Function Words in Surface-Syntactic Universal Dependencies

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, Guy Perrier
*Université Paris-Saclay, Inria Nancy, Paris Nanterre Université,
Université de Lorraine, France*

1. Introduction

This chapter is devoted to presenting the analysis of function words in SUD (Surface-syntactic Universal Dependencies, Gerdes et al. 2018, 2019, 2021a), the surface-syntactic variant of the UD (Universal Dependencies) annotation scheme (Nivre et al. 2016; de Marneffe et al., 2021, 2022). Both scheme variants aim to encode syntactic properties of words as basic units, both stick to pure tree structures as opposed to graphs that could encode multiple governors, and both aim at proposing universally applicable criteria, backed by an online forum to resolve complex issues in the most general and universal way possible.

In the remainder of this section, we oppose the concepts behind surface vs. deep syntax, content vs. functions words, and SUD vs. UD. Section 2 is devoted to the three sets of criteria at the core of what makes SUD a truly distributional approach to syntactic annotation. In Section 3, we come back to the definition of function words and explain the Tesnièrian term of *translatives* in order to present the actual analysis of different types of function words in SUD: adpositions and subordinators, auxiliaries, and relative pronouns. We conclude Section 3 with a discussion of nuclei, the analyses of which differ substantially across SUD and UD. Section 4 contains a collection of curious cases of function word analysis that deserve further discussion: substantive phrases, relative and interrogative clauses, and, last but not least, coordination. Section 5 sums up the main points of the chapter.

1.1 Surface versus deep syntax

The opposition between surface syntax and "deeper" forms of syntax appears in various syntactic theories. The distinction between deep syntax and surface syntax as we understand it stems from the

work of Igor Mel'čuk in the framework of the Meaning-Text Theory. For Mel'čuk (1988), the deep syntactic structure is intermediate between the surface syntactic structure and the semantic representation, the latter being fundamentally a predicate-argument structure. The need to define two levels of syntactic structure is also present in the Pragian opposition of analytical and tecto-grammatical layers of syntax (Sgall et al. 1986). Similarly, Lexical Functional Grammar (LFG) posits a deep f-structure and a surface c-structure (Bresnan et al. 2015), and Chomsky's Generative Grammar has the notions of deep structure and surface structure (Chomsky 1969). We refer to (Kahane & Gerdes 2022: chapter 13) for a detailed discussion of these notions. SUD differs from UD in that SUD proposes dependency trees close to surface-syntactic analysis. UD is a "deeper" annotation scheme as it uses more semantic criteria; in particular it relies on the distinction between content and function words to establish dependency relations.

1.2 SUD

The Surface-syntactic Universal Dependency scheme (SUD) pursues an analysis that is purely syntactic, excluding semantic criteria whenever possible. As exemplified by the Vauquois triangle (Vauquois 1968) where languages differ less in semantics, a surface-syntactic approach results in greater differences across languages than a deeper approach as exemplified by the classical UD scheme. This allows for finer-grained typological measures as some distinctions between languages may only be visible on the surface-syntactic level and can be blurred in deeper structures.

The SUD scheme is designed to be near-isomorphic to UD by means of a rule-based conversion system that goes both ways (Gerdes et al. 2018). The compatibility with UD is an important constraint in the development of the SUD annotation scheme. Every SUD structure must be automatically convertible into a valid UD annotation, and inversely, even if some information is lost in the process, we also want to be able to convert every UD structure into a SUD-conform structure, Three cases where SUD is more informative than UD are developed in this paper: multiple transfer in Section 3.2; internal structure of idioms in Section 3.3; and free relative clauses in Section 4.2. To keep the conversion simple, SUD uses the same POS tagset, but includes the *ExtPos* (external part of speech) feature to capture

functional reassignment of multi-word expressions in the SUD to UD conversion (see Section 3). SUD introduces new relations for the most frequent dependencies while sharing with UD the remaining relations, and proposes different rules of establishing syntactic relations, based on distributional criteria.

SUD is motivated by two converging needs for our syntactic annotation projects:

1. The heuristic verification of many of the classical typological language universals require measures on syntactic data based on an annotation scheme where the notion of "head" follows the established notions (Gerdes et al. 2021b). In particular, so-called *head-initial languages* should have most governors preceding their dependents, and, vice-versa, *head-final languages* should have most governors following their dependents. Put simply, a prepositional phrase (PP) bears this name for distributional reasons to be reviewed in Section 2.

2. Teaching syntax by means of treebanks requires an annotation that follows established notions wherever possible. Important deviations from the standard analysis complicate significantly not only the integration of treebank-based heuristic syntax classes and treebank annotation projects into a linguistic curriculum, but also any scientific interaction and interest in treebank annotation beyond the circle of computational linguists. The fact that one annotation is convertible into another does not convince a classical syntactician that both analyses are equally correct.

It should be pointed out though that SUD is merely an annotation scheme, not a syntactic theory and even less syntactic truth. It is a compromise between a myriad of constraints (Gerdes & Kahane 2016) building on the wealth and the experience of the UD project. First and foremost, it is an operational annotation scheme that fosters ease of annotation for human annotators, trained as linguists. Nevertheless, as with any annotation scheme, SUD is based on a theoretical framework that will be presented in this article and contrasted with the theoretical framework underlying UD.

1.3 Content and function

The difficulty of defining content and function words has been raised in the introduction and elsewhere. Suffice it to say that the definitions most often refer to the notion of *meaning*, and the presence or absence thereof—a notion that is rarely part of syntactic reasoning but rather a part of lexical and semantic terminology. As we will develop further in Section 3.1, the invention of the term *function word* is often attributed to Fries (1952). Beyond their description as words that do not have an autonomous meaning, independent of the context in which they are used and of the structures in which they participate, Fries also proposes quantitative measures: The class of function words has a very small number of words that are very frequent in the corpora. These notions and numbers are not readily available when analyzing a sentence.

The opposition between content word and function word, foundational for the UD scheme, creates much hesitation among annotators as evidenced by many discussions in the UD forum that can be traced back to the difficulty of establishing this contrast. As a glaring example, consider the many multi-word expressions that are in the distributional paradigm of function words: the complex prepositions *in accordance with*, *in order to*, *as opposed to*, *on behalf of*, *out of*; the complex determiner *a lot of*, or the coordinate conjunction *as well as*. As a whole, these expressions are "function words"; as individual tokens, they may or may not be content words. Depending on the "wordness", i.e., the degree of cohesion of the multi-word expression, parts of the expression may be separated or coordinated independently, creating gaps in the resulting analysis (also called *catastrophes* in the strictly mathematical sense of Thom's catastrophe theory, cf. Saunders 1980 and Gerdes & Kahane 2016).

The annotator has to cut into the continuum between compositionality and idiomaticity to set heads (content words) apart from dependents (function words). This could be solved by a list of multi-word lexemes that trigger a different analysis than the compositional words, but this is neither language independent nor operational in a real-world treebank project setting, as the annotator constantly has to consult such a list (which, moreover, is difficult to establish on purely linguistic criteria, see Savary, Cordeiro, & Ramisch 2019 for details).

However, going the SUD way and annotating tokens based on their syntactic distributions only results in a more homogeneous

annotation with fewer catastrophes; i.e., fewer breakpoints amidst syntactically similar structures.

2. Distributional criteria

Following Mel'čuk (1988), Gerdes & Kahane (2013), Kahane & Gerdes (2022), we consider that the syntactic structure must be defined in three steps using three sets of criteria.

- Criteria A define the syntactic units, that is the text fragments we consider as relevant from a syntactic point of view (which are more numerous than the constituents of phrase structure grammars).
- Criteria B decide what is the head for each syntactic fragment.
- Criteria C are used to categorize the sets of dependencies and to decide which ones must be grouped and labeled with the same relations.

With criteria A and B, we obtain a syntactic structure that is not necessarily a tree. We will say only a few words about criteria A and C and focus on criteria B instead, which are the most relevant for the the analysis of function words.

Our methodology will be illustrated by the following example from Naija, a pidgin creole of English from Nigeria spoken by more than 100M people (Caron et al. 2019).

(1)   you   sabi   di   people   wey   dey   play?
      you   know   the   people   that IPFV   play
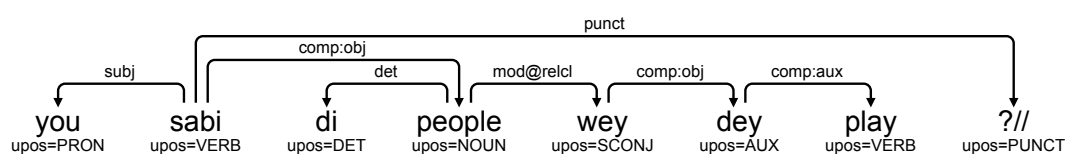   'Do you know the people that are playing?'

Figure 1- The SUD dependency tree for Example (1)

2.1 Criteria A

A fragment X of an utterance U is a syntactic unit if X is a sub-sign of the sign U, that is, if both the form and the meaning of X are a sub-part of the form and the meaning of U.[1] Moreover, X must combine freely with the rest of U, which means that X belongs to a regular paradigm of units that can commute with X. The semantic contribution of X must be the same in U as in other contexts. In particular, if X can stand alone (that is, form a speech turn on its own), it must have the same semantic contribution when used alone as in a text.[2]

Applying this principle to (1), we observe that the relative clause *wey dey play* or the substantive phrase *di people wey dey play* are syntactic units of (1).[3] The same holds for *di people* or *you sabi*, which are not considered as constituents in traditional phrase structure grammars. There are also fragments that cannot stand alone, such as the words *di* or *sabi* or the so-called VP *sabi di people wey dey pay*.

