SUD Annotation Guideline
Author: Cheikh Bamba Dione
University Gaston-Berger, Senegal

Contents

1 Introduction

This guideline introduces basic concepts of syntactic annotation in SUD. Using most examples from Wolof, this guide aims to generalize enough for other African languages that have similar phenomena as those found in Wolof.

2 POS annotation

Recall: In the first stage of the project, we proceeded to the part-of-speech (POS) tagging of words. The list of the tags used is given in table 1:

Table 1: Universal POS tags

ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Note: The POS tags are always written in **capital** letters.
Example: let's consider the Wolof sentence in (1).

(1) *xale bi gis na nag wi.*
child the see 3SG cow the
'The child has seen the cow.'

For this sentence, we have the following POS tags (upper line) and glosses (bottom line):

NOUN	DET	VERB	AUX	NOUN	DET	PUNCT
xale	bi	gis	na	nag	wi	.
child	the	see	3SG	cow	the	.

3 Syntactic annotation

In the second phase of the project, we're going to annotate **syntactic relations** (given in Table 2) between words.

Table 2: Syntactic relations

Relation	**Subcategories**	**Description**
Root		root of the sentence
subj		subject
comp	comp:aux	auxiliary complements
	comp:obj	object complements
	comp:obl	oblique complements
	comp:pred	predicative complements
	comp:cleft	cleft complements
mod		modifiers
udep		portmanteau dependency for mod and comp:obl
compound	compound	regular compounds
	compound:prt	verb particle compounds
	compound:svc	serial verb compounds
appos		appositional modifiers
conj		coordinate conjuncts
cc		coordinate conjunctions
flat		names
fixed		fixed grammatical expressions
dislocated		dislocated elements
punct		punctuation

For a better understanding of syntactic annotation, it may help to first introduce some basic notions: constituents or phrases, head vs. dependent, arguments vs. modifiers.

Syntactic phrases

A sentence structure can be decomposed into phrases or constituents. For instance, an English sentence like "The child has eaten the big apple" can be seen as composed of two main constituents:

1. a nominal phrase (NP): "the child"
2. a verbal phrase (VP): "has eaten the big apple"

In turn, the VP can be decomposed into an auxiliary (*has*), a main verb (*eat*) and an NP (*the big apple*).[1]

Head vs. dependent

The syntactic relations in SUD are in the form **head-dependent**. The head is the main word of a phrase. A head is required for a phrase (it is obligatory in the phrase, i.e. can't be omitted).

With respect to our English example, the head phrase is the VP; the two NPs being its dependents. Then, we need to identify the main word of the VP. Here, we should decide between the main verb and the auxiliary. As a general rule, SUD considers the auxiliary (here *has*) the head of the VP, and therefore the head of the entire sentence. Accordingly, the auxiliary has three dependents: the first NP, the main verb, and the second NP.

Subsequently, we will need to identify the syntactic relation between these constituents, as described below.

Argument vs. modifier

Arguments are inherent to the meaning of the head of the phrase or clause. In contrast, modifiers merely supplement the head with additional information. This means that arguments are generally obligatory, while modifiers are optional.

[1]The NP (the big apple) can be further decomposed into a determiner (the) and an adjectival phrase (big apple).

Examples In (2), the subject (John) and the object (pizzas) are arguments of the verb (loves). These nouns are required by the verb “loves", which is a transitive verb, meaning it requires a subject and a direct object.

(2) **John** loves **pizzas**.

In (3), “tasty” is not an argument, but rather a modifier of the noun “pizzas". Removing this adjective will not affect the grammaticality of the sentence.

(3) John loves **tasty** pizzas.

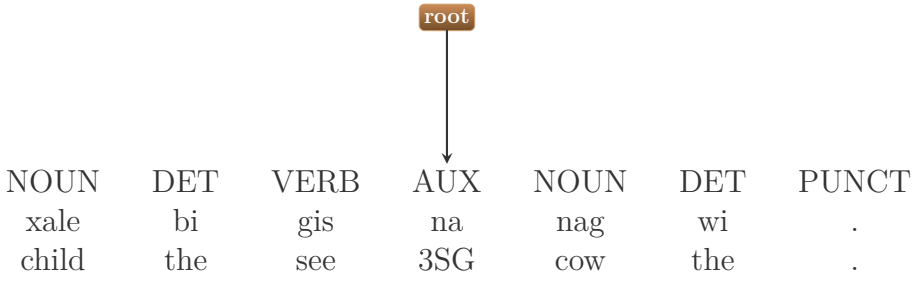
Below we will describe the most common syntactic relations in SUD.

