Function Words in Universal Dependencies[1]

Marie-Catherine de Marneffe
FNRS – UCLouvain – The Ohio State University

Joakim Nivre
Uppsala University – RISE Research Institutes of Sweden

Daniel Zeman
Faculty of Mathematics and Physics, Charles University

1. Introduction

Languages differ in the way they encode grammatical structure, drawing on a wide range of strategies including the linear arrangement of words and phrases, morphological processes such as inflection, derivation, compounding and incorporation, as well as the use of specialized particles which may be realized as clitics or independent words. The latter are commonly known as *function words* and constitute the theme of the articles in this volume. Most theories of grammar have the ambition to capture what is common to all languages despite the large variation in surface structure. Such an ambition is evident, for example, in the work of Tesnière (2015[1959]), who proposes a theory of syntax where the notion of *dependency* plays a central role, and draws on examples from over sixty languages to illustrate its universal applicability. In a similar vein, Croft (2001, 2016) gives a construction-based account of typological variation based on two types of comparative concepts: universal *constructions* and language-specific *strategies*.

Universal Dependencies (UD) is a grammatical theory that has developed in parallel with a framework for morphosyntactic annotation, having universal applicability and cross-linguistic consistency as central goals. As the name implies, UD incorporates a notion of dependency in its analysis of grammatical structure, a notion that is quite similar to Tesnière's original notion in that it primarily applies to the grammatical relations involved in predication and modification, relations that are also central to Croft's universal constructions, while

[1] The authors have made equal contributions. We thank Adam Moss for his help with computing some of the statistics used in the empirical study and producing the diagrams.

Linguistic Analysis, 43 3-4
2024 Linguistic Analysis
8832 SW Dilworth Rd, Vashon, WA 98070

the internal structure of phrases involving function words and other language-specific means of morphosyntactic realization is analyzed in slightly different terms. UD posits a special set of functional relations loosely corresponding to Tesnière's notion of *transfer* and Croft's notion of strategy.[2] As a consequence, UD does not attempt to capture in its structural representations all aspects of surface realization and constituency, and these representations therefore may appear quite different from frameworks that use dependency structure primarily to analyze surface syntactic structure.[3] Thus, the structure of UD representations needs to be interpreted together with the relation labels used to distinguish different kinds of grammatical relations – some of which are not dependency relations in the narrow sense corresponding to Tesnière's original notion.[4] This point is crucial in order to understand the UD approach in general, and its analysis of function words in particular.

In this article, we explain, motivate, and exemplify the UD analysis of function words against the broader theoretical background of the UD framework. We begin by examining the core theoretical assumptions of UD and explain how they naturally lead to a treatment of function words in terms of special functional relations. We then characterize and discuss these functional relations at a theoretical level, before embarking on a large-scale empirical investigation of four of the most important relations based on the current repository of UD treebanks.[5] We conclude with some reflections on the UD approach to cross-lingual syntax and its relation to other frameworks.

2. Universal Dependencies

Universal Dependencies has developed primarily as a framework for cross-linguistically consistent morphosyntactic annotation, which has at the time of writing been applied to 122 languages (UD v2.9) (Nivre et al. 2016; Nivre et al. 2020). This framework was to a large extent created by merging three pre-existing frameworks: the Stan-

[2] For a detailed discussion of the relation between UD relations, on the one hand, and constructions and strategies, on the other, see Croft et al. (2017).

[3] For a historical overview of the treatment of function words in different dependency grammar frameworks, see Osborne and Maxwell (2015).

[4] The French term used by Tesnière for this notion is connexion.

[5] The UD treebanks are documented on the UD website (https://universaldependencies.org) and are released through LINDAT/CLARIAH (https://lindat.mff.cuni.cz).

ford Typed Dependencies for syntactic relations (de Marneffe et al. 2006; de Marneffe and Manning 2008; de Marneffe et al. 2014), the Google Universal Part-of-Speech Tagset for word classes (Petrov et al. 2012), and the Interset system for fine-grained morphological classification (Zeman 2008). However, it is important to note that these frameworks have undergone substantial revision and harmonization in the development of UD. Thus, while 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. This theory is laid out in de Marneffe et al. (2021) and we will not be able to discuss it in detail here, but we believe that a review of the basic tenets is necessary to provide context for the discussion and investigation of function words in later sections.

UD assumes that grammatical structure is essentially about information packaging, and that the organization of all human languages reflects a basic world view where entities (or objects) participate in events (actions, states). We therefore expect all languages to have two fundamental linguistic units: *nominals*, canonically used for representing entities, and *clauses*, canonically used for representing events. In addition, both nominals and clauses can be refined by *modifiers*, which describe attributes of entities or events.

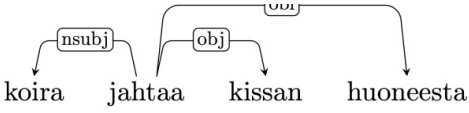
All three fundamental linguistic units may have internal structure. This is most obvious for clauses, which are organized around a *predicate* expressing a state or action, but which may also include nominals, modifiers and other clauses. Nominals and modifiers can also contain all three fundamental linguistic units, although (nonclausal) modifiers mainly contain other modifiers. To describe this hierarchical structure, UD adopts a dependency grammar perspective. A phrase has a *head*, and other elements are *dependents* of that head. The head of a nominal is canonically a noun or a pronoun. 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 words are adjectives and adverbs.

Example (1) shows how UD represents the dependency structure of a simple main clause in Finnish, consisting of a predicate verb (*jahtaa* 'chase') with three nominal dependents (*koira* 'dog.Nom',[6]

[6] We use UD-defined feature values where useful in the glosses (e.g., "Nom" = nominative). We capitalize them following the UD convention, to distinguish them from Leipzig glosses, with which they are not always compatible.

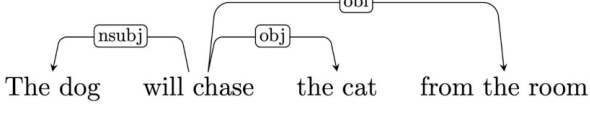
kissan 'cat.Acc', *huoneesta* 'room.Ela'). Dependency relations are typed with grammatical relation labels, as discussed in further detail below, and hold between elementary syntactic units which, in this example, correspond to single words. This type of unit is what Tesnière (1959) calls a *nucleus*.

(1)



Example (2) shows the dependency structure for a possible translation of the Finnish sentence into English, again consisting of a verbal predicate (*will chase*) with three nominal dependents (*the dog*, *the cat*, *from the room*). The difference is that in English all four elements are realized as multiword expressions, where the linguistic head functions are divided between a semantic center – the verb *chase*, the nouns *dog*, *cat* and *room* – and one or more function words – the article *the*, the auxiliary *will*, and the preposition *from*. This type of realization is what Tesnière (1959) calls a *dissociated nucleus*.

(2)



For Tesnière, the elements of a dissociated nucleus are not related to one another by dependency relations of the same kind as those connecting predicates, nominals and modifiers. Instead, he uses the concept of *transfer* to analyze their internal structure, essentially treating function words as category-changing operators. For example, the addition of the preposition *from* in the example above turns the nominal *the room* into an expression that can appear as a modifier rather than as a core argument of the predicate. Tesnière's concept of nucleus is useful for capturing the essential dependency structure in terms of nominals, clauses and modifiers in a way that abstracts over the concrete morphosyntactic realization of nuclei in different languages. The UD analysis of grammatical structure is largely

compatible with this view of syntactic structure, and we will refer to the relations connecting predicates, nominals and modifiers as *central dependency relations*.

However, because UD representations are formally spanning trees over the words of a sentence, one of the elements of a dissociated nucleus has to be formally treated as the head (or parent) in the tree structure. In these cases, UD consistently chooses the lexical or content word as the head, and makes function words dependents of the head with special functional relations to indicate their status as nucleus elements. This choice follows naturally from the decision to prioritize predicate-argument and modifier relations in the syntactic structure. It also makes cross-linguistic similarities more transparent, since direct relations between the semantic cores of nuclei are more likely to be parallel across languages, whereas function words frequently correspond to morphological inflection (or nothing at all) in other languages. Thus, what counts as the head of a nucleus is likely to be more parallel across languages (and sometimes also within languages) if the content word is consistently analyzed as the head. The analysis of function words in terms of special functional relations will be discussed in detail in Section 3. Before that, we want to review some of the other fundamental assumptions of UD that are important as background.

One basic assumption that has been implicit in the discussion so far is that UD follows traditional grammar in giving primary status to *words*. Words are the basic building blocks of grammatical structure; they have morphological properties and enter into syntactic relations with other words. This view can be seen 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 cross-linguistically consistent manner (Haspelmath, 2011a), 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. It is important to note, however, that the relevant morphosyntactic notion of word does not always coincide with orthographical or phonological units. This means, among other things, that clitics (Spencer and Luís, 2012) often have to be separated from their hosts and treated as independent words even if they are not recognized as

such in conventional orthography. Similarly, compound words may need a special treatment depending on orthographic conventions: compare, for example, *night school* in English with *Abendschule* 'night school' in German.