2.2 Criteria B

The criteria B, which identify the head of a unit, will be presented in detail. The situation is as follows: we consider three syntactic units U, A, and B such that U = AB or U = BA. This means that A and B combine together to form U, and we can postulate a connection between A and B. To convert this connection into a dependency between A and B, we must decide which of the two is the head of U. Note that it is equivalent to say "A is the head of AB", "A governs B", or "B depends on A".[4]

The main criteria we use for the identification of the head of a unit are distributional criteria. What we call the *distribution* of a unit X is the set of syntactic positions that X can occupy, also named the

---

[1] A fragment of U can be a discontinuous part of U. For instance, *the dog* is a sub-sign of *the white dog*.

[2] We say that a unit X can stand alone when X can be used as a speech turn, such as *two* (*How many siblings do you have? – Two*.) or *we can* (*Can we do that? – We can*). Finite verbal forms cannot stand alone in English (*wants, was*), as well as most function words (*the, to, if*).

[3] We use the term *substantive phrase*, a neutral term that presupposes neither the noun or the determiner as the head of the phrase.

[4] We do not restrict the notion of dependency to the connection between two words. A connection can also link units that have an internal structure, composed of several words: If A combines with B and $A = A_1A_2$, the connection between A and B can be refined into a connection between A_1 and B or A_2 and B (Gerdes & Kahane 2013).

passive valency by Mel'čuk (1988:114) and the *superior valency* by Kahane & Gerdes (2022). The distribution was first defined by Bloomfield (1933), under the name of *form-class* and defined in the following terms:

> "Certain English words and phrases can appear in the actor position, certain others in the action position. The positions in which a form can appear are its *functions* or, collectively, its *function*. All the forms which can fill a given position thereby constitute a *form-class*." (*ibid*: §12.2)

In dependency grammar, the distribution of X can also be described as the set of possible governors for X, the function of X being its relation with its governor.

As noted by Mel'čuk (1988:115), the need to know where the governor of the unit under investigation is located opens up an apparent circularity in the definition of *head*: The head notion is based on the notion of governor, which is equivalent to the notion of head. In fact, there is no paradox in this definition because we can start with a particular case which does not involve a governor strictly speaking. This is the case of the head of the sentence, the root of the dependency structure. A sentence is an illocutionary unit, which gives us the following criterion.[5]

> Criterion for the head of a sentence
> The head of a sentence is the element that bears the illocutionary force and will be affected if the illocutionary force is modified.

Let us see how this criterion applies to English. English has special constructions for a question or a negation involving a particular position of an auxiliary verb. In (2), *came* is replaced by *did ...come* when the assertion becomes a question. According to the previous

---

[5] Mel'čuk (1988:115) does not give any criterion, considering the question obvious. He writes:

"this circularity is eliminated because the concept of passive SS-valency includes the capacity of the unit in question to be the absolute head of an utterance. This capacity is postulated in this book for the finite verb. The dominant syntactic role of the finite (i.e., tensed) verb is intuitively evident. Taking it for granted, one proceeds by induction, the finite verb constituting the induction basis. The other cases of syntactic dependency are reduced, sometimes in several steps, to dependency on the finite verb, that is, on the grammatical predicate of the sentence."

criterion, this shows that the verb bears the illocutionary force and is affected when the illocutionary force changes. Moreover, the negation is carried by the auxiliary verb, indicating that it is the head.

(2) a. Peter came.
     b. **Did** Peter come?
     c. Peter **didn't** come.

Note also that in many cases, the question of the governor is not controversial. For instance in *Mary talks to Peter*, it is clear that *talks* is the governor of U = *to Peter* and the only question is whether *to* or *Peter* is the head of U.

   We will consider three distributional criteria for the choice of the head. When deciding which part of a combination U = AB is the head, the simplest case is when A and B can stand alone, that is, when A can be used without B and B without A. If A and B can stand alone, it is possible to study the distributions of A and B independently.

> Positive distributional criterion with deletion
> If U = AB, A can stand alone (that is, B can be deleted), and U and A have the same distribution, then A is a head of U.[6]

If B can also stand alone and the distribution of B is different from the distribution of U we can state that A is *the* head of U, but the criterion does not exclude the possibility of both A and B being heads of U. For practical reasons (simplicity of annotation, compatibility with most tools including UD, etc.), SUD adopts a tree structure and needs to choose a head, which means that we must decide which of A or B fulfills the criteria best.

   Let us consider a simple example illustrating how the criterion works. Consider U = *di people wey dey stay*, A = *di people*, and B = *wey dey stay*. As U = AB, A and B are connected. Clearly, U and A have the same distribution and B has a different distribution. We deduce a dependency A → B, where A is the governor and B the

---

[6] We should say *roughly* the same distribution. As Hockett (1958:184) put it: "Some constructions are such that the form-class of the constituents is similar to the form-class of at least one of the immediate constituents. Here "similar" means that the two ranges of privileges of occurrence largely overlap. The grammarian would prefer to speak of identity rather than similarity, but languages are not completely tight-knit, and greater precision would be spurious."

dependent. This dependency can be refined, in particular by identifying the head of B.

The positive distributional criterion with deletion was first stated by Bloomfield (1933: §12.10):

> "Every syntactic construction shows us two (or sometimes more) free forms combined in a phrase, which we may call the *resultant* phrase. […] In subordinative endocentric constructions, the resultant phrase belongs to the same form-class as one of the constituents, which we call the *head*: thus, *poor John* belongs to the same form-class as *John*, which we accordingly call the head; the other member, in our example *poor*, is the attribute. The attribute may in turn be a subordinate phrase: in *very fresh milk* the immediate constituents are the head *milk*, and the attribute *very fresh*, and this phrase, in turn, consists of *ranks* of subordinative position: in *very fresh milk* there are three: (1) *milk*, (2) *fresh*, (3) *very*."

The limits of this criterion were also pointed out by Bloomfield (1933: § 12.20), who stated that some phrases do not have a head:

> "The resultant phrase may belong to a form-class other than that of any constituent. For instance, *John ran* is neither a nominative expression (like *John*) nor a finite verb expression (like *ran*). Therefore, we say that the English actor-action construction is *exocentric*: the resultant phrase belongs to the form-class of no immediate constituent."

But there is one problem in Bloomfield's reasoning: We do not know how the distributional class of *ran* can be compared with the distributional class of *John ran*, due to the fact that *ran* cannot stand without a subject. In other words, the distribution of a unit can only be considered if this unit can stand alone. This leads us to the two following criteria.

Our second distributional criterion can be applied to a head that cannot stand alone, provided that its obligatory dependent can stand alone.[7]

---

[7] The criterion can be reformulated as follows:

Variant of the negative distributional criterion with deletion. If B can stand alone and A is a dependent of B, the combination of A with B cannot change the distribution of B.

> Negative distributional criterion with deletion
> If U = AB, B can stand alone, and U and B do not have the same distribution, then A is a head of U.

The Negative distributional criterion with deletion can be applied to Bloomfield's example of U = *John ran*. As B = *John* can stand alone and does not have the same distribution as U, then A = *ran* is the head of U. The fact that the distributions of U and B are very different clearly indicates that A plays an important role in the distribution of U. As with all our criteria, the Negative criterion with deletion is gradual: The more the distribution of U and B are different, the more A is likely to control U's distribution and have head properties.

We can also extend the Negative criterion for applying it to the internal structure of the relative clause U = *wey dey play* in example (1), with A = *wey* and B = *dey play*. B cannot really stand alone, but it can be used without A, as soon as a subject is added. Without A, B has a completely different distribution from U and cannot modify a noun. In consequence, *wey* is analyzed as the head of the relative clause and categorized as a subordinating conjunction (SCONJ). This analysis for Naija, which uses only the relativizer *wey,* will be contrasted with the case of relative pronouns in English in Section 4.

There is a third distributional criterion that does not refer to the distribution of A and B and does not need to hypothesize that A or B can stand alone.

> Distributional criterion without deletion
> If U = AB, A can commute with an A' and U and U' = A'B does not have the same distribution, then A is a head of U.[8] In

---

The equivalence of the two variants can be done in two steps: first "if p and q, then r" can be replaced by "if p, then [if q, then r]" and then by "if p, then [if no r, then no q]", which, starting from the Negative distributional criterion with deletion, gives us "If U = AB and B can stand alone, then [if A is not a head of U (i.e., A is a true dependent of B), then U and B must have the same distribution]."

[8] When saying that A' can commute with A, we are only considering the commutation in the context of B. In other words, this means that A'B is a valid combination and that A and A' exclude each other in this context (i.e. AA'B is not valid). One of our reviewers pointed to the following problem: In Russian, the copula is omitted in the present tense:

(i) моя мама   учительница математики
   moja mama učitel'nica   matematiki
   my   mother teacher   math.GEN
  'my mother is a math teacher'

One could replace моя мама *moja mama* with, say, an adjective, creating a

other words, if B depends on A, then B must not modify the distribution of A and a commutation on B does not change the distribution of the unit it forms with A.

To our knowledge, this criterion was first stated by Garde (1977:8).[9] A particular case is Zwicky's (1985) Morphosyntactic locus criterion: the head of a unit U is "the constituent on which inflectional marks will be located. […] One way in which a constituent can 'characterize' a construct is that it can be the bearer of the morphosyntactic marks of syntactic relations between the construct and other syntactic units." This argument is used to choose the verb as the head of the clause (rather than the subject) and the auxiliary as the head of a complex verbal form (rather than the lexical verb), because the verb, or its auxiliary in the case of complex forms, carries the mode imposed by the governor of the clause. And a change in the mode changes the distribution of the clause. Zwicky (1985:7) himself extended the Morphosyntactic locus to constructions not involving inflection and which fall under the Distributional criterion without deletion:

> English P + NP has no actual inflection locus: the NP does bear the marks of person and number, but person and number play no role in the distribution of P + NP constructs; and P in English is uninflected. However, there is a phenomenon long recognized as analogous to case inflection here, namely the choice of one preposition over another in constructions like *inform Sandy of the news* and *tell the news to Sandy*, where the prepositions *of* and *to* mark particular syntactic arguments of the verbs. The analogy suggests that P is the morphosyntactic locus in P + NP.