3.1 Relation: root

The **root** relation is used to indicate the root (i.e. the head) of the whole sentence. There is always **one and only one word** that functions as the root of a sentence. Typically, it is the auxiliary that plays this role. In the absence of the auxiliary, it's usually the main verb that would function as the root.

Let's reconsider the Wolof sentence in (1), repeated in (4). The root of this sentence is the auxiliary *na* and we mark this relation as shown below:

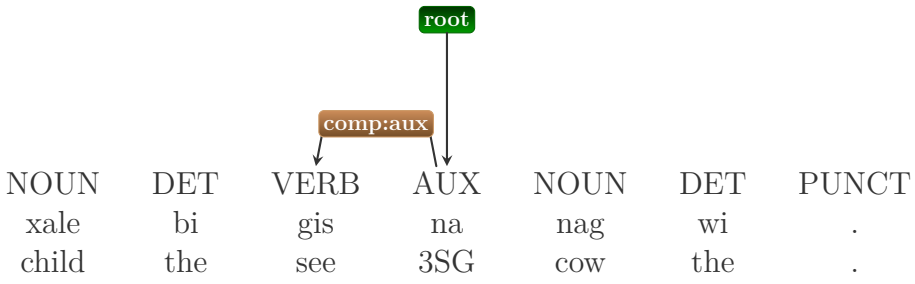
(4) *xale bi gis na nag wi.*
child the see 3SG cow the
‘The child has seen the cow.’



Note: *na* is a TAM (tense-aspect-mood) marker. it is an auxiliary that carries information about aspect (perfective) and agrees with the subject in person and number (third singular). It may translated into English as ‘has’.

3.2 Relation: comp:aux

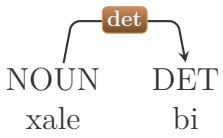
When functioning as a head, an auxiliary will have one or more dependent(s). One typical dependent of the auxiliary is the main verb. The relation used to mark such dependency is called **comp:aux**, which is used for the argument of auxiliaries. Thus, for the Wolof example, the verb *gis* is treated as a comp:aux of the auxiliary *na*.[2]



Note: The arrow that marks a relation **runs from the head to the dependent**. This is a general principle in dependency grammar.

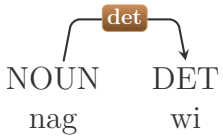
3.3 Relation: det

The relation determiner (**det**) holds between a nominal head and its determiner. For instance, the Wolof sentence in (1) has two noun phrases (NP). The first NP is: **xale bi** (child the). The head of this NP constituent is the noun (*xale*). Conversely, the article (*bi*) functions as the dependent of that noun. We call this a *determiner* relation (**det**). In general, this is the relationship we choose for words whose POS tag is DET.

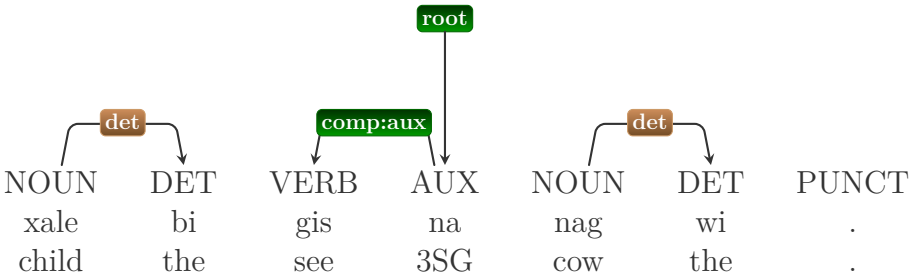


Likewise, we have the same type of relation for the second NP, i.e. between *nag* (cow) and *wi* (the).

[2]In general, the POS tag for complements of auxiliaries are VERB and AUX.

