusative alignment found in many languages, but it is challenged by the ergative–absolutive alignment that groups S and P together. For many languages, ergative–absolutive case marking appears to be only a morphological feature, which we handle at the level of the `Case` feature. Basque, below, is an example. For other languages, ergativity has been argued to extend to the treatment of grammatical relations (Dixon 1994). There are then multiple possible analyses (and different ones may apply to different languages). One choice is to regard the ergative as an oblique (Mel'čuk 1988), essentially analyzing all sentences in the language as intransitive, with only one core argument marked in the absolutive, which is used for intransitive arguments and the patient-like argument of transitive verbs. A more frequent analysis is to

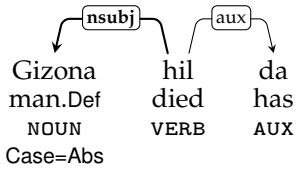
say that such syntactically ergative languages treat the intransitive core argument and the patient-like argument of transitives together as a "pivot" (Dixon 1994), which we would analyze as a subject (nsubj), and then the agent-like argument of transitives is also a core argument, which we would analyze as an object (obj). The unusual thing, then, is the reversed alignment between semantic roles and grammatical relations. This is a place where the relation subtype :pass can be usefully used in an extended sense. If we regard it as marking not only passives but all cases where the nsubj does not mark the agent-like argument of the verb, then all transitive subjects in such a language are nsubj:pass. In addition, we can reuse the subtype :agent, which in other languages is optionally used for an oblique modifier denoting a demoted agent, to mark the ergative core argument as obj:agent.

Basque. In Basque (Zúñiga and Fernández 2019), nominal case morphology is the main indicator of core argument relations. However, instead of nominative–accusative, the core pair of cases is ergative–absolutive. Most two-argument verbs have the more agentive argument in the ergative and the patient-like argument in the absolutive case, while single argument verbs usually use the absolutive for their single argument. Nevertheless, there is no evidence that absolutives form a coherent grammatical relation. Rather, the ergative argument is treated as subject (nsubj), while the absolutive argument of transitives is object (obj), as in Example (67).

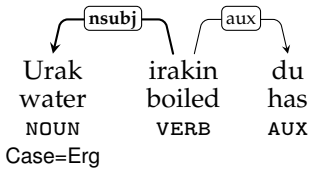


(67) 'The storm has sunk the ship'

The single argument of intransitive verbs takes mostly the absolutive Example (68) but sometimes the ergative form Example (69). It is labeled as subject (nsubj) in both cases.

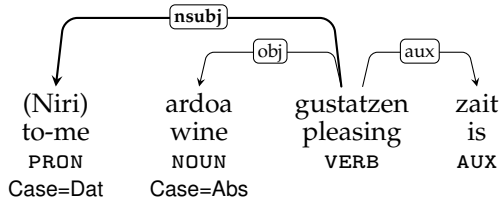


(68) 'The man has died'



(69) 'The water has boiled'

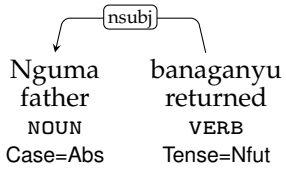
The third core argument case is the dative. Arguments in all three core cases are cross-referenced on finite verbs and can be omitted. Some experiencer-subject two-argument verbs take dative + absolutive, instead of ergative + absolutive, as in Example (70).



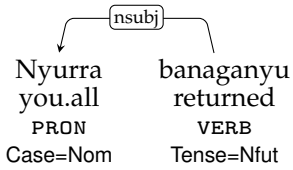
(70) 'I like wine'

According to Zúñiga and Fernández (2019), the dative encodes the A function in such constructions, which makes it subject in UD. Supporting evidence for this is provided by causativization, a valency-changing operation that takes a transitive clause, adds a third, ergative argument, and switches the original subject to the dative (unless it already was in dative). The fact that causativization is available for dative–absolutive clauses supports our treatment of the dative argument as the subject.

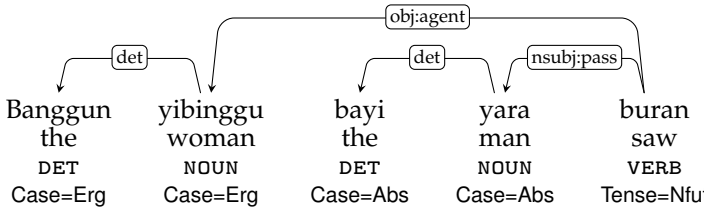
Jirrbal. Jirrbal or Dyirbal (Pama-Nyungan, Australia) (Dixon 1972, 1994) is a famous case of a language that has been argued to have transitive clauses with an S and P pivot. It has a combination of ergative–absolutive case marking on nouns (similar to Basque), as in Example (71a) and Example (72a), and nominative–accusative case marking on pronouns, as in Example (71b) and Example (72b), a common pattern of split ergative case marking. In both cases, in transitive clauses, we treat the P pivot core argument as the `nsubj` and the A core argument as an `obj`, but we mark them for unusual semantic role alignment with `nsubj:pass` and `obj:agent`, respectively.