To describe words and their internal structure, UD uses a combination of universal part-of-speech tags (UPOS) and morphological features. Partly for broad comprehensibility, the inventory of UPOS tags stays fairly close to traditional parts of speech, 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 categories shown in Table 1.The first 6 rows are part-of-speech tags often associated with function words.

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

Table 1: Universal part-of-speech tags (UPOS)
(de Marneffe et al, 2021)

The categories in Table 1 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. Moreover, the exact criteria for drawing

the line between different categories are by necessity language-specific. For example, the category **AUX** (for auxiliary) is reserved for words encoding the tense, aspect, mood or evidentiality status of the predicate of a clause, but the extent to which these functions are expressed by grammaticalized particles or auxiliary verbs varies across languages, as do the criteria for identifying these words. Thus, while the lack of *do*-support is a typical feature of auxiliary verbs in English, which groups modal verbs like *can* and *must* together with temporal auxiliaries like *be* and *have*, the criteria available in other languages may not group these semantic classes of verbs together. In addition, as will be explained in Section 3, we think the class of function words is better analyzed as words realizing certain functional relations, rather than belonging to particular parts of speech.

The UPOS categories are deliberately coarse-grained to be broadly applicable, but in many languages words participate in paradigms of forms that express extra features, such as number or tense. There is therefore a need to further subdivide the appropriate UPOS classes into subclasses according to features which express paradigmatic position. For this purpose, UD defines a system of feature-value pairs that are attested in multiple languages and also allows language-specific features to be defined if necessary. This system is compatible with other initiatives to define a universal set of morphological features, such as UniMorph (Sylak-Glassman et al., 2015) and the GOLD Ontology (Farrar and Langendoen, 2003). We refer the reader to de Marneffe et al. (2021) for more details.

Morphological features play an important role in capturing cross-linguistic variation in the realization of syntactic nuclei and can be said to be in complementary distribution with functional syntactic relations. For example, in the Finnish sentence in (1), all the nouns are inflected for the grammatical category of case, which is captured by morphological features on the nouns, such as 'Case=Ela' for *huoneesta*. By contrast, when case is marked by an adposition, forming a dissociated nucleus with the noun, as in the English nominal *from the room* in (2), the noun instead has a dependent with the functional relation **case**. In both cases, the grammatical marker is anchored in the noun, which forms the semantic core of the nucleus. Note, however, that grammatical markers are not always in one-to-one correspondence. For example, the English definite articles have no counterparts in Finnish, because definiteness is not grammaticalized in that language, and two of the English nouns (*dog*, *cat*) have neither morphological nor syntactic case markers.

A second basic tenet of UD, besides the primacy of words as grammatical units, is a commitment to *grammatical relations* as 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 this respect, UD adheres to a long tradition represented by theories like relational grammar (Perlmutter 1983), lexical-functional grammar (LFG) (Kaplan and Bresnan 1982; Dalrymple 2011; Bresnan et al. 2016), word grammar (Hudson 1984; 1990), functional generative description (Sgall, Hajičová and Panevová 1986), meaning-text theory (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 cross-linguistic studies is rarely questioned.

It is important to note that, while a commitment to grammatical relations is naturally compatible with a dependency-based view of grammatical structure, it actually goes beyond it. First of all, as noted earlier, not all grammatical relations distinguished in UD are central dependency relations. Secondly, even for those relations that are, grammatical relations provide a more fine-grained classification than the bare dependency structure. This is perhaps most obvious for core argument relations like *subject* and *object*, which in the prototypical case are both dependency relations holding between a predicate and a nominal but which are nevertheless distinct from each other as grammatical relations.

Dependent → Head ↓	Nominal	Clause	Modifier Word	Function word
Clause core	nsubj obj iobj	csubj ccomp xcomp		
Clause other	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal	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

Table 2: The UD taxonomy of universal grammatical relations.

To categorize grammatical relations between words, UD provides a taxonomy of 37 universal relations, illustrated in Table 2.[7] The central part of this taxonomy is organized by two main principles. The first is the core-oblique distinction (Thompson 1997; Andrews 2007), which 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 second is the recognition of the three fundamental linguistic units: nominals, clauses and modifiers. Thus, the first three rows of Table 2 list relations used to classify (i) core arguments, (ii) other dependents at the clause level, and (iii) dependents inside nominals. Each row has one column each for dependents in the form of (i) nominals, (ii) clauses, and (iii) modifier words. The relations in these groups—shaded green areas in Table 2—can all be considered central dependency relations.

The fourth column in each of these rows contains function word relations that occur in clauses and nominals, respectively. As discussed earlier, these relations have a special status as they occur in dissociated nuclei rather than in traditional dependency structures following Tesnière, and will be discussed in detail in the rest of this article. The bottom row of Table 2 contains relations that are necessary to analyze various types of constructions in natural language,

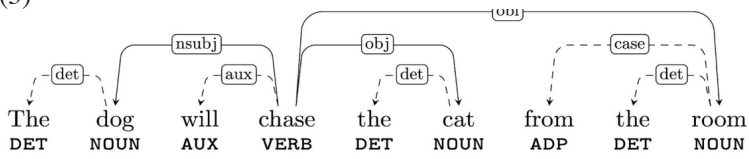
but which do not clearly satisfy the criteria for central dependency relations. One of these relations, the **cc** relation, is used to link coordinating conjunctions to coordinated phrases and will also be treated as a function word relation in the next section. For a more detailed discussion of the UD taxonomy of grammatical relations, we refer to de Marneffe et al. (2021).

Summing up, the UD analysis of grammatical structure is based on grammatical relations between words (and part-of-speech tags and features to classify the words themselves). The structural representation of a sentence forms a spanning tree over the words of the sentence, and the core of this tree structure consists of typed dependency relations for core arguments and oblique modifiers. However, this representation may also contain special relations that encode other types of relations, in particular relations that combine the elements of dissociated nuclei in Tesnière's sense. These relations are fundamental to the UD analysis of function words, to which we turn next.

3. Function Words in UD

We adopt the view of a function word being the realization of some grammatical category as an independent word form as opposed to morphological realization. Thus, by essence, function words constitute key differences in morphosyntactic realization of universal constructions across languages. As mentioned in Section 2, UD, therefore, in its desire of promoting parallelism between languages, does not choose function words as heads of constituents. If such a treatment of function words departs from many dependency grammars, it aligns with Tesnière's notion of *transfer* and Croft's notion of strategy. By prioritizing predicate-argument and modifier relations, and using special functional relations when a grammatical category is realized by a word (instead of morphologically – which UD encodes as features), UD gives parallel representations to universal constructions which vary in their instantiation strategies for different languages. (3) illustrates the grammatical and functional relations for the English sentence *The dog will chase the cat from the room*.

(3)



The solid arcs represent the dependencies between content words (the nouns and the verb), while the dashed arcs represent the relations linking a function word to its head: determiners are attached to the noun they determine, auxiliaries to the verb they modify, and prepositions to their complements. (3) thus shows a complete dependency tree (where each word in the sentence is the end point of a labeled relation) compared to (2), and illustrates three of the seven functional relations in UD: **det**, **aux**, **case**. As seen above, (1) shows the sentence in Finnish, which has the same syntactic structure as the English sentence (a predicate with a subject, an object and a locative modifier), but differs in its morphosyntactic realization. Indeed, Finnish does not explicitly encode definiteness nor future tense, and it further uses case makers on the nouns (as for instance the elative case on *huoneesta* 'room' to indicate its locative function). Despite adopting different strategies, the English and Finnish dependency structures produced by UD are consistent.

UD thus defines function words in terms of special relations rather than in terms of part-of-speech tags or lemmas. In particular, function words in UD cannot be identified with words belonging to closed classes: pronouns are a prime example of closed class words that normally do not act as function words according to UD as they typically function as nominal arguments or oblique modifiers. And even though there is some correspondence between function words and certain part-of-speech tags (such as **ADP**, **AUX**, **DET**, **CCONJ**, **SCONJ**, and **PART**) as mentioned in Section 2, the mapping is not one to one. For example, while adpositions (**ADP**) prototypically occur with the **case** relation, they can also be elements of lexical compounds in particle-verb constructions. Function words also cannot be defined in terms of individual lemmas, since many lemmas alternate between function word uses and other functions. A typical case is the verb lemma *have* in English, which is used as a function word in *they have gone home*, but not in *they have money.*

It is however important to note that, although function words are defined in terms of special relations, these relations are sometimes not directly visible in the dependency representations, notably in cases of ellipsis. For example, in a sentence like *Mary may buy a book and Sam may too*, both instances of *may* are auxiliary verbs, realizing the **aux** relation. However, since the basic UD representation does not permit empty elements, the second instance of *may* has to replace the omitted main verb *buy* as the root of the clause, through a process called promotion. As a result, the second instance of *may* will overtly have the label **conj** on its incoming arc. Implicitly, however, it is still an auxiliary verb realizing the *aux* relation.