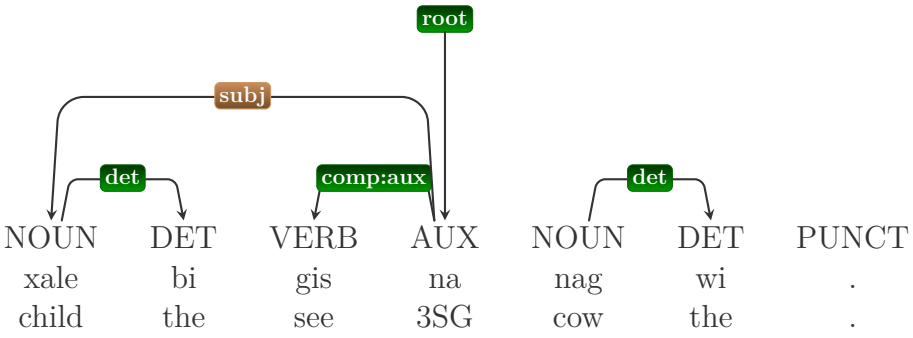


Until now, the syntactic annotation of our Wolof sentence looks like this:



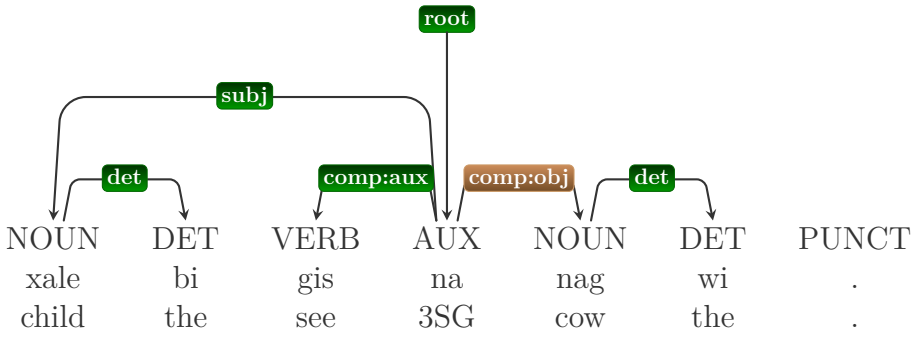
3.4 Relation: subj

The **subj** relation is used for subjects. In our Wolof example (1), the subject is “xale" (child), i.e. the one who is performing the seeing action. However, note that, since the head of the verbal phrase is the auxiliary (and not the main verb), we mark this relation between the subject and the auxiliary. Also, there is some logic going on here, because in Wolof (and in many other languages), the subject agrees with the auxiliary in person, number, etc.



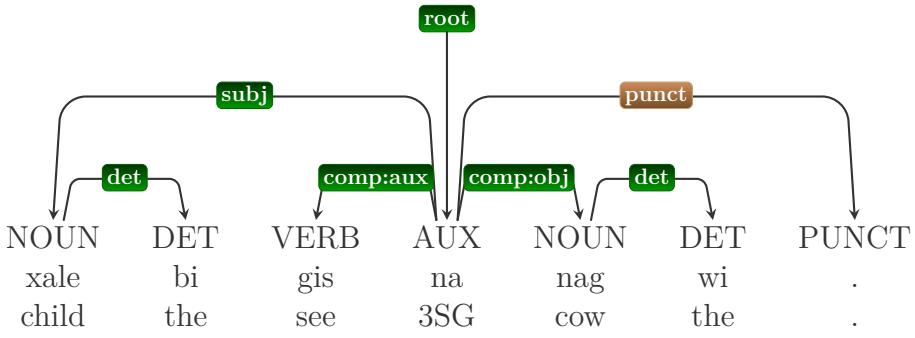
3.5 Relation: comp:obj

The **comp:obj** relation is used for **direct object** complements.[3] With respect to our Wolof example, the direct object complement is the NP "nag wi" (cow the). As with the subject, we need first to identify the head of that NP (which is *nag* "cow"). So, we will connect the auxiliary with the head of that NP through the comp:obj relation.



3.6 Relation: punct

Punctuation is represented by the relation **punct**. Here, the period is treated as a dependent of the sentence's head, i.e. the auxiliary (*na*). Generally speaking, there are specific rules for attaching punctuation marks in SUD. Some of them (in particular for apposition, dislocation and coordination) are discussed below.