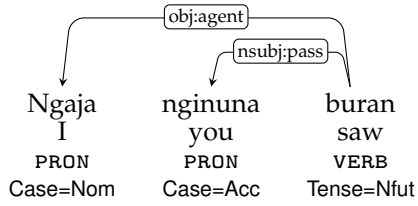
(71) a. 'Father returned'



b. 'You all returned'



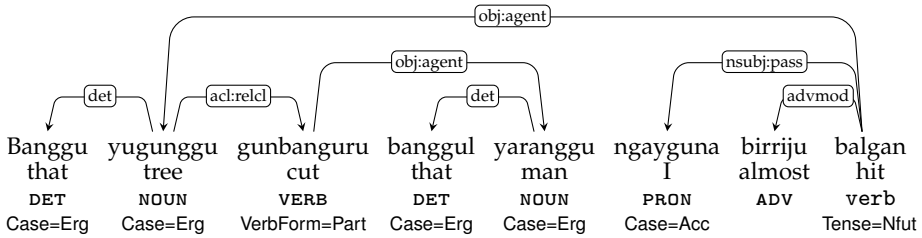
(72) a. 'The woman saw the man'



b. 'I saw you'

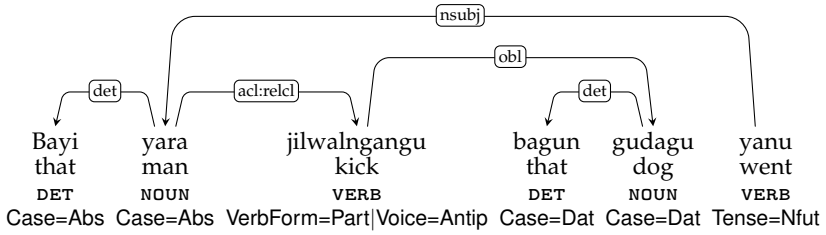
There are several grammatical processes, such as relativization, which, when restricted in application in a language, frequently only apply to subjects. The motivation for the above analysis is that in Jirrbal these processes apply to the S and P core arguments. For instance, the role of the head noun in a relative clause must be S or P, allowing relative clauses like Example (73a) where the relativized role is P, but not a relative clause where the relativized role is A. To express such an idea, the relative clause must be antipassivized, making the previous P into an oblique and the previous A into an S pivot, as in Example (73b).

(73) a.



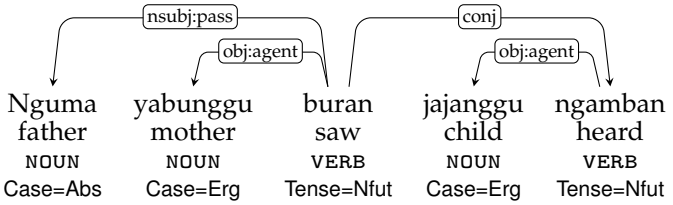
'The tree which the man had cut nearly fell on me'

b.



'The man who kicked the dog went'

As another example, the shared arguments in coordinated clauses must be S or P pivot core arguments, allowing the normally unexpected coordination in Example (74) but not allowing 'Mother saw father and heard the child' with a shared argument, except by antipassivization of the second clause. Again, this is most naturally handled by recognizing an S/P pivot which is analyzed as the `nsubj` in UD.



(74) 'Mother saw father and the child heard (him)'

4.5 Other Voice Systems

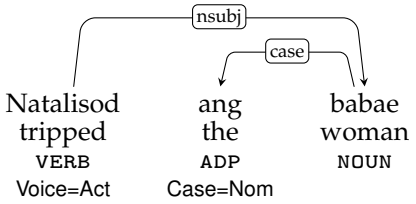
In European languages, the contrast between the active and passive voice is an important factor in categorizing simple clauses and their arguments. Ergative languages sometimes have an analogous contrast between the active and the antipassive. Yet there are languages whose voice systems do not seem to fit easily into either of these patterns. In this subsection, we first look at Tagalog, a representative of the Philippine-type languages, which are sometimes subsumed in a larger group of **symmetrical voice languages** (Himmelmann 2005). Then, we will discuss the direct–inverse voice system of Algonquian languages, exemplified by Plains Cree.

Tagalog. The arguments in Tagalog are marked by function words that could be analyzed as either prepositions, or case-bearing determiners; the former analysis is adopted here.[17] Although adpositions are often associated with oblique arguments and adjuncts, we have seen that it is not a universal rule. Spanish marks an animate direct object with the preposition *a*, and in Japanese all arguments are marked by postpositions, including the subject and the direct object.

The most subject-like argument (again called the pivot) is marked by the preposition *ang*. Other core arguments (if any) are marked by the preposition *ng* (Kroeger 1993, pp. 40–47). A different set of prepositions is used with proper nouns. Personal pronouns are not used with prepositional markers but inflect for case. Verbs are marked with infixed voice markers.