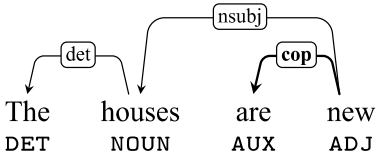
One limitation of the UD analysis of function words is that it does not directly capture the fact that nuclei can have a nested hierarchical structure, since all function words associated with a given content word are attached directly to their head in a flat structure. Thus, in a sentence like *they could have gone home*, the nucleus *could have gone* is assigned a structure where the auxiliary verbs *could* and *have* are both direct dependents of *gone*. For languages like English, the nested structure can often in practice be inferred from the dependency tree in combination with the linear order of the dependents, but in principle there is no guarantee that this is always possible.

In the rest of this section, we discuss the seven UD functional relations in more detail: the three relations which appear in clauses: **cop**, **aux**, **mark**; the three relations which appear in nominals: **case**, **clf**, **det**; and the **cc** relation used for coordinators.

3.1 The cop Functional Relation

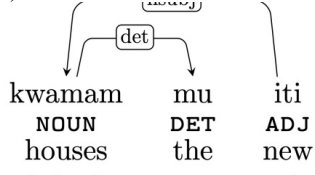
When a language uses a function word ("copula") to connect a nonverbal predicate with its subject, this function word is attached to the nonverbal predicate via the **cop** relation. In (4), for instance, the copular verb *are* is linked to the nonverbal predicate *new*.

(4)



In this UD representation, the subject *the houses* thus directly depends on the nonverbal predicate *new*. This direct relation between the subject and the nonverbal predicate allows one to obtain a parallel dependency structure in languages which employ another strategy for nonverbal predication. For instance, some languages, like Waskia (a language of Papua New Guinea), Russian or Chinese, do not use a copular strategy like English for expressing nonverbal predication, but a zero strategy where only the argument and the predicate are overtly expressed. (5) shows a translation of (4) in Waskia, where no copula is used. Except for the presence of the functional relation **cop** in (4), both structures in (4) and (5) are the same.

(5)

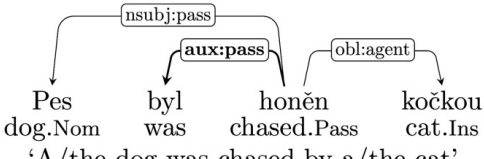


'The houses are new'

3.2 The aux Functional Relation

The **aux** relation links a function word that expresses tense, mood, aspect, voice or evidentiality to a predicate. For instance, in the Czech sentence in (6), the passive auxiliary *byl* is attached to the predicate *honěn*.

(6)

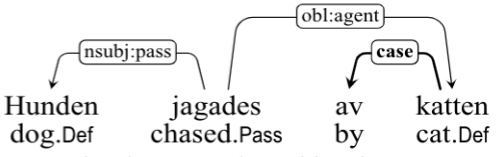


'A/the dog was chased by a/the cat'

As can be seen in this example, UD allows the specification of valency-changing operations, as subtypes of the relations: the relations are subtyped with **pass** and **agent** to signal that the mapping of the grammatical relations to semantic roles has changed.

If we compare the Swedish translation (7) of the Czech sentence above, the central dependency structures in terms of nominals and modifiers are parallel: both the subjects and agents are attached to the predicate.

(7)



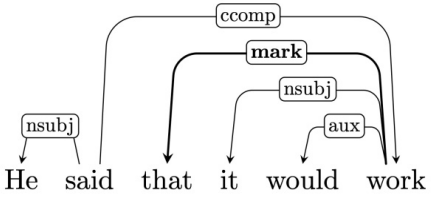
'The dog was chased by the cat'

Swedish, however, differs from Czech in its strategy for expressing the passive construction which is morphologically realized in Swedish, whereas Czech uses an auxiliary. Similarly, both languages differ in how the agent is expressed: it is introduced by a preposition in Swedish, but morphologically marked in Czech. However, by using functional relations (**aux** and **case**) to link the function words within their dissociated nuclei, the UD analysis preserves identical central dependency structures between the two sentences.

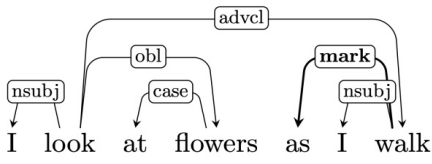
3.3 The mark Functional Relation

In parallel to what is done in matrix clauses, subordinate clauses in UD are organized around the predicate, which is taken to be the head of the clause. If the subordinate clause is introduced by a subordination marker, this marker is attached to the predicate of the clause, as in (8) in a complement clause (**ccomp**) and (9) in an adverbial temporal clause (**advcl**).

(8)

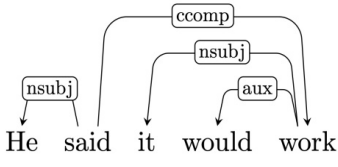


(9)



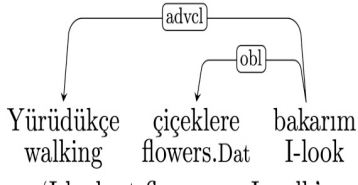
Subordination is not always overtly expressed: complementizers are indeed optional in English in some cases, as in (10), a variation of (8). By linking the subordinating conjunction within the clause it belongs to, the UD analysis preserves identical central dependency structures in terms of nominals and clauses in both sentences.

(10)



Subordinate clauses can also be morphologically marked, as in the Turkish translation of (9), shown in (11): while English uses the subordinating conjunction *as* to introduce the adverbial clause, Turkish uses the morphological marker *-çe*. Again, the UD analysis gives parallel dependency structures between sentences which differ in the strategy used for subordinating clauses: (9) vs. (11). Note that Turkish also uses morphological case instead of a preposition to mark the oblique modifier and subjects are incorporated in the predicates.

(11)



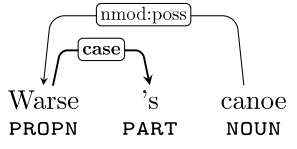
'I look at flowers as I walk'

The functional relation **mark** encompasses subordinating conjunctions introducing complement clauses (e.g, *that* in English) and adverbial clauses (e.g., *if*, *when*, *because*), as well as infinitival markers (*to* in English).

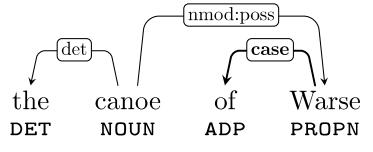
3.4 The case Functional Relation

The Swedish example (7) above illustrates the **case** relation. Case marking is one of the strategies to indicate the grammatical function of a nominal. When case marking is realized by clitics or adpositions (prepositions and postpositions, **ADP** as UD part-of-speech tag), the **case** relation is used to link the case marker to its nominal head, as in (7) where the agent *katten* in Swedish is introduced by the preposition *av*. The English examples (12a) and (12b) also illustrate the **case** relation: English can use a clitic or a preposition to express possession. Other languages, such as Asmat (a Papuan language from New Guinea) uses no marker, as in (12c). Again by linking the overt possessive marker within its nucleus, all three structures in (12) are identical.

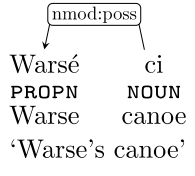
(12) a.



b.



c.



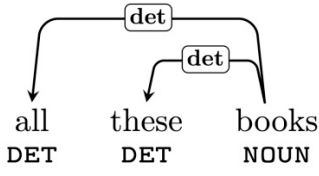
'Warse's canoe'

Note that UD follows Haspelmath (2019) in adopting a unified treatment of all case markers and adpositions, attaching them to their nominal head with the **case** relation.

3.5 The *det* Functional Relation

Properties of nominals, such as definiteness, number or gender can be indicated by a determiner (article, demonstrative, interrogative or quantifier). These determiners are linked to their nominal head in UD, as in (13), with the **det** relation.

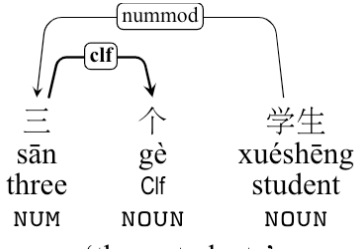
(13)



3.6 The *clf* Functional Relation

Some languages use classifiers, words that reflect a conceptual classification of nouns. Such classifiers accompany nouns in certain grammatical contexts. For instance, classifiers appear with a numeral for counting objects or with a demonstrative, as in the Chinese example (14). In UD, the classifiers are linked to the numeric modifier or the determiner in the nominal, using the **clf** relation.