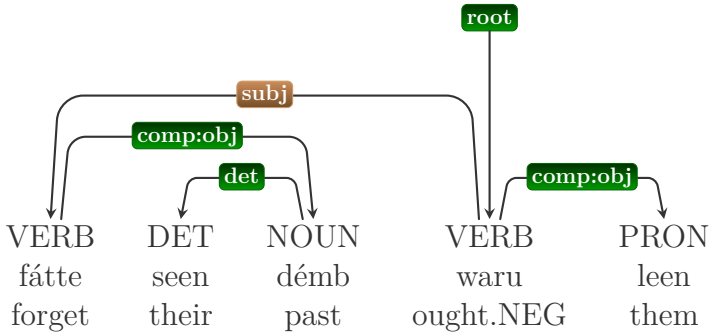
This completes the annotation of our Wolof sentence given in (1).

[3]The comp:obj relation is also used for direct complements of an adposition or a subordinating conjunction (see below).

3.7 More on the relation subj

The **subj** relation is also used for verbal phrases (VP) that act as a subject. For example, the subject of the sentence in (5) is not a nominal, but rather a verbal phrase (the underlined words). Typically these kinds of VP do not have an internal subject (i.e. are impersonal). In this example, the subjectless VP functions as subject of the sentence just as a regular NP would do.

(5) *Fàtte seen démb waru leen*
forget their past ought.NEG 3PL.OBJ
Lit. 'forget their past ought not them'
'They ought not forget their past'



Remarks: Note the following

- The subject of the sentence is the entire VP (*fàtte seen démb*) "forget their past".
- In this example, there is no auxiliary, so the main verb (ought) is the root of the sentence.
- The form *leen* is an object form, ruling out the idea to treat this as a subject. Such a form is always an object in Wolof (never a subject).

3.8 Relation: comp

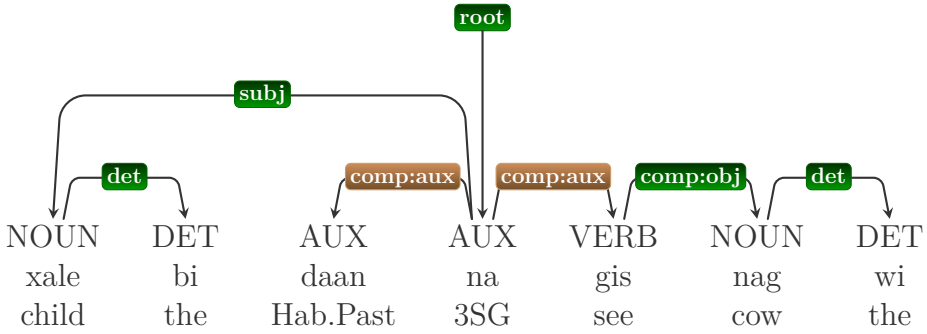
The **comp** relation is a generic relation used for arguments of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions. In SUD, comp is refined into several sub-categories: comp:aux (auxiliary argument), comp:cleft (cleft clauses), comp:obj (direct object), comp:obl (oblique argument), comp:pred (predicative argument).

Below, we discuss these different subcategories of comp.

3.8.1 More on comp:aux

The comp:aux relation has already been discussed in section 3.2. We should just add that in some cases **multiple auxiliaries** may appear. They are then annotated as shown in (6). In this example, we have two auxiliaries: *daan* which conveys a habitual past meaning and *na* which expresses perfective third singular. Note also the change in word order: the main verb (*gis*) follow the sequence of auxiliaries. In this case, the head auxiliary (here *na*) takes the other auxiliary as its dependent through the comp:aux relation.

(6) *xale bi* ***daan*** ***na*** *gis nag wi*
child the Hab.Past 3SG see cow the
'The child used to see the cow'



3.8.2 More on comp:obj

Beside direct objects (see section 3.5), the **comp:obj** relation is also used for:

- subordinate clauses without subordinating conjunction
- subordinate clauses with subordinating conjunction
- subordinate clauses without an internal subject