There is disagreement about whether the pivot is a subject and whether Tagalog has a subject at all. Andrews (2007, pp. 210–211) distinguishes two grammatical relations, the a-subject and the p-subject, each having some properties that are often associated with subjects in European languages. He also says that the actor "has subject-like properties regardless of whether or not it is the pivot. "For the purpose of easy and consistent annotation of UD, it is advantageous to follow the analysis of Manning (1996) and to always reserve the `nsubj` relation for the *ang*-phrase (the pivot), as in Example (75a). In the transitive sentences in Example (75b–c), different voices give different alignments of semantic roles to grammatical relations. We mark prepositions and personal pronouns with the `Case` feature: the pivot with nominative, and the other core argument(s) with genitive.[18]

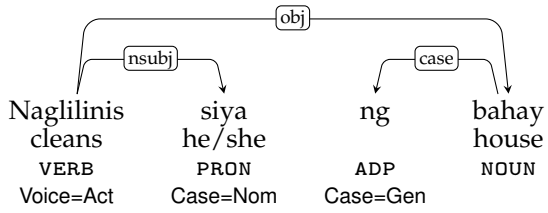
(75) a.



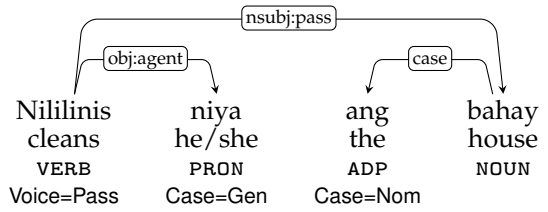
'The woman tripped'

17 There is no standard terminology for these words in the literature. Some authors classify them as prepositions (e.g., Schachter and Shopen 2007, p. 35), some as articles or determiners (e.g., Dryer 2007, pp. 94–95 and 121–122), and many authors avoid either of the terms and use the term "markers" instead (e.g., Andrews 2007, p. 203).

18 The names for the cases are not without controversy either. If the subject is nominative, the other core argument could be expected to be accusative, but due to its other functions, Tagalog *ng* is often glossed as genitive (Himmelmann 2005). The nominative–accusative analysis has been advocated by some authors (e.g., Guilfoyle, Hung, and Travis 1992), while others prefer to analyze Tagalog as an ergative–absolutive language (e.g., Payne 1982; De Guzman 1988; Gerdts 1988), which would mean that the pivot is in the absolutive and the *ng*-phrase in the ergative.



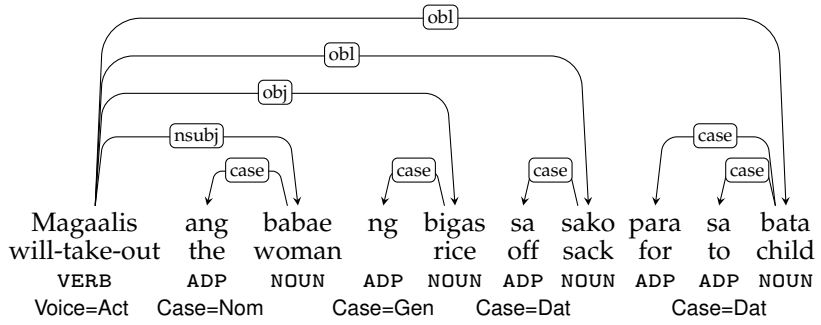
b. 'She cleans a/the house'



c. 'She cleans the house'

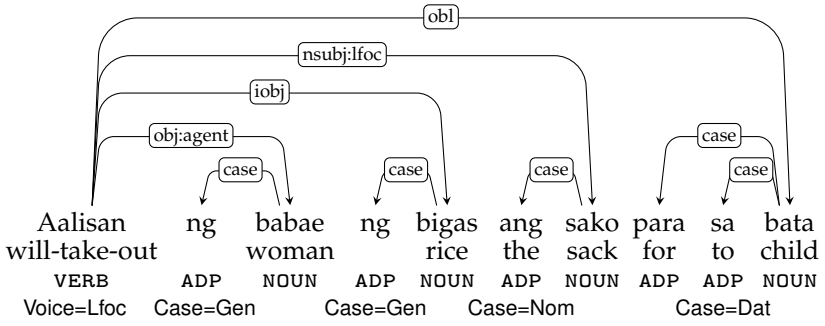
Despite the fact that we conveniently reuse the active and passive voice labels, it has to be understood that this alternation is significantly different from the active–passive alternation in English. Both clauses are transitive, as the non-subject argument stays core; in an English passive clause, the actor would be demoted to an oblique dependent. The construction in Example (75c) is neither less frequent nor morphosyntactically more complex than Example (75b). That is why the Austronesian voice system has been described as symmetrical; rather than "active" and "passive," the voice labels should be read as "agent/actor-focus" and "patient/theme-focus," respectively.