The Negative distributional criterion *with* deletion can be seen as a borderline subcase of the Distributional criterion *without* deletion. Indeed, when B can stand alone and A is deletable, it means that A

---

substantive phrase with a very different distribution from the original sentence: хорошая учительница математики *horošaja učitel'nica matematiki* (good teacher math.GEN). Does it follow from the Distributional criterion without deletion that моя мама *moja mama* 'my mother' is the head of the non-verbal sentence (i)? The answer is no, because the subject моя мама *moja mama* of (i) does not commute with an adjective. Commutation implies exclusion. And both elements can cooccur in a sentence such as Моя мама хорошая учительница математики *moja mama horošaja učitel'nica matematiki* 'my mother is a good math teacher'.

[9] Garde's paper, written in French by a Slavist, remains quite confidential, but is nevertheless a seminal paper.

has the "empty unit" in its positional paradigm. The Distributional criterion without deletion says that A is the head if at least one commutation on A changes the distribution. If the deletion of A is seen as a commutation with emptiness; the fact that B does not have the same distribution as AB means that at least one commutation on A changes the distribution.

Most of the time, our criteria coincide to give clear results. In Section 4, we will discuss some constructs where the criteria do not yield immediate results and must be assessed against each other to make a choice of analysis.

2.3 Criteria C

Criteria C for categorizing syntactic dependencies will not be developed much here (see Mel'čuk (2021) for a general discussion on the criteria and Gerdes et al. (2018) or the SUD website https://surfacesyntacticud.github.io/ for syntactic relations in SUD). Phrases that occupy the same position receive the same syntactic function, even if they have different POS. For instance, a unique relation *subj* is used for both *that you came* and *your arrival* in (3) because they commute and exclude each other, which means that they occupy the same position.[10]

(3)   a. That you came surprised me.
      b. Your arrival surprised me.

Constructions that have similar syntactic properties will be labeled with the same relation. Moreover, for the sake of simplicity of the annotation scheme, we must try to avoid having too many syntactic relations. As a consequence, we decided to use the same *comp:obj* relation for the direct object of verbs and for the direct complement of adpositions and subordinating conjunctions. Some additional facts justify merging these different constructions.

First, some adpositions are grammaticalized forms of transitive verbs in some languages: En. *regarding, considering,* Fr. *durant* 'during', lit. lasting, *vu* 'in view of', lit. seen, De. *während* 'during',

---

[10] UD uses two different relations here, *nsubj* for a substantive phrase and *csubj* for a clause. Similarly, a direct object in UD can be *obj*, *xcomp* or *ccomp* according to its POS, while it will always be *comp:obj* in SUD. See the UD webpage https://universaldependencies.org/u/dep/index.html for a precise description of every relation.

lit. lasting, Zh. 到 dào 'to/arrive', 给 gěi 'for/give', 在 zài 'at/be'. Second, complements of adpositions can share properties with direct objects of verbs in some languages. For instance, in English, indirect object, which are complement of the preposition in the SUD analysis, can be coordinated with the direct object of a verb or passivized:

(4)   a. Peter met and talked to Mary.
       b. Mary was talked to by Peter.

Finally, in some languages, some words can be both adpositions and subordinating conjunctions, such as En. *since, before*, Zh. 自从 zìcóng 'since'.

(5)   a. since he left
       b. since his departure

(6)   a. 自从   他 离开
          zìcóng   tā líkāi
        'since he left'

       b. 自从   秋天
          zìcóng qiūtiān
        'since autumn'

3. SUD analysis of function words

After a general analysis of the notion of function word, we will focus on particular cases that highlight the specificity of the SUD analysis: translatives, adpositions and subordinators, auxiliaries, and finally, pronouns. We conclude the section by a comparison with UD in case of interaction between multiple function words.

3.1 The general notion of function word

As stated in Section 1.3, the notion of *function word* was first introduced by Fries (1952). In a distributional approach to syntax, he uses *test frames* to classify words into different POS and to distinguish *function words* from *content words*. A test frame is a minimal sentence where some positions are replaced with POS, and a specific

position is a hole dedicated to receive the words to be checked. With this method applied to English, he defines four open classes, which roughly correspond to nouns, verbs, adjectives, and most adverbs, that he calls the content words, and fifteen closed classes, that he calls the function words.

Besides the fact that they belong to closed classes, function words appear more frequently than content words in corpora and have a common semantic property: it is impossible to express their meaning independently of the syntactic structure in which they participate, but they contribute to the structural meaning of the phrase in which they participate. In order to illustrate his purpose, Fries gave the following example, where the form of the auxiliary determines the diathesis of the verb *given*: passive with *was* and active with *has*.

(7)   a. The boy was given the money.
        b. The boy has given the money.

Even if we agree with this characterization of function words, the reality is often more complicated than a binary division between function words and content words. Some words have properties that make them similar to function words and other properties that make them similar to content words. This is for example the case of light verbs. Consider the following example.

(8)   Peter took a shower this morning.

We can consider *took* as an ordinary lexical verb, with two arguments, but, as it has little semantic content, we can consider it as a function word that verbalizes the lexical word *shower*, which gives the semantics of the sentence predicate. Corver & van Riemsdijk (2001) qualify the categories of such words as *semi-lexical categories* and studied them in three domains: nominal, verbal, and adpositional, following a generativist approach. Osborne & Gerdes (2019) go even further, arguing that "the distinction is, rather, more accurately captured in terms of a continuum, whereby prototypical function words and content words appear at opposite ends of the continuum, non-prototypical cases appearing somewhere on the continuum in-between." If one agrees with this point of view, the consequence is that the connections of the function words with their context must be represented in the same way as the connections of the content words.

Thus, if a function word satisfies the same distributional criteria as a content word in its relation to the environment, this is expressed by the same dependency with the same syntactic relation. It is not important, from the SUD perspective, to draw a precise boundary between function words and content words, especially since such a goal seems to be impossible to achieve. We apply our distributional criteria to all words in the same way, even for words that fall in the gray area between function words and content words.

3.2 Translatives in SUD

Beyond the words that are generally considered as function words, there is one big family that Tesnière (1959[2015]) called *translatives*. A translative is a syntactic unit (generally a word, but it can also be an inflectional morpheme or, on the contrary, a multiword expression) that transfers content words from one category to another in order to expand the syntactic functions they can perform. Typical translatives are adpositions, subordinators (subordinating conjunctions) and auxiliaries. Words that have only a translative function are called *pure translatives* (*ibid.*: Chapter 267, §11) (and are then typical function words), but many translatives can also bear a (strong) semantic content, as is the case of most adpositions and subordinating conjunctions (*before, after, behind, because, since,* etc.). Consider the following example.

(9)   the house of my father

The preposition *of* is a pure translative that transfers the nominal *my father* into a phrase that can modify a noun and is then in the paradigm of adjectives. Such a word is called a translative of nouns into adjectives by Tesnière.

Translatives, because they modify the distribution of the phrase they combine with, are considered as heads in SUD. Indeed, according to the Negative distributional criteria with deletion, if B can stand alone and the combination of B with the translative A has a different distribution, then A is the head.

Let us check this regarding the three typical translatives: adpositions, subordinators, and auxiliaries.

3.3 Adpositions and subordinators

Adpositions as heads. In SUD, adpositions are heads of adpositional phrases. Let us explain why on the example below, which is extracted from the corpus SUD French-GSD@2.8.

(10) Les chambres d' internat   ressemblent
     The rooms   of boarding school   resemble
     à de   grandes cellules.
     to IND.PL large   cells.
    'The boarding rooms look like large cells.'

The prepositional phrase *d'internat* does not have the same distribution as the noun *internat*. For example, *les chambres internat* is not acceptable. Therefore, the head of *d'internat* is the preposition *d'*, according to the Negative distributional criterion with deletion. With the same reasoning, we conclude that the head of the prepositional phrase *à de grandes cellules* is the preposition *à*. For this last case, we can also use the Distributional criterion without deletion: if we replace the preposition *à* by another preposition, the prepositional phrase can no longer be the complement of the verb *ressembler*. In other words, the verbal form *ressemblent* governs the preposition *à* and it must therefore be linked directly to it. Figure 2 gives the SUD annotation of the sentence (10), with the relevant dependencies in italics.
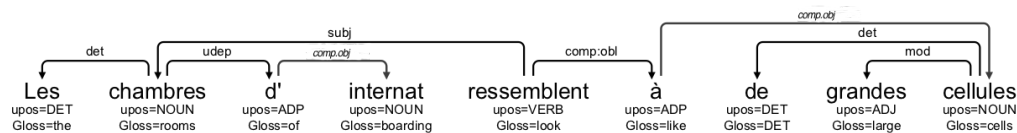

Figure 2 - SUD annotation of Example (10)

Preposition stranding. The choice of prepositions as heads of prepositional phrases is compatible with several phenomena, such as preposition stranding, which is mainly found in English. The object of a preposition is detached and appears at the beginning of the clause containing it, while the preposition remains in its normal position.

Let us consider the following example extracted from the corpus SUD_English-GUM@2.8.[11]

---

[11] The corpus SUD_English-GUM@2.8 is an automatic conversion of UD_English-GUM@2.8. All UD treebanks are available on the SUD website and can be queried on http://match.grew.fr/ with Grew-match (Guillaume 2021).