Subordinate clause without subordinating conjunction

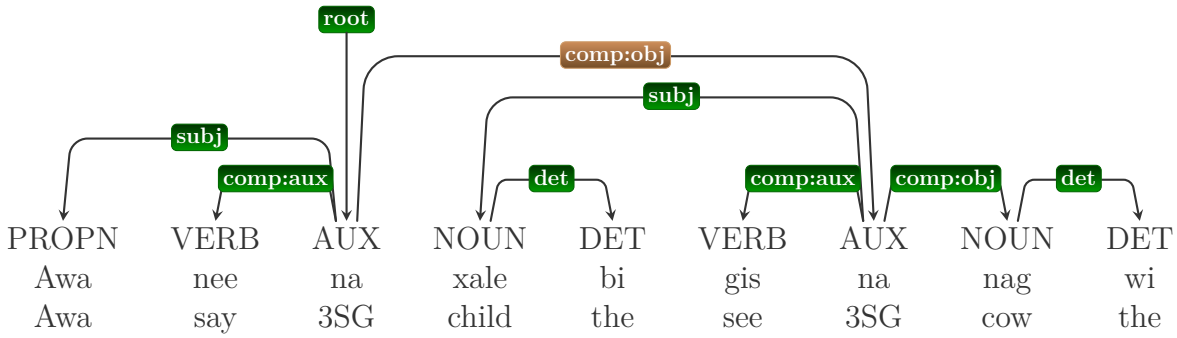
Let's reconsider our Wolof sentence in (1), repeated in (7).

(7) *xale bi gis na nag wi.*
child the see 3SG cow the
'The child has seen the cow.'

The sentence in (7) may appear as a **subordinate** clause of a main clause, as shown in (8).

(8) *Awa nee na xale bi gis na nag wi.*
Awa say has child the see has cow the
Lit. 'Awa has said the child has seen the cow.'

In cases like (8), the subordinate clause is treated as a **comp:obj** of the main clause. Since in our case, each of these clauses is headed by an auxiliary (the first *na* auxiliary is the head of the main clause, and the second *na* auxiliary is the head of the subordinate clause), the comp:obj relation is used to link these two auxiliaries.



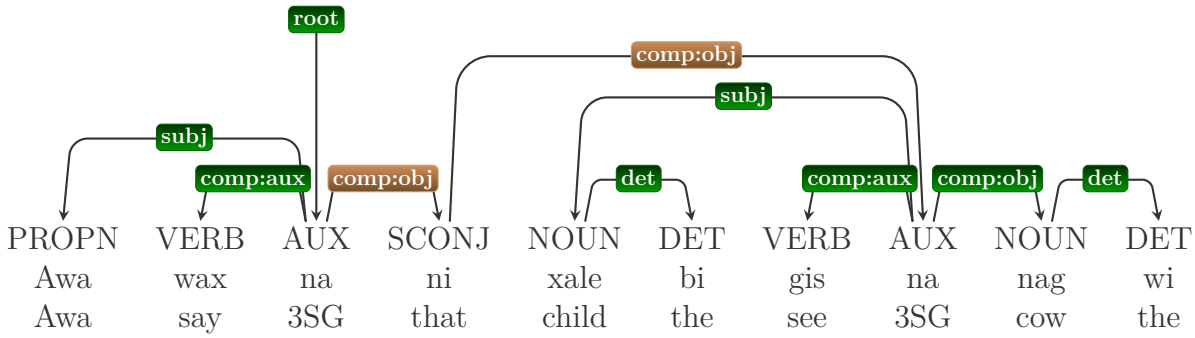
Note that, as this example shows, it is possible in Wolof to leave out the subordinating conjunction (SCONJ), e.g. ***that***, which typically introduces such subordinate clauses as in "Awa has said **that** ...' .

Subordinate clause with subordinating conjunction

In many languages, subordinate clauses are usually introduced by a subordinating conjunction (e.g. ***that*** in English). Likewise, Wolof has markers like "ni" and "ne", which play the same role, as illustrated in (9).

(9) *Awa wax na* ***ni*** *xale bi gis na nag wi.*
Awa say 3SG that child the see has cow the
'Awa has said that the child has seen the cow.'

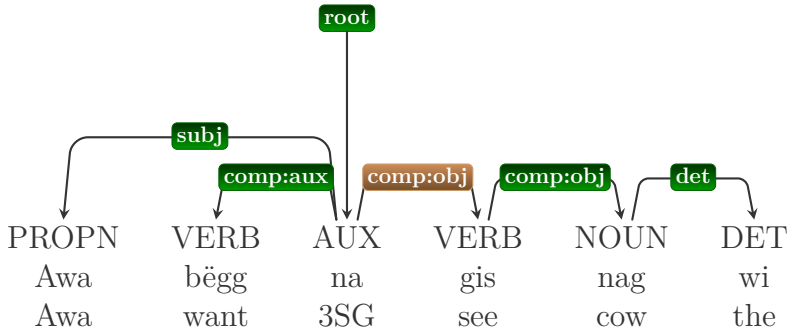
Like the preceding example, in sentence (9), we have a comp:obj relation. However, unlike in the preceding example, in (9), the comp:obj relation **applies between the main auxiliary and the subordinating conjunction** *ni* (“that"). This is because SUD considers the subordinating conjunction (when present) as the head of the subordinate clause. Furthermore, as the head of the subordinate clause, the subordinating conjunction takes the auxiliary *na* (i.e. the second *na*) as its dependent. Here, too, SUD uses the comp:obj relation to mark such a dependency.