(14)

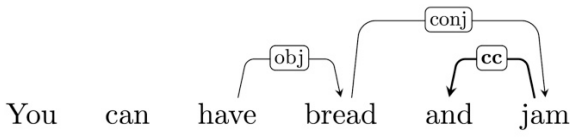


'three students'

3.7 The *cc* Functional Relation

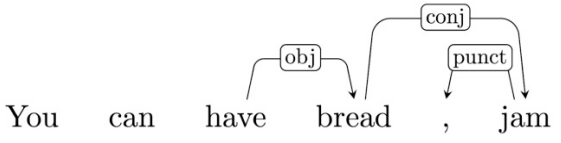
There are several ways of expressing coordination between elements. One way is to use a coordinating conjunction (e.g., *You can have bread and jam*), but it can also be left implicit and indicated simply with punctuation (*You can have bread, jam*). When a coordinating element is used between conjuncts, it is attached to the following conjunct. The relation **cc** is used when the coordinating element is a coordinating conjunction, as in (15).

(15)



(16) shows the same example with a punctuation (comma) between the coordinated elements.

(16)

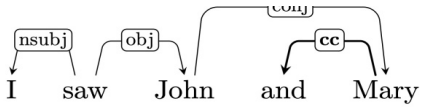


Coordination can also be marked with a clitic (as in Latin *Senatus Populusque Romanus* 'the Senate and the people of Rome') or morphologically (as in Japanese where different morphological markers are used depending on the word categories of the conjuncts). Besides the consideration of obtaining parallel analyses between languages with different strategies for expressing coordination, there are also structural properties of many languages which motivate attaching a coordinating element to an adjacent conjunct as they constitute a phrase together (see Gerdes and Kahane (2016) for discussion of this point).

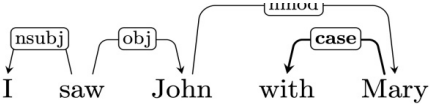
As illustrated by all the examples in this section, what the seven UD functional relations have in common is that they link a function word to the core lexical element of the nucleus or clause it belongs to. These functional relations should therefore not be interpreted as indicating subordination between their elements (contrary to the standard grammatical relations of subject, object, modifier, etc.) but simply as indicating the presence of a functional word that could be expressed by a different strategy in another language.

The analysis of function words in UD also leads to assigning essentially similar structural representations to different syntactic constructions with shared meanings (while the relation labels capture nonetheless the distinct constructions): for instance, associating two elements with the coordinating conjunction *and*, or the preposition *with* as in (17), or subordinate clauses and prepositional modifiers (18). Languages can differ in their preferences for one construction, which can lead to variations in the frequencies of functional relations in languages.

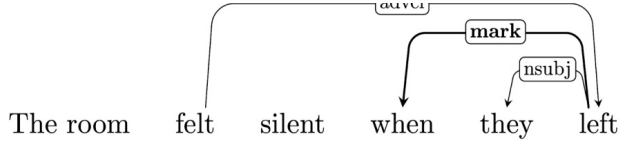
(17) a.



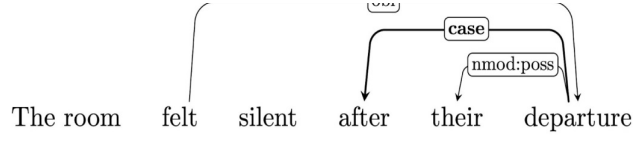
b.



(18) a.



b.



Finally, it should be noted that the distinction between dissociated nuclei involving function words and other constructions is not always clear-cut, nor is the distinction between function words, clitics, and inflectional morphemes. 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 not straightforward to classify the components of the construction. However, the fact that borderline cases exist, where in linguistic annotation we are forced to make a more or less arbitrary decision, does not invalidate the fact that grammaticalized function words need to be analyzed in a different way than the constructions that constitute their historical origin.[7]

[7] The existence of partial grammaticalization is one reason that UD sometimes fails to give parallel analyses to constructions with similar meaning. A prime coun-

4. An Empirical Study of Function Words in UD

UD is not just a theory but also a large collection of data annotated in accordance with that theory, available for many languages. This enables us to conduct a quantitative study of individual categories of function words across languages. Due to limited space, we focus only on four major categories here: **adpositions, subordinators, coordinators,** and **auxiliaries.** In the first part of the study, we investigate the frequency of these functional relations across a broad sample of languages to find out how much variation there is both across relations and across languages. In the second part of the study, we focus on the linear position of function words and its correlation with other word order patterns. In both cases, the purpose is mainly to illustrate how UD resources can be used as the basis for large-scale cross-linguistic comparison.

All statistics in this section are collected on the UD release 2.8 from May 2021 (Zeman et al. 2021). The release contains 202 treebanks for 114 languages. For various reasons, we decided to omit some treebanks from consideration:

- Learner corpora are omitted because they specialize in collecting text produced by non-native speakers, typically with a significant amount of errors. There are three such treebanks: Chinese CSL, English ESL, and Italian Valico.

- Twitter-based corpora are omitted because the language used on Twitter significantly differs from the standard language, and we hypothesize that the focus of this study, i.e., the distribution of function words, is among the areas most affected by the genre specifics.[8] We are aware that many other UD treebanks contain some proportion of user-generated content, weblogs, reviews, as well as various other text types (or spoken data)

terexample is the periphrastic future in French (*Je vais me promener dans la forêt* 'I will walk in the forest' to be contrasted with the simple future *Je me promènerai dans la forêt*) where the root of the sentence is taken to be the periphrastic element *vais* and not the core lexical element *se promener* 'to walk', because the former has not fully grammaticalized into an auxiliary. Light verb constructions (e.g., *to take a picture* vs. *to photograph*) are another common case.

[8] A comparison of Twitter and non-Twitter treebanks for the two languages where such a comparison is possible (Irish, Italian) confirms that function words have lower frequency in the Twitter treebanks. The difference is especially notable for case markers.

that diverge from the standard. We avoid having to quantify the level of "non-standardness" and only exclude treebanks that consist exclusively of Twitter data. There are four such treebanks: Hindi-English HIENCS, Irish TwittIrish, Italian PoSTWITA, and Italian TWITTIRO.

- Code-switching corpora are omitted because analyzing their contents would require extra care to evaluate the contribution of the source languages vs. the impact of code switching. While a limited amount of code-switching may occur in any treebank, there are three treebanks that identify themselves as focusing on code-switched data: Frisian-Dutch Fame, Hindi-English HIENCS, and Turkish-German SAGT.

- The Swedish Sign Language treebank is omitted because it is based on glosses rather than on an actual representation of the sign language.

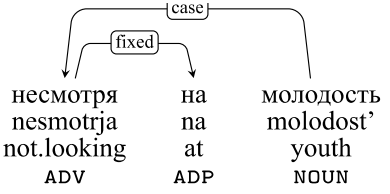
- The validation process of UD data is constantly improved, which also means that new violations of the annotation guidelines may be found in previously released treebanks. We check these legacy treebanks for error types that are related to the analysis of function words: treebanks that have 1% or more of such errors (in relation to the number of non-punctuation tokens) are omitted from the current study.

After applying the filters described above, 169 of the original 202 treebanks remain. We merge treebanks of the same language, yielding datasets for 98 languages. Finally, to avoid drawing conclusions from low numbers of instances, we remove 26 datasets that have less than 5,000 non-punctuation tokens (part-of-speech tag other than **PUNCT**) each. The resulting data contains 72 languages from 16 language families.

When searching for a function word category, we rely mostly on relation labels rather than on part-of-speech tags. The relation connects the function word to the content word in the same nucleus, and the relation label characterizes the grammatical function of the function word within this nucleus. In contrast, part-of-speech tags in UD are less dependent on context and may point to instances that we do not want to include in the statistics. For example, the **ADP**

tag denotes adpositions, most of which will be attached to a nominal via the **case** relation. However, some adpositions in some languages may also be used as lexical morphemes modifying the meaning of a verb (e.g., English *come on*). Despite being traditionally called verbal 'particles', they retain the **ADP** part-of-speech tag in UD; yet their incoming relation will be **compound:prt** rather than **case**. On the other hand, many languages allow other parts of speech or multi-word expressions to act as secondary adpositions. For example, in the Russian nominal *несмотря на молодость (nesmotrja na molodost')* 'despite (his) youth' (19), *несмотря на (nesmotrja na)* is analyzed as a fixed multi-word preposition, that is, *несмотря* and *на* are connected via a technical relation **fixed**, *несмотря* is taken as the technical head and attached via the **case** relation to *молодость (molodost')* 'youth'. Nevertheless, *несмотря* itself is tagged as an adverb rather than as an adposition.

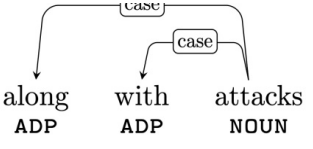
(19)



'despite his youth'

An alternative to the fixed expression analysis is to attach two **case** dependents as siblings to the same content word. There are no UD-wide rules that would determine which annotation alternative should be used; the criteria must be specified for each language separately. Attaching the **case** words as siblings is more natural when each of them can also function as an independent preposition; thus the English preposition collocations *along with*, *out of*, *from within* are all annotated as sibling dependents, as shown in (20).

(20)



A special case in English is when a preposition combines with a genitive clitic (tagged **PART**), both of which use the **case** relation, as in *like Applebee 's*.