(11) What are they fighting for?

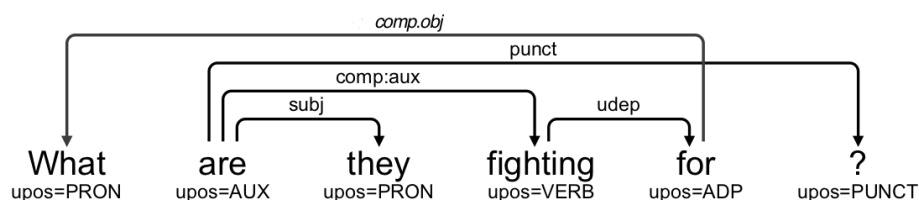The SUD annotation of this sentence is given Figure 3.

Figure 3 - SUD annotation of Example (11)

If we had assumed that in the connection between the preposition *for* and the interrogative pronoun *what*, the pronoun is the head of a dependency with *for* as its dependent, this would violate the general rule that in the detachment of a linguistic unit, all dependents of the head move with it.[12]

Words that behave both as adverbs and as prepositions. Another argument for prepositions as heads of prepositional phrases concerns words that are adverbs in some contexts and prepositions in other contexts. This is the case for the word *before*. Consider the following examples.

(12) a. John is coming before.
       b. John is coming before his brother.

In the first sentence, *before* is an adverb and in the second sentence, it is a preposition. In SUD, there is a *mod* dependency from *coming* to *before* in both sentences, because *before* in the second sentence is the head of the prepositional phrase *before his brother*. If we had assumed that the head of this prepositional phrase was the content word *brother* (as done in UD), there would be an inconsistency between the analyses of the two sentences, only the first sentence containing a direct link from *coming* to *before*. On the opposite assumption, the application of criteria B entails a continuity of analysis

---

[12] Some dependent phrases can stay in the canonical position, as the relative clause in (i). But the fact that a dependent function word stays alone is less expected, because such words have less autonomy (they can generally not stand alone, they cannot form a separate prosodic unit, they do not have an autonomous semantic interpretation).
 (i) Qui as-tu   invité   que je connais ?
     'Who have you invited that I   know?'

across (12a) and (12b), by considering that *before* is the head of the prepositional phrase.

Subordinators as heads. The reasons for choosing subordinators as heads of subordinate clauses in SUD is the same as for adpositions. Let us consider the following example extracted from the corpus SUD_English-GUM@2.8.

(13)   You could just scoop them out when the color dissolves.

The subordinating conjunction *when* is a translative that transfers the clause *the color dissolves* into the adverbial clause *when the color dissolves*. The distribution of *the color dissolves* is very different from the distribution of *when the color dissolves*. The Negative distributional criterion with deletion entails that the head of the subordinate clause is *when*.

Even for the tricky case of the subordinating conjunction *that*, we can notice that, although *that the color dissolves* and *the color dissolves* can commute in some positions, there are positions where they cannot, such as the subject position: *That the color dissolves is a surprise*. In consequence, *that* is also the head of the clause it introduces.

As for prepositions, the dependency from the subordinator to the main verb of the subordinate clause is labelled with *comp:obj*. Hence, the SUD annotation of Example (13) on Figure 4.
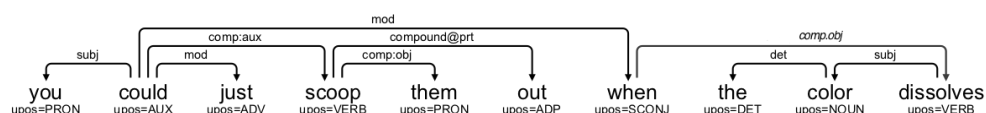

Figure 4 - SUD annotation of Example (13)

Some particles play the role of subordinators of infinitives (*to* in English, *zu* in German). They are labeled PART and not SCONJ in UD and SUD, but for the same reason as other subordinators, they are heads of the infinitives they subordinate.

Prepositional and conjunctive idioms. A way of dealing with prepositional and conjunctive idioms is to consider them as fixed prepositions and conjunctions and then apply the same principles as for simple prepositions and conjunctions. But their presence in coordination requires another representation.

Consider the following simplified example extracted from the corpus UD_Polish-PDB@2.8

(14) *ze   względu na* brak   środków,
       from due   to   a.lack of.resources,
       ale *na* niedobór siły   roboczej
       but to   a.shortage of.force labor
       'due to a lack of resources, but to a shortage of labour'

If we consider *ze względu na* as a fixed prepositional idiom (as done in UD), we cannot represent the second preposition *na* as a repetition of the first one.[13] Our solution is to consider that *ze względu* is an adverbial idiom with an argument introduced by *na* in its valence. Hence, the SUD annotation of (14) below.
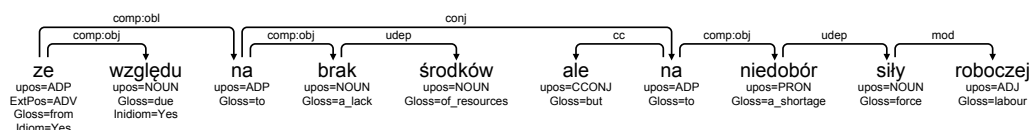
Figure 5 - SUD annotation of Example (14)

Figure 5 requires explanations on how to annotate adverbial idioms and more generally idioms in SUD. In order to distinguish the internal structure of an idiom from its participation in the syntax of the sentence, the head carries a feature *ExtPos* (for external POS) with the POS of the idiom considered as an atomic unit of the syntax. In our example, since *ze względu* is considered as an adverb, *ExtPos=ADV*. Another feature *Idiom=Yes* is placed on the head word of the idiom. This second feature is necessary because we use the same mechanism in a more general way to annotate expressions that have an internal structure that is not integrated in the syntactic structure of the sentence but marks a break with it. This is the case for titles: their heads are marked with the features *ExtPos=PROPN* and *Title=Yes*. Whether they are idioms or titles, it is necessary to distinguish the nodes internal to their structure. For idioms, they are marked with the feature *InIdiom=Yes*. In our example, there is only one internal node, *względu*, because *na*, which is repeatable as seen in (14), is analyzed as the régime of the idiom.

---

[13] The internal structure of idioms is missing in the UD annotation, where all elements of the expression are linked to the first word by a fixed relation. Note that the conversion into SUD will give us an underspecified structure from the viewpoint of SUD.

3.4. Auxiliaries

Auxiliaries as heads. In UD, an auxiliary verb (AUX POS) is a copula or a word that is attached to a full verb in order to express tense, aspect, modality, or voice (see UD webpage https://universaldependencies.org/u/pos/AUX_.html). Although the definition of auxiliaries seems very semantic, the idea is to unite in this category words whose behavior depart from ordinary verbs or adverbs. For French, we have only kept three grammaticalized verbs in the category of auxiliaries: the copula ê*tre*, the tense auxiliaries *avoir* and *être*, the passive auxiliary *être*, and the causative auxiliary *faire*. These three verbs share one unique property, namely clitic climbing as illustrated by (15).[14]

(15) Ce livre lui   *a   été*   donné par sa fille.
      this book to.him has been given by his daughter
    ‘This book was given to him by his daughter.’

In English, the list of auxiliaries must be extended to the whole paradigm of modal verbs, which share placement properties in questions and negations.

According to Fries' (1952) definition of function words, as well as UD choices (de Marneffe et al., 2021, 2022), auxiliaries are function words. The problem is that it is difficult to draw a clear line between auxiliaries and content verbs. In the intermediate area, we find semi-auxiliaries and light verbs. As we will show below, this is not a problem for representation in SUD. Consider the very simple French example.

(16) a. Il fait rire.
       ‘He makes (people) laugh.’
    b. Il aime rire.
       ‘He likes to laugh.’

The first verb *fait* is usually considered a causative auxiliary, while the second verb *aime* is considered a content verb. The auxiliary is the head because it bears the illocutionary force and is the morphosyntactic locus (Distributional criterion without deletion). For instance, it shows subjunctive mood if the clause is subordinated to

---

[14] Marginally, other verbs allow for clitic climbing such as *laisser* 'let' and *entendre* 'hear'.

a verb requiring the subjunctive, e.g., *J'ai peur qu'il* fasse *rire* 'I'm afraid he will make (people) laugh'. Moreover, the words *il* and *rire* are not connected according to Criteria A; *il rire* is not an acceptable syntactic unit. Therefore, there is a dependency in SUD from *fait* to *rire*, labeled *comp:aux@caus*.[15] The same reasoning identifies a dependency from *aime* to *rire* in (16b). The only difference with (16a) is that the dependency is labeled *comp:obj*. The choice of different relations is justified by differences in the behavior of the two constructions, such as regarding clitic climbing.

What is important is that the two annotations are homogeneous, which would not be the case if one followed the UD annotation in assuming that the auxiliary depends on the content verb. In SUD, the only difference between (16a) and (16b) is in the dependency labels, while in UD there is a difference in the dependency structures because auxiliaries depend on main verbs in UD.

Copula as head. The same distributional criteria as for auxiliaries lead us to consider the copula as head with respect to its complement. Consider the following example.

(17) John is a popular teacher.

In its combination with *a popular teacher*, the copula *is* is considered as the head because of the Negative distributional criterion with deletion: The noun phrase *a popular teacher* has a distribution, which is totally different from the distribution of *is a popular teacher*. See Note 8 for the case of predicative constructions without copula. While UD aims at having similar analyses for predicative constructions with or without the copula (in both cases, the subject depends on the predicate), SUD on the contrary contrasts the two constructions (the subject depends on the copula if there is one, and on the predicate if there is not).