Infinitive complements

Verbs that allow people to express wishes, hopes, wants, and expectations usually trigger infinitive complements. For instance, in (10), the complement VP *gis nag wi* “see the cow" is an infinitive complement of the verb *bëgg* ‘want’. But, because SUD considers the auxiliary (*na*) rather than the main verb as the head of the VP, the relation applies between the auxiliary and the complement VP. Here too, the comp:obj relation is also used to mark this dependency.

(10) *Awa bëgg na gis nag wi*
Awa want 3SG see cow the
‘Awa wants to see the cow.’



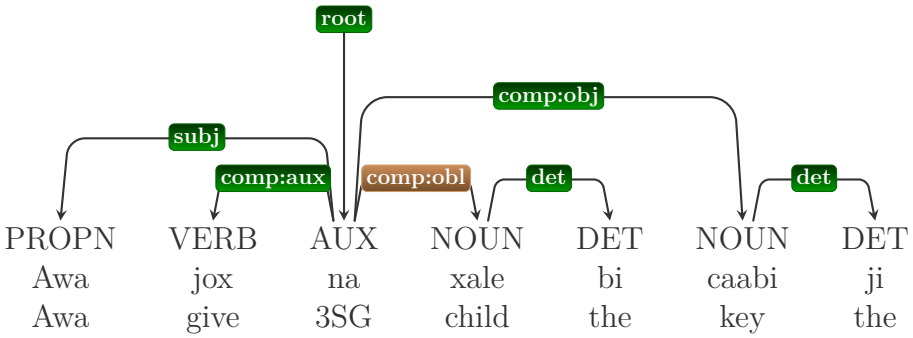
3.8.3 Relation: comp:obl

The **comp:obl** relation is used for (1) **indirect** objects and (2) **oblique** arguments.

Indirect objects

In general, an indirect object is the word or phrase that receives the direct object. For example, in the Wolof sentence in (11), the word *xale* 'child' is considered the indirect object (comp:obl); *Awa* being the subject (subj) and *caabi* 'key' being the direct object (comp:obj).

(11) *Awa jox na xale bi caabi ji*
Awa give 3SG child the key the
'Awa has given the key to the child'

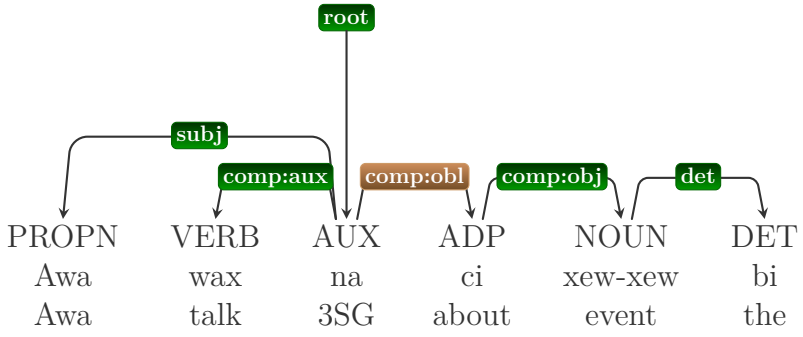


Oblique arguments

An oblique argument is an argument of a relation that is typically marked with a preposition[4] (or by case in some languages). Syntactically speaking, oblique arguments are not direct arguments (i.e. they are not subjects or direct objects or second objects). However, SUD does not make the distinction between oblique and indirect objects. Both are marked with the **comp:obl** relation.

For instance, the Wolof sentence in (12) contains a prepositional phrase (PP): *ci xew-xew bi* "about the event". Here, the preposition (*ci* "about") is considered the head of the PP, playing the role of **comp:obl** with respect to the head of the sentence (i.e. the auxiliary *na*). In turn, within the PP, the noun *xew-xew* "event" functions as the comp:obj of the preposition.