These examples demonstrate that there are two possible ways to count function words, yielding different results: 1. count every occurrence of a function word of the given type; 2. count every node that has one or more functional dependents of the given type (here, a nominal that has one or more **case** dependents). The latter approach consistently yields lower numbers; however, the difference in the ranking of languages is negligible, so in the subsequent text we only report counts according to the simpler first approach.

4.1. Relative Frequency of Function Words

Figure 1 shows the languages ordered by the relative frequency of **case** dependents. Languages from the same family and genus share the same color. There are clear clusters of phylogenetically related languages, with most Indo-European and Afro-Asiatic languages occupying the lower half of the scale (with high frequencies), while Turkic and Uralic languages appear near the top of the diagram (with low frequencies). In general, languages with more morphological cases need fewer case-indicating function words. Language development from synthetic to analytical types can be also observed: within the Indic languages, Hindi and Urdu rank among the highest rates of **case** dependents (21% and 20%, respectively), while their predecessor Sanskrit appears at the other end of the scale (less than 1%). Similarly, the Romance languages generally have a very high proportion of **case**, led by Portuguese with 17%, but their predecessor Latin has only 9%.

Note that the distinction between a bound case affix and a function word is also greatly affected by word segmentation, which is challenging in languages whose writing system does not use spaces between words. Hence, for example, in Japanese, there are three different traditions of defining words (Murawaki, 2019). The UD treebanks follow one of the traditions, leading to high numbers of function words (Japanese leads the **case** dependent ranking with 23%). If a different tradition were followed, Japanese would be an agglutinative language with a low proportion of function words.

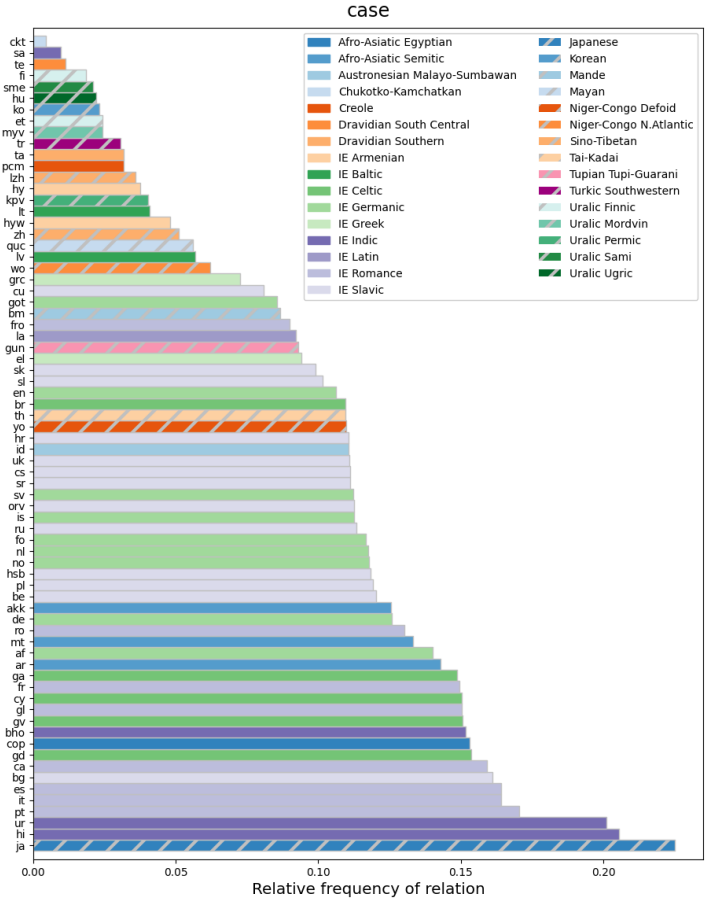


Figure 1:Relative frequency of nodes attached as **case** dependents. Each bar is one language.

An interesting oddity with respect to the above observations is found in the two Sino-Tibetan languages, Chinese and Classical Chinese which have low proportions of case, respectively 5% and 4%. They have no case morphology but they also employ comparatively few adpositions. A possible partial explanation could be that Chinese tends to prefer core predicate-argument relations over oblique modifiers. So we have *I go to* Beijing in English but 我去北京 (*Wǒ qù Běijīng*, lit. 'I go Beijing') in Chinese. Indeed, both the Chinese languages have a much higher proportion of the obj relation (for direct objects) than English.

We now move on to the other three functional relations that we selected for our study: subordinators (**mark**), coordinators (**cc**), and auxiliaries (aux). Two general observations can be made: 1. the relative frequencies are considerably lower than those of case; 2. phylogenetic relationship no longer plays a crucial role.

Figure 2 shows the languages ordered by the relative frequency of mark dependents. The largest number of mark dependents (12%) occur in the Tupian language Mbyá Guaraní. It is the second most-frequent non-punctuation relation in the UD treebanks of this language, owing to the fact that it is used for a wide range of functions there: adverbializers, relativizers, as well as nominalizers (Thomas, 2020). The next ranks are occupied by two Celtic languages (Scottish Gaelic 9%, Irish 6%) and two Afro-Asiatic ones (Coptic 8%, Maltese 7%).

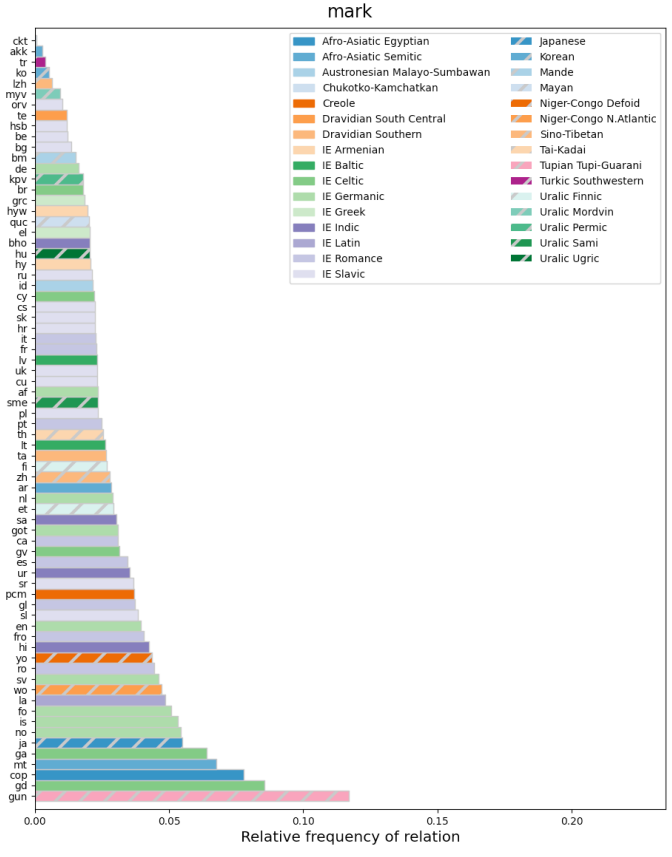


Figure 2. Relative frequency of nodes attached as **mark** dependents. Each bar is one language.

At the other end of the scale, the proportion of subordinators in Turkish, Akkadian, and Chukchi rounds down to 0%. Besides language typology, other factors may affect the counts, such as the genre of the text. In genres where sentences tend to describe relationships between events and can thus be complex, subordinators are more likely to occur than in simple sentences and therefore more mark relations (as in (21) from the UD English EWT corpus, where we have three **mark** relations). However, the exact influence of genres on our study cannot be quantified, as genres are not annotated on a per-sentence basis.

(21)

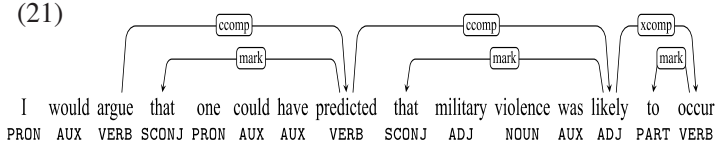


Figure 3 shows the languages ordered by their proportion of **cc** dependents. Here the striking commonality of high-ranking languages is that a significant part of the data consists of old texts. Old East Slavic (9%), Gothic (9%), Old Church Slavonic (8%), Ancient Greek (7%), Latin (6%), and Old French (6%) are all historical language varieties. Faroese (7%) and Icelandic (6%) have both modern and historical data in UD, but the historical parts are much larger. And Arabic (7%), despite being based on modern texts, is grammatically very close to the classical language of the Quran. At the other end, there are five languages whose coordinator rate rounds down to zero. Among them are Japanese and Korean where semantically coordinate structures are analyzed syntactically as subordination.