Locative, directional, and benefactive nominals are normally coded as oblique (e.g., the dative *sa sako* 'from sack' in Example (76a)). However, there are additional voices where these nominals become subjects, such as the location-focus voice in Example (76b) and the beneficiary-focus voice in Example (76c). One of the reasons why a dependent is promoted to the subject is that the subject is understood as the topic of the sentence.[19]

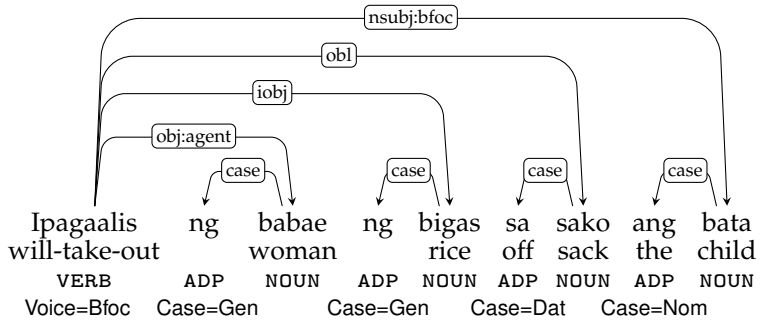


(76) a. 'The woman will take some rice out of a/the sack for a/the child'

19 The "focus" in the names of the voices indicates that the verb "focuses" on a particular semantic role and it should not be confused with pragmatic focus, which is the opposite of "topic."

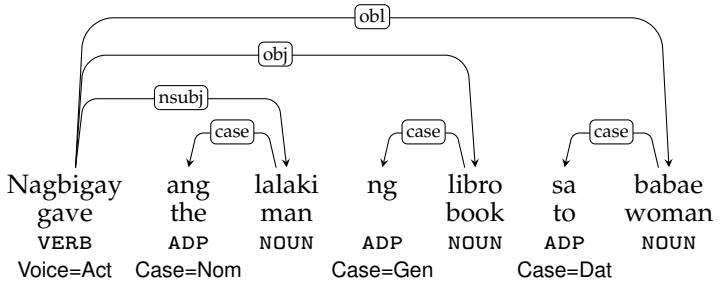


b. 'A/the woman will take some rice out of the sack for a/the child'



c. 'A/the woman will take some rice out of a/the sack for the child'

Because the agent and patient stay core arguments even in the locative and beneficiary voices, Example (76b) and Example (76c) are ditransitive clauses with three core arguments. In contrast, the verbs of giving, which are typical representatives of ditransitive predicates in other languages, form a standard transitive clause in the "active" and "passive" voices, with the recipient coded as a directional (dative) oblique dependent, as in Example (77).



(77) 'The man gave a book to the woman'

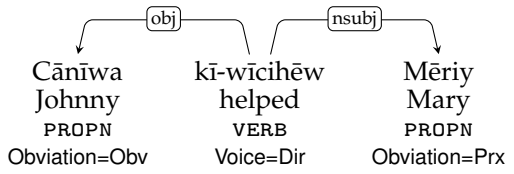
Plains Cree. The Algonquian (North American) language Plains Cree (Wolvengrey 2011) cross-references one or two core arguments by verbal inflection, which is sufficient to allow for a relatively free word order. As in many other languages where person and number of an argument is cross-referenced by the verb, the argument does not need to appear overtly. The distinguishing feature of the verb forms in Example (78) is voice: Example (78a) is in the direct voice (Dir), where higher arguments in the obliqueness hierarchy are taken to be more agent-like, whereas Example (78b) is in the inverse voice (Inv), where lower arguments are taken to be more agent-like. Given that first person arguments are higher than third person arguments, the agent is 'we' and the patient is 'they' in Example (78a), and inversely in Example (78b).

(78) a. Niwīcihānānak
1Plur[high]-help-Dir-3[low]-Plur[low]
'We help them'

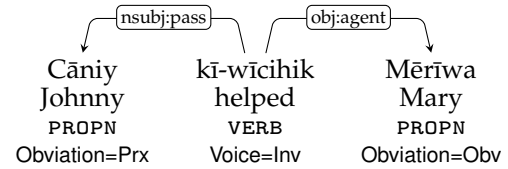
b. Niwīcihikonānak
1Plur[high]-help-Inv-3[low]-Plur[low]
'They help us'

Arguments cross-referenced by the verb are without doubt core arguments. It is not so obvious how to label the two arguments, as Plains Cree does not clearly have a subject in the Indo-European sense. It is one of a number of languages where evidence for differentiating core grammatical relations except via semantic role seems limited or non-existent. Nevertheless, it seems best to postulate that the argument higher in the obliqueness hierarchy should get the label `nsubj` in UD; the other core argument then gets `obj`. Such a distinction can be annotated easily and consistently. The subject will be more agent-like in the direct voice, and more patient-like in the inverse voice. This can be signaled by labeling non-agentive subjects as `nsubj:pass` without explicitly claiming that such sentences are passivized, unlike Dahlstrom (1991).

If two animate third-person arguments are involved, one of them is considered *proximate* (more topical, higher in the obliqueness hierarchy) and the other is considered *obviative* (less topical, lower in the obliqueness hierarchy). The obviative noun is marked morphologically by the suffix *-a*. We define a language-specific morphological feature, Obviation, with the values Prx and Obv, to represent this. In Example (79a), *Mēriy* is proximate, hence it is the subject, and it is also the agent because the verb is in the direct voice. In Example (79b), *Cāniy* is proximate and thus the subject; however, *Mēriy* is still the agent because the verb is in the inverse voice.