(12) *Awa wax na ci xew-xew bi*
Awa talk 3SG about event the
'Awa has talked about the event'

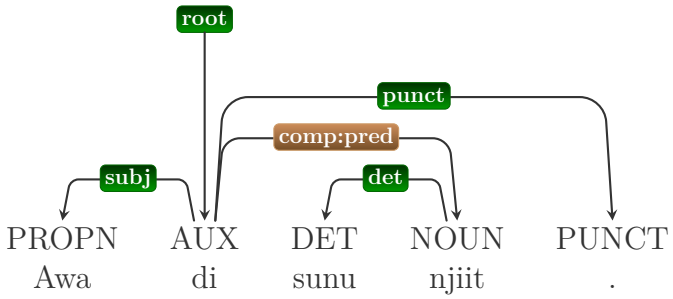


3.8.4 Relation comp:pred

The relationship **comp:pred** is used for predicative verb arguments. For example, the verb *to be* is a copula and its argument is typically marked as comp:pred. Example: *John is the president* (here *president* is a comp:pred of 'is'). In Wolof, in a sentence like (13), *njiit* "leader" is considered the comp:pred of the verb *di* (to be) and Awa as its subject. The copula *di* links the two and is treated as the head of the sentence.

[4]Recall: that in (S)UD, prepositions belong to the category Adposition (ADP).

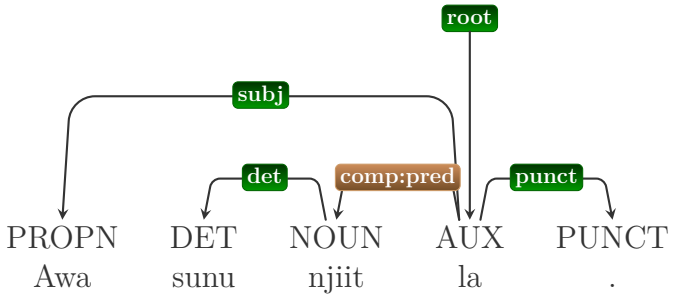
(13) *Awa di sunu njiit.*
Awa be our leader
'Awa is our leader.'



Example (14) shows a similar structure, except that

- the copula is ***la***, which puts some focus on the comp:pred.
- there is a change in word order: the comp:pred occurs before the copula.

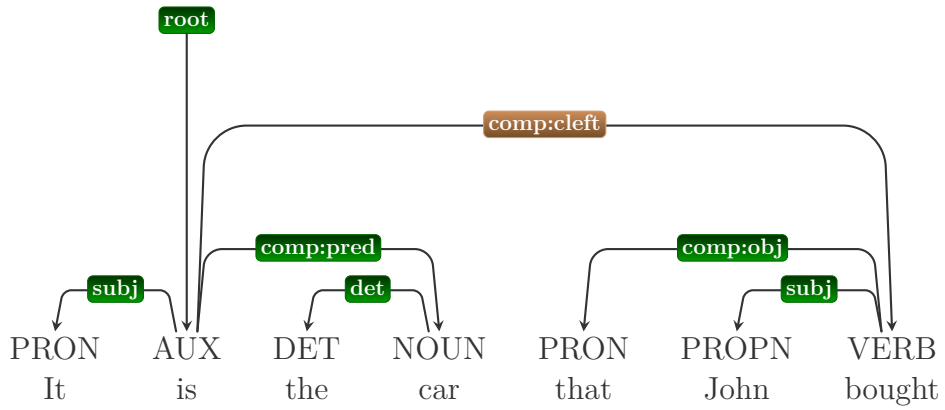
(14) *Awa sunu njiit la.*
Awa our leader be.3SG
'Awa is our leader (with a focus on leader)'



3.8.5 Relation comp:cleft

The **comp:cleft** relation is used in cleft sentences for the dependency from the head of the sentence to the head of the complement clause. An example of this relation is given in (15).