Coordination of a full verb with an auxiliary or a copula. If an auxiliary or a copula is not considered the head of the syntactic unit it forms with its complement, and if it is coordinated with a full verb, any annotation of the resulting syntactic construction is inconsistent. Consider the following example extracted from the corpus SUD_French-GSD@2.8.

[15] In SUD, an extension in the name of relation introduced with the @ symbol indicates a property of the deep syntax

(18) Andy Kaufman *fut* et *reste* une énigme.
       'Andy Kaufman was and remains an enigma.'

In French, *fut* 'was' is a copula and *reste* 'remains' a full verb. The predicative complement *une énigme* is shared by the two verbs; it therefore forms the syntactic unit *fut une énigme*. Let us suppose that the head of this unit is *une énigme*. On the other hand, this noun phrase forms the syntactic unit *reste une énigme* with the verb *reste*, which is the head of this unit. It is therefore impossible to find an annotation that reconciles the two requirements.

On the other hand, if the copula *fut* 'was' is considered to be the head of the unit it forms with its complement *une énigme* as in SUD, then the dependency between the two elements is similar to the dependency formed by the full verb *reste* with its complement énigme; in both cases, we use the same relation *comp:pred* for the predicative complement. So, the coordination of *fut* with *reste* can be annotated without problem, as shown in Figure 6.[16]
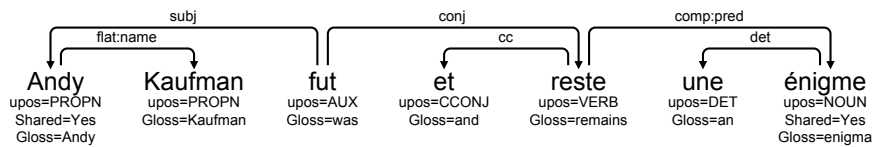

Figure 6 - SUD annotation of Example (18)

In SUD, dependents shared by the conjuncts of a coordination are attached to the closest conjunct (Gerdes et al. 2018). Thus, *une énigme* is attached to *reste*, but the feature *Shared=Yes* indicates that this complement is dependent on the whole coordination.[17]

Predicative clauses with copula. This construction raises two issues simultaneously: the choice of a head in the relations of auxiliaries with content verbs and the choice of a head in the relations of subordinators with the head of the clauses they subordinate. If in both cases we use the content word as the head, we arrive at aberrant annotations. Let us take an example from the corpus SUD_English-PUD@2.8.

---

[16] We postulate that an element that can be shared should have the same syntactic function towards its two potential governors. That was one of our arguments, based on (4a), to consider that the direct object of a verb and the complement of a preposition should have the same function *comp:obj*.

[17] There are other ways to indicate the dependent sharing in coordination. Tesnière (1959[2015]: Chapter 143) attempted to link all conjuncts to the shared dependent, but the tree structure is lost.

(19) Her reason is that it provides insufficient protection to American workers.

Assume for a moment that the copula *is* depends on the head of its complement, the subordinate clause *that it …workers*. Therefore, the head of the sentence in our example is the head of the subordinate clause. Further, assume that subordinators depend on the main verbs of the clauses they subordinate. In our example, the verb *provides* is then the head of the subordinate clause, and because of the first assumption, it is also the head of the whole sentence. From a distributional point of view, this is quite aberrant. Moreover, it implies two subjects for *provides*: the subject of the main clause *reason,* and the subject of the subordinate clause *it*.[18]

On the opposite assumption of SUD, the auxiliaries and subordinators are the heads, and hence we have the annotation of Figure 7, which is consistent with the distributional criteria.

Figure 7 - SUD annotation of Example (19)

3.5 Pronouns

Pronouns are often considered as function words, but not always (e.g., Fries (1952) ranks them in the same POS as nouns). This does not matter for the SUD annotation: Pronouns are analyzed, as all words, in accordance with the distributional principles. Anaphoric personal pronouns, which refer to an antecedent, do not entail any

[18] Note that this aberrant analysis is not the current UD analysis. The problem has been identified in the UD guidelines on the page http://universaldependencies.org/u/overview/complex-syntax.html (retrieved on August 23, 2021) retaining a SUD-style analysis:

"In these cases, the copula is treated as a head to preserve the integrity of clause boundaries and prevent one predicate to be assigned two subjects."

They further concede:

"This is not an optimal solution given the analysis of equational constructions involving nominals, where one of the nominals is treated as the head, but it is the preferred solution for now."

annotation difficulty because they behave like any substantive phrase. Relative and interrogative pronouns are more problematic because they play a double role: As pronouns, they refer to an antecedent, but they also give the interrogative or relative characteristic to the clauses in which they are present. We will address this issue in Section 4.

That leaves the expletive pronouns, i.e., pronouns without antecedent and without semantic content of their own. They fill the position of a syntactic argument that is required by a verb to express various features: pronominal verbs, impersonal verbs, impersonal diathesis, pronominal passive diathesis, … In SUD, they are annotated just as syntactic arguments which are content words. One reason for doing this in French is that they have the same grammatical properties. Consider the following example.

(20) a. Il arrive   deux personnes aujourd'hui.
        It is.arriving two   persons   today
        'Two persons are arriving today.'

     b. Il en   arrive   deux aujourd'hui.
        It of.them is.arriving two   today
        'Two of them are arriving today.'

In both sentences, the singular verbal form *arrive* agrees with the expletive pronoun *il*, which has the same form and position as a subject anaphoric singular pronoun *il*. It is for this reason that it is considered the subject of *arrive*. The noun phrase *deux personnes* is considered the direct object of the verb *arrive* because it has roughly the same properties as the object of a transitive verb. In particular, the second sentence shows that *personnes* can be pronominalized in the pronoun *en* like any (indefinite plural) direct object, as shown in (20b).

3.6 Multiple transfer and the internal structure of nuclei

For Tesnière (1959), the translative and the unit it transfers form a unit that he calls a (translative) nucleus.[19] He only represents relations between nuclei by dependencies and considers that the relations inside the nucleus are more cohesive and not hierarchized. Based on this idea, UD separates these two kinds of relations: criteria for

headedness are only applied for relations between nuclei and inside the nucleus, the content word is taken as the head and all translatives are dependents of the content word (de Marneffe et al. 2021, 2022). This has the pedagogical advantage of making very visible the link between nuclei by just pruning the function words from the UD tree. But it has an important disadvantage: The internal structure of the nucleus is completely lost.

As stated by Tesnière himself—who dedicates hundreds of pages to this question (Tesnière 1959[2015]: Part III)—multiple transfers are very common in natural languages. For instance, instead of saying *a nice house*, we can say *a house that is nice*, where the adjective *nice* is transferred into a verb by the copula BE and then into an adjective by the relativizer THAT (one of the reasons to do this is to introduce the possibility of adding tense on the copula and to contrast this with *a house that was nice*). Another example is given by adverbial clauses introduced by adpositions:

(21) a. avant   son depart
         'before his departure'

      b. avant qu'il parte
         'before   he leaves'

The French preposition *avant* can be followed by a substantive phrase or by a clause. But in the second case, the verb must be first transferred into a substantive by the complementizer *que* 'that'. See the SUD analysis in Figure 8.

Figure 8 - The SUD analysis of (21a) and (21b)

---

[19] The notion of "translative nucleus" is debatable. Some head-marking languages (Nichols 1986) mark the transfer by an affix on the governor rather than by a case marker on the dependent. This is the case of Wolof for instance. Compare (ia) and (b).
    (i) a. *xaj bi* 'the dog', lit. dog the
       b. *xaju Peer bi* 'Peter's dog', lit. dog-TR Peter the
In English too, the fact that a preposition forms a nucleus with its complement can also be questionned, because in some constructions, such as relative clauses, the preposition can be separated from its complement and remain with the verb, its governor. This is called piep-piping and is illustrated by Example (28).

In the UD analysis of (21b), both of these translatives, *avant* and *que*, are dependent on the verb. This flat structure does not indicate that the verb has been first transferred into a substantive by *que* and then into an adverb by the preposition *avant*. Some UD trees have up to seven dependents in the same nucleus. In some cases, the internal structure of the nucleus can be recovered by using a simple heuristic considering that the closer a function word is to the content word, the sooner it combines with it. But this heuristic does not always work. In Gerdes et al. (2021), we show problematic examples from Wolof and German.

4. Problematic constructions involving function words

In this section, we present our analysis concerning three important families of constructions involving function words, where the syntactic position of the function word needs a more thorough discussion: substantive phrases, relative and interrogative clauses, and coordination.

4.1 Substantive phrases

In the last decades, much debate has taken place about the head of a substantive phrase composed of a determiner and a noun. Here, we consider a determiner to be a syntactic unit that occurs together with a noun and cannot be suppressed in certain positions, contrary to modifiers.[20]

Due to the mandatory nature of the determiner, it is, for such a simple phrase as *the dog*, difficult to decide whether the head is *the* (making the whole a determiner phrase or DP) or *dog* (making the whole a noun phrase or NP). There is extensive research discussing these alternatives, but we can only give a very short review here (see Hudson 1984 for a first discussion and Hudson (forthcoming) and Osborne 2021 for a comprehensive survey).

---

[20] Note that this definition is narrower than the one used in UD, which is only semantic ("Determiners are words that modify nouns or noun phrases and express the reference of the noun phrase in context." https://universaldependencies.org/u/pos/DET.html retrieved on August 23, 2021). For instance, we do not consider that Italian possessive *mio* 'my' in *il mio libro* 'my book', lit. the my book, is a DET because it needs to combine with an article (*mio libro) and can always be suppressed (*il libro*).

With the criteria defined earlier, it is not possible to decide on the head of a substantive phrase. In sentences such as *He was here this morning*, or Fr. *Il est venu trois fois* 'He came three times', we can apply the distributional criterion without deletion, which shows that the distribution is controlled by the noun (*morning* or *fois*) so that the noun should be the head (*a morning*, *some morning*, *next morning* have the same distribution as *this morning*, whereas *this house*, *this dog*… do not).