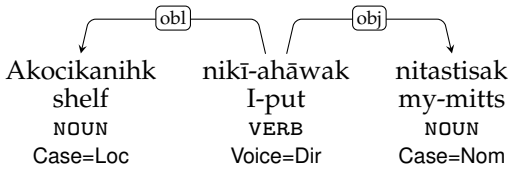


(79) a. 'Mary helped Johnny'



b. 'Mary helped Johnny / Johnny was helped by Mary'

Even though Plains Cree does not use morphological cases to distinguish agents from patients, nouns have a locative case (Case=Loc) that marks the noun as oblique and unable to be cross-referenced by verbal inflection.



(80) 'I put my mitts on the shelf'

While much work remains to be done in descriptive linguistics and its implementation in UD, we hope that this survey of typologically different languages has shown that UD provides a workable framework for the description and annotation of a broad range of clause-marking choices.

5. Design Principles of UD

There are many different ways that UD could have been designed. In this section, we briefly motivate and explain the design principles that guided us. Importantly, what UD seeks to achieve is rather different to what a grammar formalism in theoretical linguistics typically seeks to achieve, and thus the outcome is quite different.

The overarching goal of UD is a crosslinguistically consistent universal grammar that is suitable for use by the common person. That is, UD should be informed by our linguistic knowledge and the typology of language variation, but it should be something simple and interpretable enough that a psychologist, a software engineer, or a high school English teacher can comfortably use it. Behind this goal is a belief that there is something in common between human languages to be captured; as Bresnan et al. (2016, p. 1) argues, "there must be … a common organizing structure of all languages that underlies their superficial variations in modes of expression." From a linguistic point of view, such a common organizing structure is necessary for comparative linguistic studies and a substantive theory of crosslinguistic typology. From a practical NLP viewpoint, a common framework is needed to make it easy to build and maintain multilingual NLP systems, to allow effective crosslinguistic transfer learning, to enable meaningful crosslinguistic comparisons of parsing difficulty, and to approach the goal of a universal parser that works for all languages based on modern universal neural encodings of text (see, e.g., Kondratyuk and Straka 2019).

In choosing a common organizing structure for human language, UD applies a version of the Goldilocks principle: We should aim to maximize the commonality between languages but not to an extent that it obscures genuine differences between languages. Seeking commonality, it is a mistake if a parallel morphosyntactic notion is unnecessarily annotated inconsistently across different languages. Seeking fidelity, we avoid annotating things that are actually different (such as morphological vs. periphrastic expression of tense) as if they were the same. As a special case, UD eschews annotating things that are not there (empty items), because this is usually an artificial device to increase parallelism. Practically, we deal with quirky features of particular languages by insisting on use of a universal taxonomy of categories, features, and relations, but allowing the use of language-specific elaboration via subcategories. While the pressure in theoretical linguistics is for representations to become more and more detailed and complex over time, for UD, we realize that often less is better.

The secret to understanding the design and success of UD is to realize that the design is a very subtle compromise between a number of competing criteria:

1. UD needs to be reasonably satisfactory on linguistic analysis grounds for individual languages—a journeyman's universal grammar.
2. UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by non-linguist users with prosaic needs.

5. UD must be suitable for computer parsing with high accuracy.
6. UD must support well downstream language understanding tasks, such as relation extraction, reading comprehension, machine translation, and so on.

We observe that it is very easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part of developing UD has been working to improve the scheme and annotation guidelines while remaining sensitive to all these dimensions. Compare the analogy that school children are taught that English has eight parts of speech: Noun, Verb, Adjective, Adverb, Pronoun, Preposition, Interjection, Conjunction. This is not really true, but it has enough fidelity, enough simplicity, and enough comprehensibility to satisfy most people.

Many of the high-level design decisions of UD can be motivated in terms of these criteria. Making UD a monostratal theory—a theory with one representation (cf. Ladusaw 1988)—facilitates easy annotation and parsing. The emphasis on grammatical relations works well for both comparative linguistics and usage by non-linguists. Preferring relations between content words rather than mediated by function words increases crosslinguistic parallelism and within language parallelism (simple vs. periphrastic tenses become more parallel), and makes relation extraction easier (fewer, smaller patterns will cover a broader range of data). For example, the construction of predicating a property of a nominal (*the sky is blue*) is universal, while the strategy of achieving this via an auxiliary or copula verb is not. We increase parallelism by having a dependency between the nominal and the predicate. It also has the effect of more perspicuously revealing predicate–argument structure to the benefit of downstream processing. By mainly adopting terminology from traditional (European) grammar, we make it easier for non-expert users to comprehend UD representations, but we still make some changes, such as using the term **adposition**, to make UD more satisfactory on cross-linguistic grounds.

A key choice was between dependency representations and constituency representations (also known as phrase structure grammar, context-free grammar, or immediate constituency representations). One motivation here was simply the direction of the field of computational linguistics. While the famous early treebanks of modern empirical NLP, the Lancaster/IBM Treebank (Black, Garside, and Leech 1993) and the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993), and many treebanks that followed thereafter were constituency treebanks, by the early 2000s, there had been a huge shift to the use of dependency treebanks in computational linguistics. This was not altogether a new thing. David Hayes, a founder of the Association for Computational Linguistics, had strongly advocated for the use of Dependency Grammar in the 1960s (Hayes 1964). And it was not a random shift: The adoption of a dependency representation was driven by several of the ideas that underlie our design principles, such as simplicity, easy cross-linguistic applicability, interpretability by non-linguists, and usefulness for downstream applications.