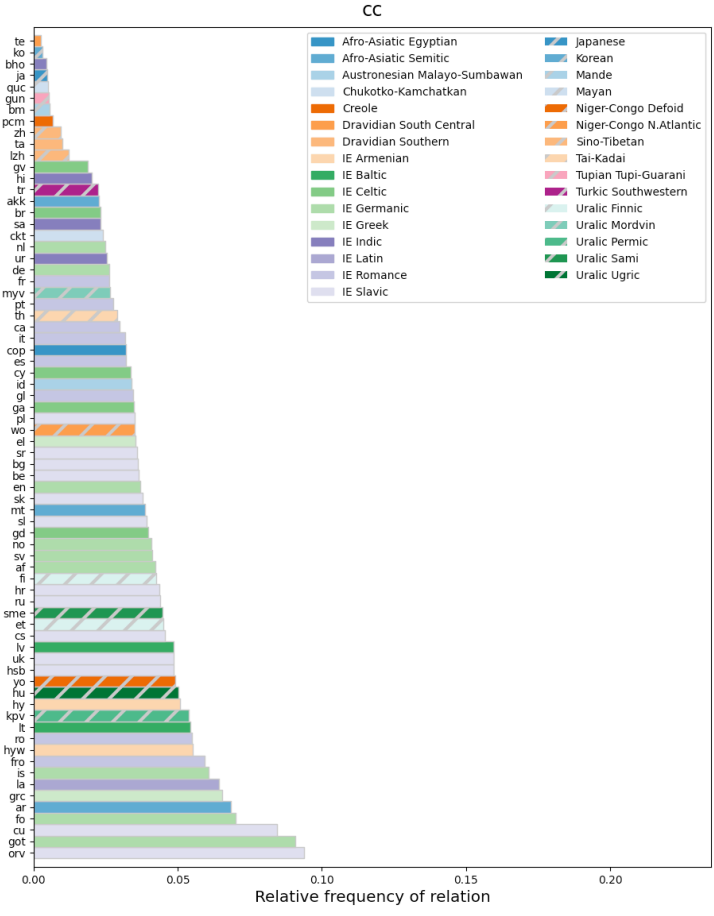


Figure 3: Relative frequency of nodes attached as **cc** dependents. Each bar is one language.

Figure 4 shows the percentage of nodes with the relation **aux** (or one of its subtypes, such as **aux:pass**), i.e., auxiliary verbs and particles. It does not include copulas, which share with auxiliaries the part-of-speech tag **AUX** but have their own relation **cop**. The scale is framed by the Celtic languages, whereas the Goidelic branch (Irish, Scottish Gaelic, and Manx) uses no auxiliaries at all, while Welsh (in the Brittonic branch) has 4%, and Breton's 12% is the largest proportion of auxiliaries observed. Other auxiliary-heavy languages are Bambara (12%), Japanese (9%), and Wolof (8%); other auxiliary-free languages are Akkadian, Gothic, and Telugu.

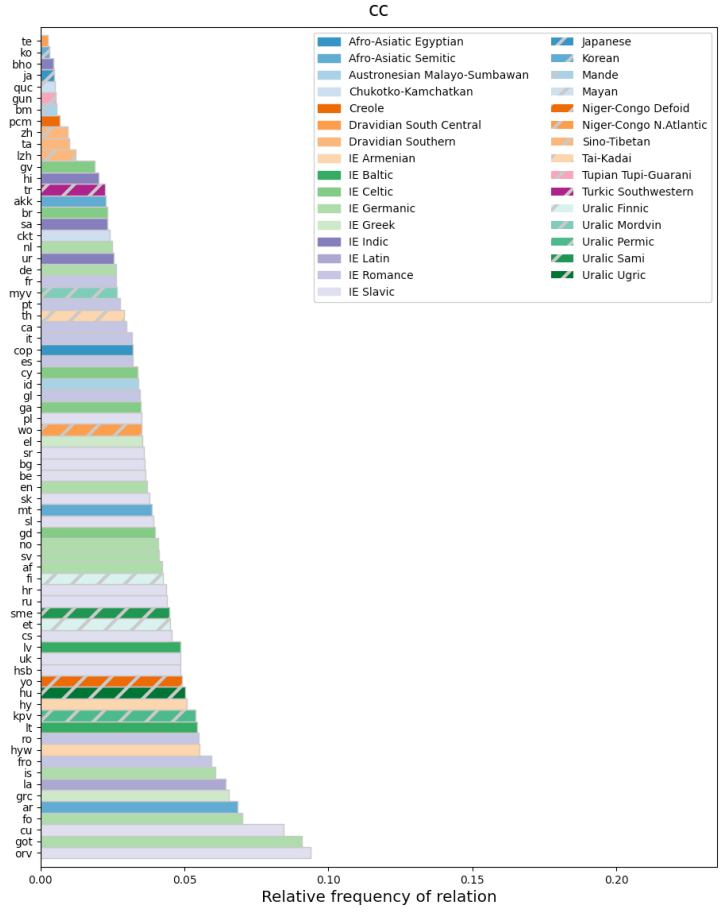


Figure 4: Relative frequency of nodes attached as **aux** dependents. Each bar is one language.

4.2. Linear Order of Function Words

The relation between word order and grammatical functions is one of the cornerstones of syntax, and cross-linguistic comparison of such relations is an important topic in linguistic typology (Futrell et al. 2015; Alzetta et al. 2018; Levshina 2019; Yu et al. 2019; Gerdes et al. 2021). The obvious research question in the context of this volume is how the position of function words correlates with the language's preference for head-initial or head-final dependencies.

UD is particularly suited for such typological exercises, provided that the different types of UD relations are properly acknowledged. To establish whether a language is head-initial or head-final, we need relations that are linguistic dependencies and where the notion of head and subordination exists. We must not consider punctuation, coordination, certain technical relations, and function word attachments in dissociated nuclei. In our experiments, we take predicate-argument and predicate-modifier relations in verbal clauses, that is, the following relation types: **nsubj, csubj, obj, iobj, ccomp, xcomp, obl, advcl, advmod**.[9] We then take our selected functional relations, measure the likelihood that the function word precedes its lexical counterpart, and compare these statistics among head-initial and head-final languages. We only conduct this experiment with the functional relations **case**, **mark**, and **aux**. The coordinators (**cc**) are special because they normally pertain to two (or more) conjuncts, and the UD guidelines dictate that their technical parent node be the immediately following conjunct if possible; their position is mostly between two conjuncts.

Figure 5 shows the percentage of head-final dependencies in our sample of 72 languages. We have 10 **strongly head-final** languages (more than 75% of the examined relations go right-to-left) in our selection: Korean, Tamil, Telugu, Japanese, Turkish, Akkadian, Hindi, Urdu, Bhojpuri, and Sanskrit. Conversely, there are 7 **strongly head-initial** languages (more than 75% of the examined relations go left-to-right): Manx, Scottish Gaelic, Irish, Welsh, Breton, Arabic, and K'iche'. About half of our languages are spread around the center, ranging from Hungarian (61% head-final) to Swedish (62% head-initial).

[9] There are several other possible approaches to establishing the dominant dependency direction in a language. One alternative would be to consider all central dependencies, including those inside nominals. Another alternative would be to only look at the verb-object relation. This approach (which is often the basis for word-order universals in typology) would lead to a quite different language ranking. There would be 11 strongly head-final (OV) languages (Japanese, Korean, Bambara, Tamil, Telugu, Turkish, Bhojpuri, Akkadian, Sanskrit, Afrikaans, Hindi) and as many as 32 strongly head-initial (VO) languages (Naija, Coptic, Manx, Classical Chinese, Scottish Gaelic, Indonesian, Chinese, Arabic, English, Maltese, Swedish, Norwegian, Yoruba, Thai, Portuguese, K'iche', Breton, Faroese, Italian, Spanish, Serbian, French, Irish, Galician, Greek, Icelandic, Bulgarian, North Sámi, Polish, Russian, Welsh, Belarusian).

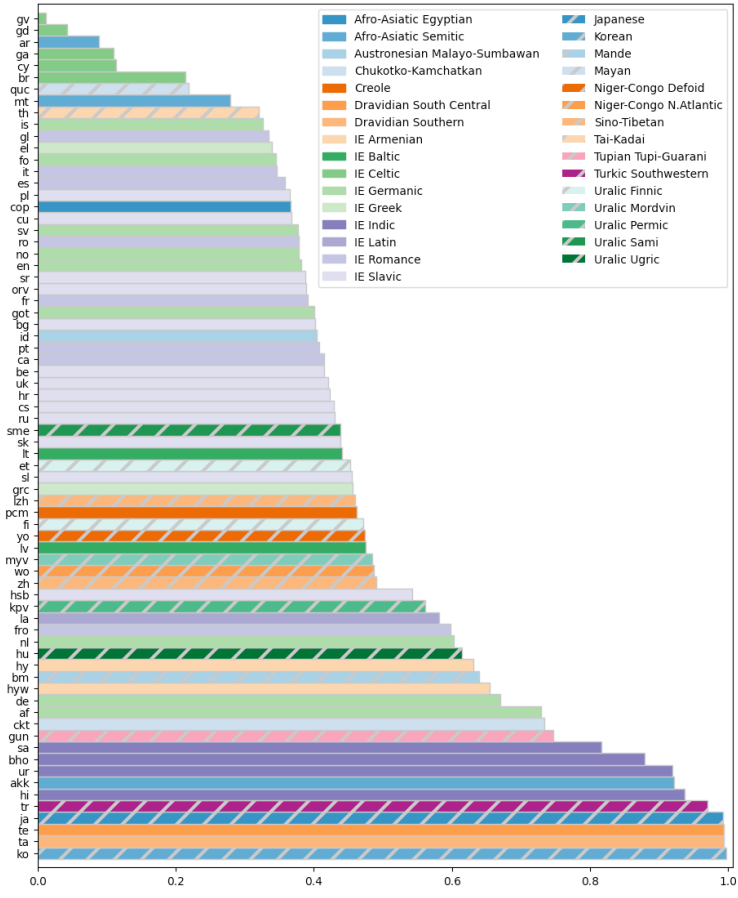


Figure 5: Percentage of head-final dependencies. Each bar is one language.