On the other hand, using the Negative distributional criterion with deletion on the French phrase *la syntaxe* with the examples below, we observe that the distribution of the phrase is controlled by the determiner *la*. This would argue in favor of the determiner as the head of the phrase.

(22) a. J'enseigne la syntaxe. / *J'enseigne syntaxe.
        'I teach syntax.'

      b. On parle syntaxe. / *On parle la syntaxe.
        'We talk (about) syntax.'

      c. un livre de syntaxe / *un livre de la syntaxe
        'a book of syntax'

In conclusion, there is no definitive and uncontroversial choice between the two alternatives. In Hudson's (2004) words: "I shall argue that the determiner and the common noun […] each depend on the other, so either (but not both) of them may be the head of the noun-phrase." As syntactic criteria do not allow us to decide, we favor the semantic criteria that make the noun the head of the substantive phrase. This leads us to follow the UD choice on this issue, taking the noun to govern the determiner on this issue.

In French, some quantifications can be expressed by the ADV *de* NOUN construction: *beaucoup de chiens* 'a lot of dogs', *peu de chiens* 'few dogs'. In order to be consistent with the basic NP analysis above, it would be appealing to decide that *chiens* is the head in these two phrases. In the case of coordination, however, the preposition *de*

In the same way, French *tous* 'all' in *tous les étudiants* 'all the students' is categorized as an adjective because it can be suppressed and it cannot combine with the noun without a DET (**tous étudiants*). With such a definition of determiners, languages where the noun can always stand alone do not have determiners.

is generally repeated: *beaucoup de chiens et de chats* 'a lot of dogs and of cats', *peu de chiens et de chats* 'few dogs and cats'. These facts make the analysis with the noun as the head difficult to manage. Moreover, the same adverbs can be used alone in the same contexts in *beaucoup sont venus* 'a lot came' or *peu sont venus* 'a few came'.

For these specific constructions, we make the ADV the head of the structure and use the canonical way to deal with coordination *de chiens et de chats* (Figure 9). In order to keep a regular construction on the remaining part of the sentence, we mention the nominal aspect of the phrase with the feature *ExtPos=PRON* on the adverb.
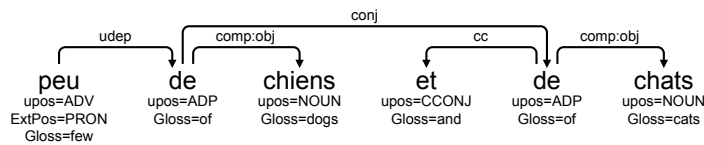
Figure 9 - SUD annotation of *peu de chiens et de chats*

Several expressions with a regular syntactic construction *un(e)* NOUN *de* NOUN can be interpreted as semantic quantifiers. For instance, *many students* can be translated into French by *un tas d'étudiants* (lit. a bunch of students) or *une foule d'étudiants* (lit. a crowd of students). But, as in the previous case with adverbs, a conjunction can be used in the second part in *un tas d'étudiants et de professeurs* 'many students and professors' and so, we annotate these phrases with *tas* or *foule* as heads of the structures. Moreover, this brings about an analysis which is consistent with other syntactic constructions *un(e)* NOUN *de* NOUN like Fr. *un verre de vin* 'a glass of wine'.

4.2 Relative and interrogative clauses

In Figure 2 (Section 2), the relativizer *wey* of Naija was analyzed as the head of the relative clause, because it is a pure translative that transfers a finite verbal construction into an adjectival clause that works as a noun modifier. Chinese also has a pure relativizer 的 *de*. See Figure 10 for an example from the corpus SUD_Chinese-GSDSimp@2.8.[21]

---

[21] Chinese UD treebanks use a special relation *mark:rel* for the relativizer, which gives the feature *@rel* in our SUD automatic conversion.

(23) 专利　　是　刺激　　个人　　　创造　　力　的　制度.
　　　zhuānlì　shì　cìjī　　gèrén　　chuàngzào lì　de　zhìdù.
　　　patent　is　stimulate individual creative　　force REL system
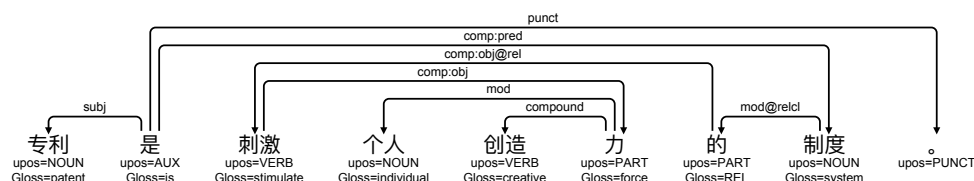　　　'Patents are a system that stimulates individual creativity.'

Figure 10 - SUD annotation of (23):
The Chinese 的 *de* is a pure relativizer

The case of a pure relativizer such as Naija *wey* or Chinese 的 *de* cannot be mixed up with relative pronouns. English for instance has two kinds of relativizers: *that*, which is a pure relativizer (Van der Auwera, 1985), and *wh*-words, which are relative pronouns.[22] They contrast by the fact that *wh*-words can involve the movement of a preposition (pied-piping, as in *the girl to whom I speak*) or they bear some morphological features, such as agreement with the noun antecedent (see the opposition between *who* and *which*) or case marking (*who, whom, whose*).

It has been argued by several authors that a relative pronoun fulfills two syntactic functions and thus occupies two positions: a translative function that makes it the root of the relative clause and a pronominal function that gives it a position inside the relative clause as any ordinary pronoun (see Tesnière 1959: ch. 246, Hudson 1984 , Kahane 2002).[23] An analysis with a double position would be possible in UD or SUD but for the sake of simplicity only one position, the pronominal position, is annotated.[24] Consider the example English-GUM_fiction_claus-39.

___

[22] In UD2.8 treebanks for English, the relativizer *that* is analyzed as *wh*-words.

[23] Analyses of extraction in generative grammars involving a movement of the extracted phrase and a trace also consider two positions. The fact that relative pronouns are also heads follows from the fact *who slept* and *she slept* have very different distributions (distributional criteria without deletion). Note that it does not follow from these data that the subject always heads the clause, because *wh*-words are the only subjects that control the distribution of the clause and moreover they change the distribution of the clause whatever their syntactic position inside the clause.

[24] For instance, an analysis with a double position has been adopted for amalgams such as Fr. *du* (= *de* + *le* 'of the') or De. *im* (= *in* + *dem* 'in the.DAT'). A *wh*-word could be analyzed as the amalgam of a pure relativizer and a pronoun.

(24)   She picked up the pans in which she'd made the potatoes and maple glaze.

Figure 11 gives the SUD analysis of (24), where *which* occupies only the pronominal function and Figure 12 an analysis where *which* occupies two positions (the translative position is called REL). It is possible to recover the translative position of the relativizer because a special relation, *mod@relcl*, is used between the governor of the relative clause and the main verb (similarly to the acl:relcl of UD).
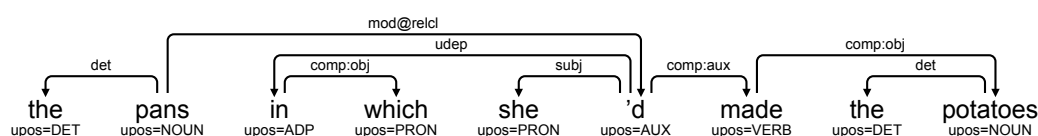
Figure 11 - SUD analysis of (24)
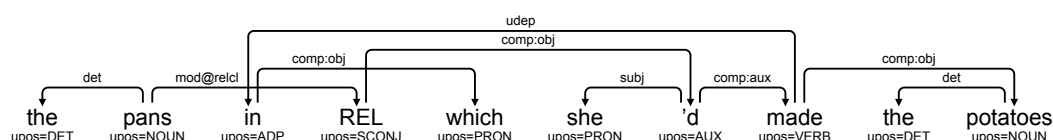
Figure 12 - Analysis of (24) with the double position of which

It is worth noting that the analysis with a unique position becomes problematic for free relative clauses. Consider the example GUM_news_iodine-3.

(25) Professor Eastman said he is alarmed by what they found. Clauses such as *what they found* here do not have an antecedent, and it is clear that the *wh*-word occupies two positions, translating the verb into a substantive phrase. In this case, UD has decided to analyze the *wh*-word as the head of the relative clause, which, after conversion into SUD, gives us the translative-based analysis of Figure 13.[25]

---

[25] The reason why UD must adopt this analysis is a consequence of the UD analysis of adpositions. In UD, adpositions are analyzed as a *case* marker of their complement, that is, as dependent on their complement by a case relation. But if the verb becomes the head of the relative clause, the adposition would be a *case* marker on this verb, which would be quite problematic.
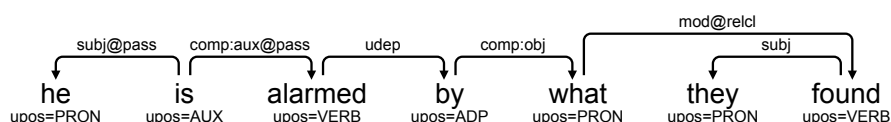
Figure 13 - SUD analysis of (25) obtained from the conversion of the UD analysis (translative-based analysis)

The analysis of Figure 13 is equivalent to the analysis of free relative clauses in Chinese with the pure relativizer 的 *de*. Figure 14 gives an example of a free relative clause in SUD_Chinese-HK@2.8.