Our goal was for UD to be a lightweight representation that is easy and satisfactory for people to work with. It is gratifying to see that many people from disparate linguistic and non-linguistic backgrounds have found UD congenial enough that they have felt able and motivated to use it.

6. Conclusion and Outlook

In this article, we have articulated the linguistic theory underlying the UD framework. After discussing basic theoretical assumptions (Section 2), we showed how the theory

applies to a wide range of linguistic constructions (Section 3), zoomed in on the treatment of core arguments in a diverse sample of languages (Section 4), and concluded by revisiting the design principles of UD (Section 5). We argued that UD provides a good foundation for crosslinguistically consistent morphosyntactic annotation, which can support research and application development in NLP, as well as typologically oriented studies in linguistics. The UD resources have already had a significant impact on NLP research, most notably for multilingual dependency parsing through two editions of CoNLL shared tasks (Zeman et al. 2017, 2018), which have created a new generation of parsers that handle a large number of languages and that parse from raw text rather than relying on pre-tokenized input. The resources have also been widely used for research on cross-lingual and polyglot parsing, as well as universal semantic parsing (see, e.g., Tiedemann 2015; Agić 2017; Kondratyuk and Straka 2019; Reddy et al. 2017), where the availability of resources with crosslinguistically consistent annotation is crucial. Among more linguistically oriented studies, we find research on psycholinguistics and especially word order typology (see, e.g., Futrell, Mahowald, and Gibson 2015; Naranjo and Becker 2018; Levshina 2019). For an overview of UD-related research, we refer to the proceedings from the annual UD workshops (de Marneffe, Nivre, and Schuster 2017; de Marneffe, Lynn, and Schuster 2018; Rademaker and Tyers 2019; de Marneffe et al. 2020).

Before we conclude, it is important to note that there are many details of the theory that still need to be worked out. Even though all major construction types are covered by the current version of the UD guidelines, there are many specific phenomena and special cases that have not been discussed in sufficient detail or received a definitive treatment in UD. Moreover, the list of such phenomena constantly grows as new languages are considered for analysis in the UD framework. Therefore, while we regard the core of the UD theory as stable, we expect the theory as a whole to continue to evolve over time, as a result of the ongoing dialogue between experts on different languages trying to find the right balance between language-specific and universal perspectives in the application of UD to their language. We look forward to continuing that dialogue and welcome everyone who is interested to take part in it.

Acknowledgments

Many people have contributed to the development of UD, and we especially want to mention our colleagues in the UD core guidelines group, Filip Ginter, Yoav Goldberg, Jan Hajič, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Sebastian Schuster, Natalia Silveira, Reut Tsarfaty, and Francis Tyers, as well as William Croft, Kim Gerdes, Sylvain Kahane, Nathan Schneider, and Amir Zeldes. We are grateful to Google for sponsoring the UD project in a number of ways, and to the *Computational Linguistics* reviewers for helpful suggestions. Daniel Zeman's and Joakim Nivre's contributions to this work were supported by grant GX20-16819X of the Czech Science Foundation and grant 2016-01817 of the Swedish Research Council, respectively.

References

Agić, Željko. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Göteborg.

Andrews, Avery D. 2007. The major functions of the noun phrase. In Timothy Shopen, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, Cambridge University Press, pages 132–223. https://doi.org/10.1017/CBO9780511619427.003

Aronoff, Mark. 2007. In the beginning was the word. *Language*, 83:803–830. https://doi.org/10.1353/lan.2008.0042

Badawi, Elsaid, M. G. Carter, and Adrian Gully. 2013. *Modern Written Arabic: A Comprehensive Grammar*. Routledge, London/New York. https://doi.org/10.4324/9780203351758

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal. https://doi.org/10.3115/980845.980860

Black, Ezra, Roger Garside, and Geoffrey Leech, editors. 1993. *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Rodopi, Amsterdam.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics*, 42:531–573. https://doi.org/10.1017/S0022226706004191

Blevins, James P., Farrell Ackerman, and Robert Malouf. 2017. Word and paradigm morphology. In Jenny Audring and Francesca Masini, editors, *The Oxford Handbook of Morphological Theory*. Oxford University Press, Oxford, pages 265–284. https://doi.org/10.1093/oxfordhb/9780199668984.013.22

Boneh, Nora and Léa Nash. 2012. Core and non-core datives in French. In Beatriz Fernández and Ricardo Etxepare, editors, *Variation in Datives*. Oxford University Press, Oxford. pages 22–49. https://doi.org/10.1093/acprof:oso/9780199937363.003.0002

Bouma, Gosse, Jan Hajič, Dag Haug, Joakim Nivre, Per Erik Solberg, and Lilja Øvrelid. 2018. Expletives in Universal Dependency Treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26, Bruxelles. https://doi.org/10.18653/v1/W18-6003

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2016. *Lexical-Functional Syntax*, 2nd edition. Wiley-Blackwell, Chichester. https://doi.org/10.1002/9781119105664

Bresnan, Joan and Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Linguistic Theory*, 13:181–254. https://doi.org/10.1007/BF00992782

Chomsky, Noam. 1970. Remarks on nominalization. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*. Ginn and Co., pages 11–61.