Having established the proportion of head-final dependencies in different languages as a basis for comparison, we now proceed to investigate how these proportions correlate with the placement of function words, using a type of diagram proposed by Gerdes et al. (2021) to discover hypothetical statistical universals. Starting with adpositions, in Figure 6, we see that, while many languages look undecided between the general directions of the main dependency relations, the preference for either prepositions or postpositions is usually very strong. 50 languages in our sample clearly prefer prepositions (89% of prepositions, observed in Yorùbá, is the mini-

mum); 19 languages strongly prefer postpositions (minimum 79% observed in Estonian); and only 3 languages (Chinese, Classical Chinese, Sanskrit) stay in the middle. In accordance with widely accepted typological findings (Dryer 2007), head-final languages prefer postpositions. In addition, we observe a preference for postpositions in Uralic languages, whose main dependencies are slightly inclined towards head-initial. On the other hand, all Germanic languages strongly prefer prepositions, but some of them (Afrikaans, Dutch, and German) come out as mildly head-final because they often put the main verb at the end of the clause.

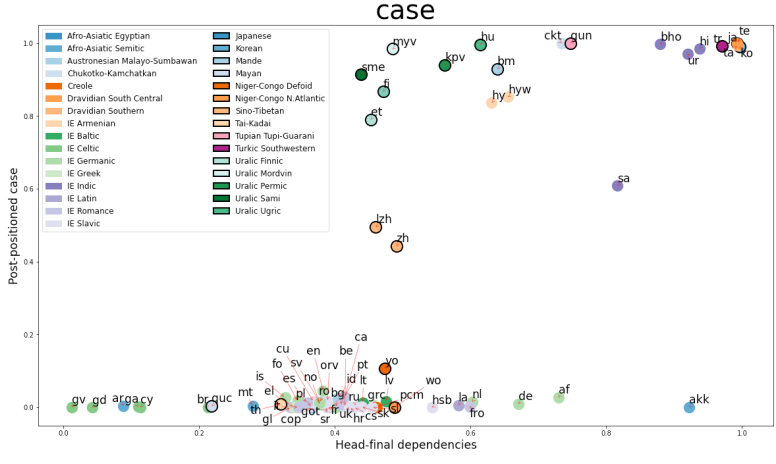


Figure 6: Percentage of post-positioned case markers (y-axis) and head-final dependencies (x-axis).

Regarding **mark** dependents (Figure 7), 6 languages strongly prefer them to follow the head of the marked clause, 6 languages show only weak preferences, and the rest strongly prefer the marker to precede the clausal head (75% for Bhojpuri, 88% for Komi-Zyrian, over 90% for the others). If a language strongly prefers post-markers, then it is also head-final, but the opposite implication obviously does not hold.

As for auxiliaries, only 66 languages actually have them. As seen in Figure 8, 8 languages strongly prefer auxiliaries after the main verb (79% for Latin, 82% for Korean, 99% for the others); again, all of them are head-final. 40 languages strongly prefer auxiliaries before the main verb (80% for Thai, 9 other languages between 80 and 90%, 14 languages between 90 and 99%, 16 languages over

99%). Out of these, 34 languages are head-initial, while 6 languages show a tendency towards head-final but not strongly head-final: Upper Sorbian 54%, Komi-Zyrian 56%, Old French 60%, Dutch 60%, Bambara 64%, Chukchi 73%.

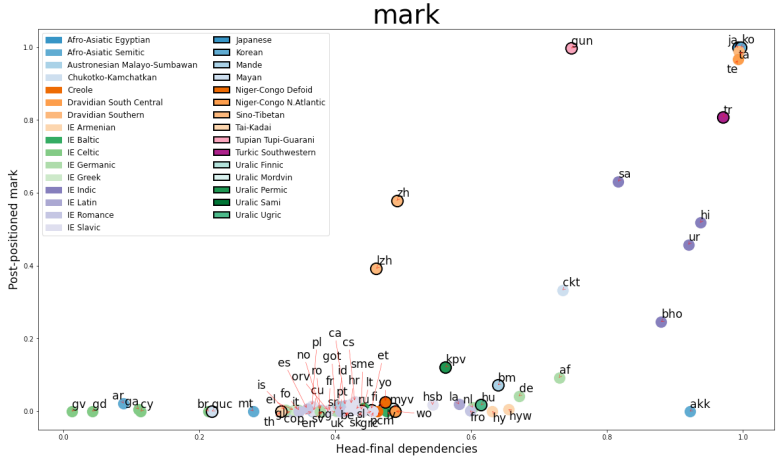


Figure 7: Percentage of post-positioned subordination markers (y-axis) and head-final dependencies (x-axis).

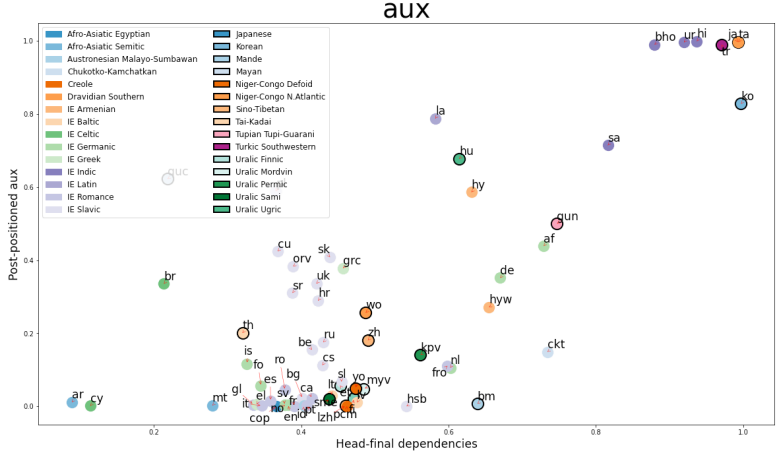


Figure 8: Percentage of post-positioned auxiliaries (y-axis) and head-final dependencies (x-axis).

5. Conclusion

In this article, we first discussed the analysis of function words within the UD framework, arguing that the emphasis on cross-linguistically identifiable constructions for predication and modification, in conjunction with a commitment to lexicalism, leads naturally to an analysis of function words as nucleus-internal elements in the tradition of Tesnière, or as elements of morphosyntactic strategies in the sense of Croft. The decision to formally treat function words as dependents of the lexical core of the nucleus is further motivated by the necessity, for cross-lingual theoretical and computational endeavors, to maximize the parallelism between central dependency structures across languages. After explaining the theoretical assumptions underlying this analysis, we then discussed the specific analysis of seven cross-linguistically identifiable function word relations, those of copulas, auxiliaries, subordination markers, case markers, determiners, classifiers, and coordination markers.

In the second part of the article, we illustrated how the application of the UD annotation guidelines to a broad range of languages enables systematic cross-lingual empirical studies of, among other phenomena, function words. We have investigated the relative frequency of four function word relations across languages (**case**, **mark**, **cc**, and **aux**) and observed considerable variation, which in some cases can be explained by reference to phylogenetic factors. We also explored word order variation specifically with respect to function word relations, as well as in relation to clause level dependency relations, confirming the expected correlation between post-posed function words and head-final dependency relations. Needless to say, these studies have only scratched the surface of what is possible given the rich UD resources, as evidenced by the increasing number of cross-lingual empirical studies based on these resources (i.a., Futrell et al. 2015; Alzetta et al. 2018; Levshina 2019; Yu et al. 2019; Gerdes et al. 2021).