(26) 餘下　的　都　　沒　用！
　　　Yúxià de　dōu　méi　yòng!
　　　leave REL entirely no　use
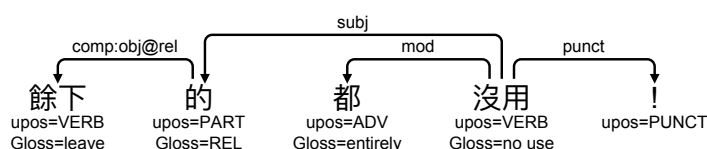　　'The rest is useless!' 'What has been left is useless!'

Figure 14 - SUD analysis of (26)

We think that the translative-based analysis of free relative clauses in English is not optimal for various reasons: first, it does not follow the pronoun-based analysis that has been chosen for standard relative clauses; second, the pronominal position of the *wh*-word is lost; third, the extension @*relcl* no longer corresponds to the relation between the relative clause and its governor (since the *wh*-word is part of the relative clause) and to the translative function.

We propose to adopt a pronoun-based analysis for the free relative clauses, with the *wh*-word in the pronominal position (see Figure 15, where *what* is the *comp:obj* of *found*) (this analysis is not implemented in the current SUD 2.8).
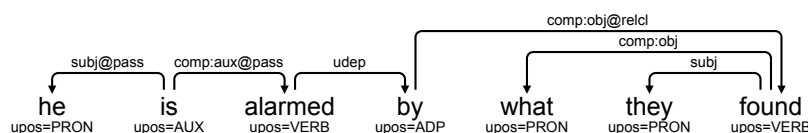
Figure 15 - SUD analysis of (25) we propose (pronoun-based analysis)

In this analysis, the main verb is the head of the relative clause, and the position of the relativizer is still indicated by a feature *@relcl* on the relation between the relative clause and its governor (in our example, the relative clause is the complement of the adposition *by*, giving us a relation *comp:obj@relcl* between *by* and the main verb of the clause). As a consequence, the additional feature *@relcl* on relations is no longer restricted to modification relations between an antecedent noun and the relative clause. Note that the pronoun-based analysis can be easily converted into the translative-based analysis, but the way back is not possible and only native SUD treebanks could have the pronoun-based analysis.

A fourth reason to adopt a pronoun-based analysis for free relative clauses is the fact that indirect interrogative clauses also adopt a pronoun-based analysis and that the frontier between free relative clauses and indirect interrogative clauses is not clear-cut.[26] To be more parallel with the analysis of relative clauses, we also propose to add a feature *@intcl* on the relation between the interrogative clause and its governor. Consider the example GUM_vlog_pregnant-22 and the SUD analysis we propose in Figure 16.

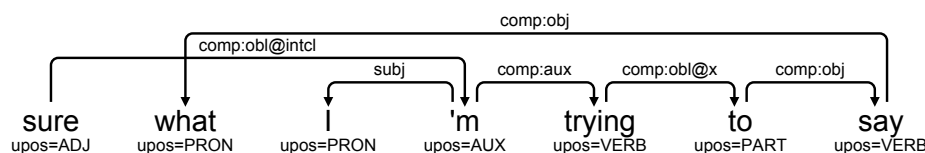(27) And, I'm not really sure what I'm trying to say.

Figure 16 - SUD analysis of an indirect interrogative clause

4.3 Coordinating conjunctions

Existing dependency annotation schemes differ a lot in the way coordination is analyzed. The goal of this section is to explain the choices of the SUD annotation for coordination structures. We will focus on basic coordination without considering gapping or non-constituent coordination (see Gerdes & Kahane 2015 for a more

---

[26] Free relative clauses are in a paradigm of substantive phrases, while indirect interrogative clauses have their own positional paradigm containing *whether*-clause. But in some positions both commutations are possible:
(i) a. I don't understand what he did.
    b. I don't understand his decision.
    c. I don't understand whether he'll come or not.

comprehensive analysis) and consider only examples in English. The same analysis may be conducted on other languages but with different conclusions: For head final languages, there are strong arguments to propose an analysis where the second argument is the head of the structure (see Kanayama et al. 2018 for Japanese and Korean and Kahane et al. 2021b for Beja).

In basic coordination structures like *Alice and Bob*, we first have to decide what are the syntactic units (Criteria A). The segment *and Bob* can stand alone in an isolated sentence or in a speech turn during a dialog whereas the segment *Alice and* cannot (Mel'čuk 2009:72). Hence, the syntactic units are the three tokens, *and Bob*, and the whole unit *Alice and Bob*. The next step is to decide the heads in combinations of syntactic units (Criteria B). Between *Alice* and the segment *and Bob*, we can clearly apply the positive criterion with deletion (*Alice* and *Alice and Bob* have the same distribution), hence *Alice* is the head. Deciding the head between *and* and *Bob* in *and Bob* is less straightforward. Both elements control the distribution of the whole segment: *and* cannot stand alone and *Bob* and *and Bob* have different distributions, therefore *and* is a head, according to the Negative distributional criteria. But at the same time, the distribution of the segment also depends on the conjunct *Bob*; *and Bob* can be combined with a proper noun in *Alice and Bob*, but not with another category (**yellow and Bob*). Therefore both elements can be considered as the head. We decide to favor the relation between the conjuncts *Alice* and *Bob*, because this relation must be present when there is no coordinating conjunction between them (for instance in *Alice, Bob, and Charlie*) (see Mazziotta 2011 for a similar argumentation in Old French, where the coordinating conjunction is particularly optional). In other words, we follow the UD analysis and consider that *Bob* is the head of the segment *and Bob*. The names of the relations are also kept from UD annotation (*conj* and *cc*). This analysis of coordination is also supported by another construction, when both conjuncts are introduced by a coordinating conjunction, as in *and Alice, and Bob*. Only the analysis with the coordinating conjunction as a dependent can provide *Alice* as the head of the coordinated phrase and a symmetrical analysis of *and Alice* and *and Bob*.[27]

---

[27] Double coordination is common in some languages, such as Fr. *soit A, soit B* 'either A or B'.

When coordination is iterated, as in *Alice, Bob and Charlie*, SUD adopts a string-analysis of coordination, where each conjunct depends on the previous one (Figure 17). The main argument for the string-analysis is that it reduces the dependency length (Gibson, 1998; Liu, 2008; Futrell et al., 2015).
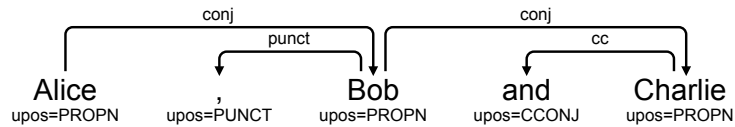
Figure 17 - Iterated coordination

One of the conjuncts of a coordination can itself be a coordination. Even if it is theoretically possible to have several layers of embeddings, the real embedding depth is at most 1 in the SUD corpora we were working on. In the example below (GUM_fiction_garden-4), the conj label does not make a distinction between embedded relations and surface relations because they form a single chain. But, the relation can be distinguished with the use of the extension @*emb* for embedded coordination.[28]
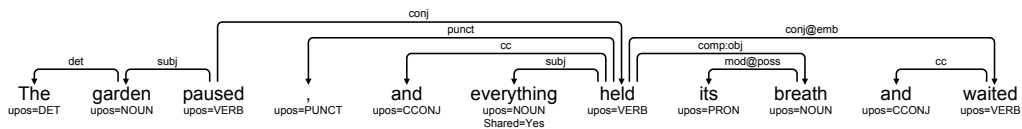
Figure 18 - Embedded coordination

Left dependents on coordination are attached to the head of the coordination which is the head of the leftmost conjunct. Hence, it is not possible to distinguish between a dependent on the first conjunct and a dependent of the whole structure. In order to be consistent with the string-analysis of coordination, right dependents on a coordination are attached to the head of the rightmost conjunct. Again, a dependent of the last conjunct only can be confused with a dependent of the whole coordination. For this latter ≈situation, in SUD, we use the feature *Shared=Yes* on the dependent when it should be distributed

---

[28] UD distinguishes embedded coordinations when they are part of the second conjunct: in "A and (B or C)", C would be attached to B. But in "(A and B) or C", both B and C would be attached to A as in an iterated coordination "A, B, and C". In SUD, both cases are treated in the same way, with a *conj@emb* relation.

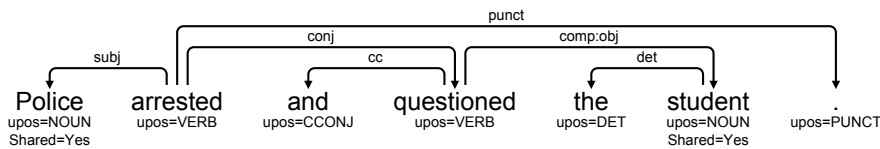on other conjuncts (Fig. 19) and no specific annotation when it is related only to the closest conjunct (Fig. 20).[29]
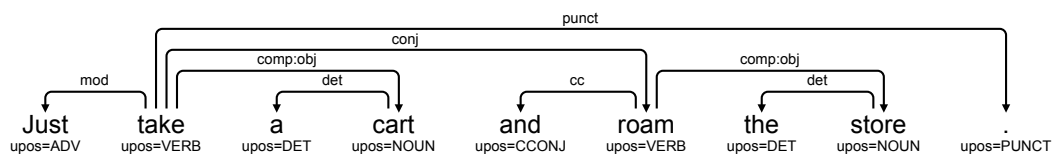
Figure 19- Shared dependents of a coordination

Figure 20 - Non-shared dependents

5. Conclusion

To conclude we recall that SUD is a simple and operational annotation scheme, which implies a tree-shaped structure with few dependency relations on a limited number of POS tags. SUD is compatible with other syntactic resources, in particular UD, to which it is aligned by means of rule-based transformations.