Comrie, Bernhard. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*. Basil Blackwell, Oxford.

Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. University of Chicago Press.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001

Croft, William. 2002. *Typology and Universals*, second edition, Cambridge University Press.

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

Dahlstrom, Amy. 1991. *Plains Cree Morphosyntax*. Garland, New York.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press. https://doi.org/10.1163/9781849500104

De Guzman, Videa. 1988. Ergative analysis for Philippine languages: An analysis. In Richard McGinn, editor, *Studies in Austronesian Linguistics*. Ohio University Center for International Studies, Athens, OH, pages 323–345.

de Marneffe, Marie Catherine, Miryam de Lhoneux, Joakim Nivre, and Sebastian Schuster, editors. 2020. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*. Online. https://www.aclweb.org/anthology/2020.udw-1.0.

de Marneffe, Marie Catherine, Teresa Lynn, and Sebastian Schuster, editors. 2018. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Bruxelles.

de Marneffe, Marie Catherine, Joakim Nivre, and Sebastian Schuster, editors. 2017. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Göteborg.

Dione, Cheikh Bamba. 2019. Developing Universal Dependencies for Wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23, Paris. https://doi.org/10.18653/v1/W19-8003

Dixon, R. M. W. 1972. *The Dyirbal Language of North Queensland*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139084987

Dixon, R. M. W. 1994. *Ergativity*. Cambridge University Press, Cambridge.

Dixon, R. M. W. 2009. *Basic Linguistic Theory. Volume 1: Methodology*. Oxford University Press.

Dryer, Matthew S. 2007. Word order. In Timothy Shopen, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, second edition. Cambridge

University Press, Cambridge, pages 61–131. https://doi.org/10.1017/CBO9780511619427.002

Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the semantic Web. *GLOT International*, 7:97–100.

Fischer, Wolfdietrich. 1997. Classical Arabic. In Robert Hetzron, editor, *The Semitic Languages*, Routledge, London/New York, pages 187–219.

Foley, William A. 2007. A typology of information packaging in the clause. In Timothy Shopen, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, second edition. Cambridge University Press, Cambridge, pages 362–446. https://doi.org/10.1017/CBO9780511619427.007

Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341. https://doi.org/10.1073/pnas.1502134112, PubMed: 26240370

Gerdes, Kim, 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, Bruxelles. https://doi.org/10.18653/v1/W18-6008

Gerdes, Kim and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, pages 131–140, Berlin. https://doi.org/10.18653/v1/W16-1715

Gerdts, Donna. 1988. Antipassives and causatives in Ilokano: Evidence for an ergative analysis. In Richard McGinn, editor, *Studies in Austronesian Linguistics*, Ohio University Center for International Studies, Athens, OH, pages 323–345.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, MIT Press, pages 73–113.

Grimshaw, Jane. 1991 [2005]. Extended projection. Ms., Brandeis University. Appears in Jane Grimshaw (2005), *Words and Structure*. Stanford, CA: CSLI Publications, pages 1–74.

Guilfoyle, Eichne, Henrietta Hung, and Lisa Travis. 1992. Spec of IP and Spec of VP: Two subjects in Austronesian languages. *Natural Language and Linguistic Theory*, 10:375–414. https://doi.org/10.1007/BF00133368

Haspelmath, Martin. 2001. Word classes and parts of speech. In *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier Science, pages 16538–16545. https://doi.org/10.1016/B0-08-043076-7/02959-4

Haspelmath, Martin. 2011a. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45:31–80. https://doi.org/10.1515/flin.2011.002

Haspelmath, Martin. 2011b. On S, A, P, T, and R as comparative concepts for alignment typology. *Lingustic Typology*, 15:535–567. https://doi.org/10.1515/LITY.2011.035

Haspelmath, Martin. 2014. Arguments and adjuncts as language-particular syntactic categories and as comparative concepts. *Linguistic Discovery*, 12(2):3–11. https://doi.org/10.1349/PS1.1537-0852.A.442

Haspelmath, Martin. 2015. Ditransitive constructions. *Annual Review of Linguistics*, 1:19–41. https://doi.org/10.1146/annurev-linguist-030514-125204

Haspelmath, Martin. 2019. Indexing and flagging, and head and dependent marking. *Te Reo*, 62(1):93–115.

Hayes, David G. 1964. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525. https://doi.org/10.2307/411934

Himmelmann, Nikolaus P. 2005. The Austronesian languages of Asia and Madagascar: Typological characteristics. In Alexander Adelaar and Nikolaus P. Himmelmann, editors, *The Austronesian Languages of Asia and Madagascar*. Routledge, London/New York, pages 110–181.

Hopper, Paul J. and Elizabeth Traugott. 2003. *Grammaticalization*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Hudson, Richard A. 1984. *Word Grammar*. Blackwell.

Hudson, Richard A. 1990. *English Word Grammar*. Blackwell.