We are aware that the UD analysis of function words is regarded as controversial by some dependency grammarians.[10] Much of the criticism seems to be based on a view of syntax where the main goal is to capture surface-syntactic structure, recognizable through traditional substitution and permutation tests, combined with the as-

[10] For a critique of UD along these lines, see Osborne and Gerdes (2019).

sumption that this structure should be directly reflected in tree-shaped representations, where all parent-child relations represent a single relation of (surface-syntactic) dependency. However, UD does not make either of these assumptions. First of all, the main goal of UD is to capture cross-linguistically relevant constructions of predication and modification, as well as the morphosyntactic strategies used to realize them in different languages. Secondly, the tree-shaped representations employed in UD must be understood as multi-relational, with relation labels crucially indicating what kind of relation is assumed to hold between a parent and a child. Some of these relation labels correspond to central dependency relations like subject and object, but many of them do not. Some are purely technical relations used to analyze phenomena like speech repairs and typographical errors. Others are relations used to encode sequential structures as tree structures, in the analysis of fixed multiword expressions, lists and paratactic constructions, to mention just a few. And some of them are relations used to combine the elements of dissociated nuclei, as discussed in depth in this article. Thus, in UD, the labels on the arcs are more important than the structure of the tree, and the structure of the tree is not designed to capture surface-syntactic structure. Instead it is designed to maximize the number of common central dependency relations across languages, while pushing language-specific realization phenomena towards the leaves of the tree.

The debate over what constitutes an adequate theory of morphosyntax in general, and how to best understand the role of function words, is unlikely to be settled any time soon. However, we hope to have demonstrated, especially through our empirical investigations, that UD can be a useful source of data for cross-lingual investigations even for researchers that do not embrace all the theoretical assumptions of UD. Moreover, it is rewarding to see that UD has inspired alternative approaches to cross-lingual annotation, in particular the Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al. 2018), which exists in a symbiotic relationship with UD, making treebanks available in a format that differs from UD by emphasizing functional heads and surface-syntactic structure. This not only makes it possible for potential users to choose the version that best suits their current purposes, but also enables a continued dialogue on the merits of different approaches to morphosyntax. We look forward to engaging further in that dialogue.

Acknowledgements

Daniel Zeman's, Joakim Nivre's and Marie-Catherine de Marneffe's contributions to this work were supported by grant GX20-16819X of the Czech Science Foundation, grant 2016-01817 of the Swedish Research Council, and a Google faculty research award, respectively. Marie-Catherine de Marneffe is a Research Associate of the Fonds de la Recherche Scientifique – FNRS.

Works cited

1. Alzetta, Ch., F. Dell'Orletta, S. Montemagni and G. Venturi. 2018. Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 4540–4549, Miyazaki, Japan.
2. Andrews, A. D. 2007. The Major Functions of the Noun Phrase. Language Typology and Syntactic Description, Volume I: *Clause Structure*, ed. by Timothy Shopen, 132–223. Cambridge University Press.
3. Aronoff, M. 2007. In the Beginning Was the Word. *Language* 83: 803–830.
4. Bresnan, J., A. Asudeh, I. Toivonen and S. Wechsler. 2016. *Lexical-Functional Syntax*, 2nd edition. Chichester: Wiley-Blackwell.
5. Bresnan, J. and S. A. Mchombo. 1995. The Lexical Integrity Principle: Evidence from Bantu. *Natural Language and Linguistic Theory* 13: 181–254.
6. Chomsky, N. 1970. Remarks on Nominalization. In *Readings in English Transformational Grammar*, ed. by Roderick A. Jacobs and Peter S. Rosenbaum, 11–61. Cambridge: Ginn and Co.
7. Comrie, B. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*. Chicago: The University of Chicago Press.
8. Croft, W. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
9. —. 2002. *Typology and Universals*, second edition. Cambridge: Cambridge University Press.
10. —. 2016. Comparative Concepts and Language-Specific Categories: Theory and Practice. *Linguistic Typology* 20(2): 377–393.

11. Croft, W., D. Nordquist, K. Looney and M. Regan. 2017. Linguistic Typology Meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, 63-75.
12. Dalrymple, M. 2001. *Lexical-Functional Grammar*. Cambridge, MA: Academic Press.
13. de Marneffe, M.-C., B. MacCartney and Ch. D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 449-454.
14. de Marneffe, M.-C. and Ch. D. Manning. 2008. The Stanford Typed Dependencies Representation. In *Workshop on Cross-framework and Cross-domain Parser Evaluation*, 1-8.
15. de Marneffe, M.C., Ch. D. Manning, J. Nivre and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2): 255-308.
16. de Marneffe, M.-C., N. Silveira, T. Dozat, K. Haverinen, F. Ginter, J. Nivre and Ch. D. Manning. Universal Stanford Dependencies: A Cross-linguistic Typology. 2014. In *Proceedings of 9th International Conference on Language Resources and Evaluation*, 4585-4592.
17. Dixon, R. M. W. 2009. Basic Linguistic Theory. Volume 1: *Methodology*. Oxford: Oxford University Press.
18. Dryer, M. S. 2007. *Word Order. Language Typology and Syntactic Description*, vol. I, ed. by Timothy Shopen. Cambridge: Cambridge University Press.
19. Farrar, S. and D. T. Langendoen. 2003. A Linguistic Ontology for the Semantic Web. GLOT International 7: 97–100.
20. Futrell, R., K. Mahowald and E. Gibson. 2015. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics*, 91–100. Uppsala, Sweden.
21. Gerdes, K., B. Guillaume, S. Kahane and G. Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An Annotation Scheme Near-Isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies*, 66–74. Bruxelles, Belgium.
22. Gerdes, K. and S. Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, 131–140.

23. Gerdes, K., S. Kahane and X. Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: A Journal of General Linguistics* 6(1): 17.
24. Greenberg, J. H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In *Universals of Human Language*, ed. by Joseph H. Greenberg, 73–113. Cambridge, MA: MIT Press.
25. Haspelmath, M. 2011a. The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax. *Folia Linguistica* 45: 31–80.
26. —. 2011b. On S, A, P, T, and R as Comparative Concepts for Alignment Typology. *Linguistic Typology* 15: 535–567.
27. —. 2019. Indexing and Flagging, and Head and Dependent Marking. *Te Reo* 62(1): 93–115.
28. Hopper, P. J. and E. Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.
29. Hudson, R. A. 1984. *Word Grammar*. Hoboken NJ: Blackwell.
30. —. 1990. English Word Grammar. Hoboken NJ: Blackwell.
31. Kaplan, R. and J. Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In *The Mental Representation of Grammatical Relations*, ed. by Joan Bresnan, 173–281. Cambridge, MA: MIT Press.
32. Levshina, N. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3): 533–572.
33. Mel'čuk, Igor. 1988. Dependency Syntax: Theory and Practice. New York: State University of New York Press.
34. Milicevic, Jasmina. 2006. A Short Guide to the Meaning-Text Linguistic Theory. *Journal of Koralex* 8: 187–233.
35. Murawaki, Y. 2019. On the Definition of Japanese Word. arXiv:1906.09719 [cs.CL].
36. Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty and D. Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*: 1659–1666.
37. Nivre, J., M.-C. de Marneffe, F. Ginter, J. Hajič, Ch. D. Manning, S. Pyysalo, S. Schuster, F. Tyers and D. Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*: 4027–4036.

38. Osborne, T. and K. Gerdes. 2019. The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies. *Glossa* 4(1): 17.
39. Osborne, T. and D. Maxwell. 2015. A Historical Overview of the Status of Function Words in Dependency Grammar. In *Proceedings of the Third International Conference on Dependency Linguistics*, 241–250.
40. Perlmutter, D. M., editor. 1983. *Studies in Relational Grammar*. Chicago: The University of Chicago Press.
41. Petrov, S., D. Das, and R. McDonald. 2012. A Universal Part-of-speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2089-2096.
42. Pollard, C. and I. A. Sag. 1994. Head-Driven Phrase Structure Grammar. CSLI Publications.
43. Sgall, P., E. Hajičová and J. Panevová. 1986. *The Meaning of the Sentence in Its Pragmatic Aspects*. Dordrecht:Reidel.
44. Spencer, A. and A. R. Luís. 2012. *Clitics: An Introduction*. Cambridge: Cambridge University Press.
45. Sylak-Glassman, J., Ch. Kirov, D. Yarowsky and R. 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), 674–680.
46. Tesnière, L. 1959. *Éléments de syntaxe structurale*. Klincksieck.
47. Tesnière, L. 2015[1959]. *Elements of Structural Syntax*. Amsterdam: John Benjamins. Translation by Timothy Osborne and Sylvain Kahane of Tesnière (1959).
48. Thomas, G. 2020. *Annotation Guidelines for the Mbyá Treebank*. Toronto: University of Toronto.
49. Thompson, S. A. 1997. Discourse Motivations for the Core-Oblique Distinction as a Language Universal. In *Directions in Functional Linguistics*, ed. by Akio Kamio, 59–82. Amsterdam: John Benjamins.
50. Van Valin, Jr., R. D., editor. 1993. *Advances in Role and Reference Grammar*. Amsterdam: John Benjamins.
51. Yu, X., A. Falenska and J. Kuhn. 2019. Dependency Length Minimization vs. Word Order Constraints: An Empirical Study on 55 Treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy), Syntaxfest*, Paris, France.
52. Zeman, D. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference*

on Language Resources and Evaluation, 213-218.

53. Zeman, D., J. Nivre, M. Abrams et al. 2021. Universal Dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague: Czech Republic.