SUD is based on distributional criteria, which allow the syntactic structure to recover the syntactic units. Let us compare the UD and SUD analyses of Example GUM_whow_joke-27 in Figures 21 and 22.

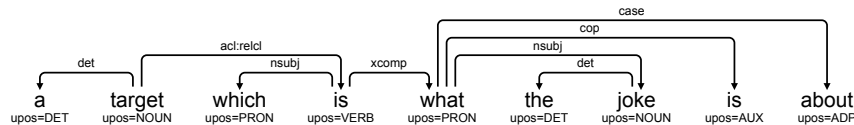(28) Every joke has a "target," which is what the joke is about.

Figure 21 - UD analysis

---

[29] The problem is avoided in UD for shared dependents on the right, because shared dependents are attached to the head of the coordinated phrase, that is the first conjunct. Nevertheless, in UD, as well as in SUD, the same ambiguity appears between a shared and non-shared dependent on the left, all attached to the first conjunct. At first, we above all wanted to be compatible with UD and we did not consider sharedness on the left. But now for the internal coherence of SUD we decided to add the feature *Shared=Yes* on shared left dependents too.
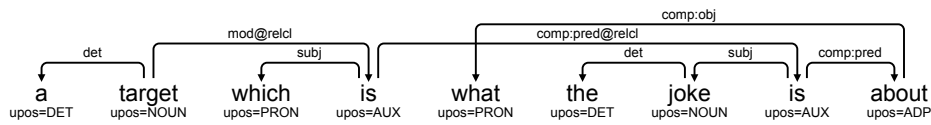
Figure 22 - SUD analysis

The relative clause contains almost only function words. In such a case, UD principles are very difficult to apply. In order to avoid having two copulas on the same word, the first copula must be treated in this case as a plain verb, which is already a problem.[30] Moreover, *what* becomes the head of the predicative complement and the subject *the joke*, the copula, and the adposition *about* are analyzed as dependents of *what*. The whole internal structure is lost. On the SUD analysis, in contrast, we indicate that we have two similar constructions "subject | verb | predicative complement" embedded, the first being a relative clause and the second a free relative clause. It is clear on the SUD analysis that *about + what* is a syntactic unit of (28) and that *what + the joke* or *what + is* are not.

Compared to UD, SUD relations are less redundant with POS annotation and they are established by means of solid distributional criteria that do not require the cumbersome distinction between function and content words. SUD has proved to be operational and used for annotation projects of diverse corpora: both written and spoken French (Kahane et al. 2021a); Naija, an English-lexifier pidgincreole from Nigeria (Caron et al. 2019); and Beja, a Cushitic language from Sudan (Kahane et al. 2021b). Thus, SUD treebanks are easier to create while allowing heuristic comparative syntax as a part of the UD project.

Although based on solid syntactic criteria inherited from a long tradition in dependency syntax (Imrényi & Mazziotta 2020), SUD does not pretend to be a syntactic theory and remains, above all, an operational and well-defined syntactic annotation scheme.

Works Cited

1. Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.
2. Bondéelle, O. & S. Kahane. (2021). Les particules verbales du

---

[30] We corrected the UD2.8 analysis, where both copulas and both subjects depended on *what*.

wolof et leur combinatoire syntaxique et topologique. *Bulletin de la Société de Linguistique de Paris*, *115*(1), 391-465.

3. Bresnan, J., A. Asudeh, I. Toivonen & S. Wechsler. (2015). *Lexical-functional syntax*. John Wiley & Sons.

4. Caron, B., M. Courtin, K. Gerdes & S. Kahane. (2019). A surface-syntactic UD treebank for Naija. *Proceedings of the 18th conference on Treebanks and Linguistic Theories (TLT)*, Syntaxfest.

5. Chomsky, N. (1969) Deep Structure, Surface Structure, and Semantic Interpretation, *Studies on Semantics in Generative Grammar*, 62-119.

6. —. (1970). Remarks on nominalization. *Studies on Semantics in Generative Grammar*, 11-61.

7. Corver, N. & H. van Riemsdijk. (2001). *Semi-lexical Categories –The function of Content Words and the Content of Function Words*. Berlin; New York: Mouton de Gruyter.

8. de Marneffe, M. C., C. D. Manning, J. Nivre & D. Zeman. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.

9. de Marneffe, Nivre, J., & D. Zeman. (2022). Function words in Universal Dependencies. *Linguistic Analysis* 43:3-4.

10. Fries, C. C. (1952). *The structure of English*. New York: Harcourt Brace.

11. Futrell, R., K. Mahowald & G. Edward. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341.

12. Garde, P. (1977). Ordre linéaire et dépendance syntaxique : Contribution à une typologie. *Bulletin de la Société de Linguistique de Paris, 72*(1), pp. 1-26.

13. —. (1994). Syntaxe et sémantique chez Tesnière. *Linguistica, 34*(1), 95-99.

14. Gerdes K., B. Guillaume, S. Kahane, G. Perrier. (2018) SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD, *Proceedings of the Universal Dependencies Workshop (UDW)*, EMNLP.

15. —. (2019) Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic functions and deep-syntactic features, *Proceedings of the 17th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

16. —. (2021a) Starting a new treebank? Go SUD!, *Linguistic Analysis* 43 3-4.

17. Gerdes, K. & S. Kahane. (2001). Word order in German: A formal dependency grammar using a topological hierarchy. *Proceedings of the 39th annual meeting of the Association for Computational Linguistics* (ACL), 220-227.

18. —. (2013). Defining dependencies (and constituents). *Frontiers in Artificial Intelligence and Applications, 258*, 1-25.

19. —. (2015). Non-constituent coordination and other coordinative constructions as Dependency Graphs. *Proceedings of the 3rd conference on Dependency Linguistics (Depling)*.

20. —. (2016). Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*.

21. Gerdes, K., S. Kahane & X. Chen. (2021b). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, 17.

22. Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1-76.

23. Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 168-175.

24. Hajičová, E., J. Hajič, B. Hladká, M. Holub, P. Pajas, V. Řezníčková & P. Sgall. (2001). The current status of the Prague Dependency Treebank. *International Conference on Text, Speech and Dialogue* (pp. 11-20). Berlin/Heidelberg: Springer.

25. Harris, S. Z. (1951). *Methods in Structural Linguistics*. The University of Chicago Press.

26. Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: MacMillan Company.

27. Hudson, R. (1984). *Word Grammar*. Oxford: Blackwell.

28. —. (1987). Zwicky on heads. *Journal of linguistics, 23*(1), 109-132.

29. —. (2004). Are determiners heads? *Functions of Language, 11*, 7–43.

30. —. (forthcoming). Arguments against the universality of D and determiners. Dans S. Armoskaite, & M. Wiltschko, *The Oxford Handbook of Determiners*. Oxford University Press.

31. Imrényi, A. & N. Mazziotta. (eds.). (2020). *Chapters of Dependency Grammar: A historical survey from Antiquity to Tesnière* (Vol. 212). Benjamins.

32. Kahane, S. & K. Gerdes. (forthcoming). *Syntaxe théorique et formelle, vol. 1: Modélisation, Unités, Structures*. Berlin: Language Sciences Press.

33. Kahane, S., B. Caron, K. Gerdes, E. Strickland. (2021a) Annotation guidelines of UD and SUD treebanks for spoken corpora, *Linguistic Analysis* 43 3-4.

34. Kahane, S., M. Vanhove & R. Ziane. (2021b) A morpheme-base treebank for Beja. *Linguistic Analysis* 43 3-4.

35. Kanayama, H., N.-R. Han, M. Asahara, J.DS. Hwang, Y. Miyao, J.D. Choi & Y. Matsumoto. (2018). Coordinate Structures in Universal Dependencies for Head-final Languages. *Second Workshop on Universal Dependencies*. Brussels, Belgium.

36. Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159-191.

37. Mazziotta, N.(2011) Coordination of verbal dependents in Old French: Coordination as a specified juxtaposition or apposition. *Proceedings of the 1st International Conference on Dependency Linguistics (Depling)*, 28-37.

38. Mel'čuk, I. (1964). *Avtomatičeskij sintaksičeskij analiz, Tom I [Automatic syntactic analysis, Vol. 1]*. Novosibirsk: Izdatel'stvo SO AN SSSR.

39. —. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

40. —. (2009). Dependency in natural language. In Mel'cuk, I & Polguère A. (eds.), *Dependency in Linguistic Description*, 1-110.

41. —. (2021). Syntactic subject: syntactic relations, once again. In *Ten Studies in Dependency Syntax*, De Gruyter Mouton, 117-178.

42. Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language*, 56-119.

43. Nivre, J., M.C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C.D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty & D. Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, 1659-1666.

44. Osborne, T. (2021). NPs, not DPs: The NP vs. DP debate in the context of dependency grammar. *Acta Linguistica Academica*.

45. Osborne, T. & K. Gerdes. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*.

46. Osborne, T., M. Putnam & T. Gross. (2012). Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4), 354-396.

47. Saunders, P. T. (1980). *An introduction to catastrophe theory*. Cambridge: University Press.

48. Savary, A., S. Cordeiro & C. Ramisch. (2019). Without lexicons, multiword expression identification will never fly: A position statement. *Joint Workshop on Multiword Expressions and Word-Net (MWE-WN 2019)*.

49. Sgall, P., E. Hajicová, E. Hajicová, J. Panevová & J. Panevova. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

50. Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck. Engl. transl.: *Element of structural syntax*, Benjamins, 2015.

51. Van der Auwera, J. (1985). Relative *that* – a centennial dispute. *Journal of linguistics, 21*(1), 149-179.

52. Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. *Proceedings of the IFIP congress*, 1114-1122.

53. Zwicky, A. (1985). Heads. *Journal of linguistics, 21*(1), 1-29.