Kaplan, Ron and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal

system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, MIT Press, pages 173–281.

Kayne, Richard S. 1984. *Connectedness and Binary Branching*. Foris Publications, Dordrecht. https://doi.org/10.1515/9783111682228

Kondratyuk, Daniel and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2779–2795, Hong Kong. https://doi.org/10.18653/v1/D19-1279

Kroeger, Paul R. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford University Press, Stanford, CA, USA.

Ladusaw, William A. 1988. A proposed distinction between levels and strata: In *Linguistics in the Morning Calm 2: Selected Papers from SICOL-1986*. The Linguistic Society of Korea, Hanshin, Seoul.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572. https://doi.org/10.1515/lingty-2019-0025

Manning, Christopher D. 1996. *Ergativity: Argument Structure and Grammatical Relations*. CSLI Publications, Stanford, CA.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330. https://doi.org/10.21236/ADA273556

Mel'čuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Milicevic, Jasmina. 2006. A short guide to the Meaning-Text linguistic theory. *Journal of Koralex*, 8:187–233.

Naranjo, Matías Guzmán and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104, Oslo.

Nivre, Joakim, 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)*, Portorož.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036. Online.

Osborne, Timothy and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa*, 4(1):17.1–28. https://doi.org/10.5334/gjgl.537

Oyharçabal, Bernard. 2003. Lexical causatives and causative alternation in Basque. In Bernard Oyharçabal, editor, *Inquiries into the Syntax-Lexicon relations in Basque*, number XLVI in Supplements of ASJU. Euskal Herriko Unibertsitatea, pages 223–253.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank. *Computational Linguistics*, 31:71–106. https://doi.org/10.1162/0891201053630264

Payne, Thomas. 1982. Role and reference related subject properties and ergativity in Yup'ik Eskimo and Tagalog. *Studies in Language*, 6:75–106. https://doi.org/10.1075/sl.6.1.05pay

Perlmutter, Davd M., editor. 1983. *Studies in Relational Grammar*. The University of Chicago Press.

Pineda, Anna. 2013. Romance double object constructions and transitivity alternations. In *Proceedings of ConSOLE XX*, pages 185–211, Leipzig.

Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI Publications.

Przepiórkowski, Adam. 2016. How *Not* to distinguish arguments from adjuncts in LFG. In *Proceedings of the Joint 2016 Conference on Head-Driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 560–580, Warszawa.

Rademaker, Alexandre and Francis Tyers, editors. 2019. *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*. Paris.

Reddy, Siva, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 89–101, Copenhagen. https://doi.org/10.18653/v1/D17-1009

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing*, pages 1–15, Mexico City. https://doi.org/10.1007/3-540-45715-1_1

Schachter, Paul and Timothy Shopen. 2007. Part-of-speech systems. In Timothy Shopen, editor, *Language Typology and Syntactic Description. Volume I: Clause Structure*, Cambridge University Press, second edition. Cambridge, pages 1–60. https://doi.org/10.1017/CBO9780511619427.001

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel.

Spencer, Andrew and Ana R. Luís. 2012. *Clitics: An Introduction*. Cambridge University Press, Cambridge.

Stump, Gregory T. 2001. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486333

Sylak-Glassman, John, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing. https://doi.org/10.3115/v1/P15-2111

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Editions Klincksieck, Paris.

Tesnière, Lucien. 2015 [1959]. *Elements of Structural Syntax*. Translation by Timothy Osborne and Sylvain Kahane of Tesnière (1959). John Benjamins. https://doi.org/10.1075/z.185

Thompson, Sandra A. 1997. Discourse motivations for the core-oblique distinction as a language universal. In Akio Kamio, editor, *Directions in Functional Linguistics*. John Benjamins, pages 59–82. https://doi.org/10.1075/slcs.36.06tho

Tiedemann, Jörg. 2015. Cross-lingual dependency parsing with Universal Dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling)*, pages 340–349, Uppsala.

Van Peteghem, Marleen. 2006. Le datif en français: un cas structural. *Journal of French Language Studies*, 16:93–110. https://doi.org/10.1017/S0959269506002286

Van Valin, Jr., Robert D., editor. 1993. *Advances in Role and Reference Grammar*. John Benjamins. https://doi.org/10.1075/cilt.82

Whorf, Benjamin Lee. 1956. *Language, Thought, and Reality*. MIT Press.

Wolvengrey, Arok Elessar. 2011. *Semantic and pragmatic functions in Plains Cree syntax*. Ph.D. thesis, LOT, Utrecht, Netherlands.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Bruxelles.

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver. https://doi.org/10.18653/v1/K17-3001

Zwicky, Arnold M. and Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English *n't*. *Language*, 59:502–513. https://doi.org/10.2307/413900

Zúñiga, Fernando and Beatriz Fernández. 2019. Grammatical relations in Basque. In Balthasar Bickel and Alena Witzlack-Makarevich, editors, *Argument selectors: A new perspective on grammatical relations*, Typological Studies in Language, volume 123 of *Typological Studies in Language*. John Benjamins, Amsterdam, pages 185–211. https://doi.org/10.1075/tsl.123.06zun