(15) It is the car that John bought



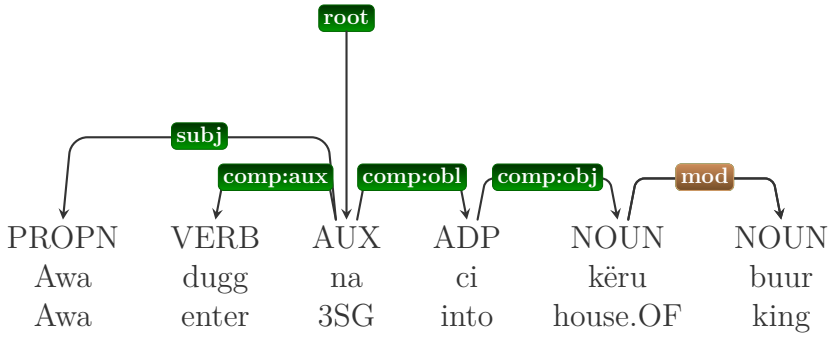
3.9 Relation: mod

The **mod** relation is used for **modifiers** of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions. For Wolof, it is used in the following cases: (1) noun modifiers, (2) non-clausal adverbial modifiers, (3) clausal adverbial modifiers, (4) clausal modifiers of noun, (5) numerical expressions.

Noun modifiers

The **mod** relation is used for nouns that modify another noun and correspond to an attribute of that noun or a genitive form. In Wolof, the genitive is often marked on the head. For instance, in the noun phrase *këru buur* 'king's house', the head is **kër** 'house' and it's the head which takes the genitive form (*-u-*). Thus, the noun *buur* is considered the dependent which modifies the head noun *kër* through the **mod** relation.

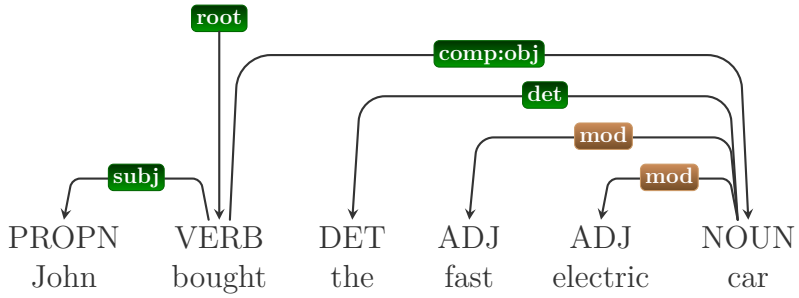
(16) *Awa dugg na ci* *këru* *buur*
Awa enter 3SG into house.GEN king
'Awa has entered the king's house'
Lit. 'Awa entered has into house.OF king'



Adjective modifying nouns

The **mod** relation is also used for adjectives that modify a noun. As Wolof lacks the category adjective, we will give an example from English.

(17) John bought the fast electric car

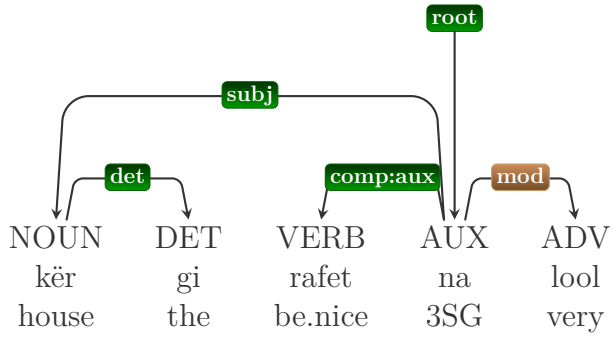


Non-clausal adverbial modifiers

The **mod** relation is also used when it comes to adverbs that modify an auxiliary, a verb, noun, adjective, adverb, ...

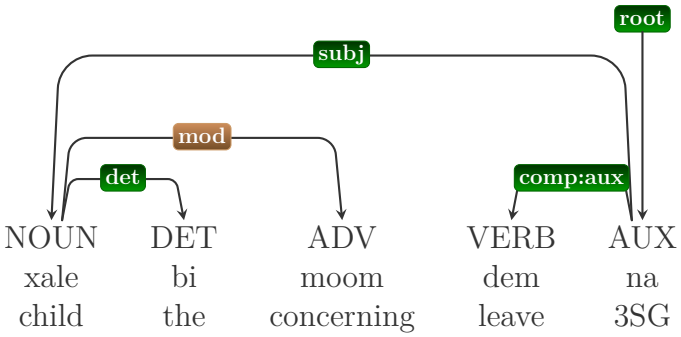
In (18), the adverb *lool* 'very' modifies the auxiliary *na*. Note that Wolof does not have the category adjective. In (18), the word *rafet* is a stative verb meaning 'to be pretty'.

(18) *Kër gi rafet na lool*
house the be.nice 3SG very
'The house is very pretty'



In (19), the adverb *moom* 'concerning' modifies the noun *xale* 'child'.

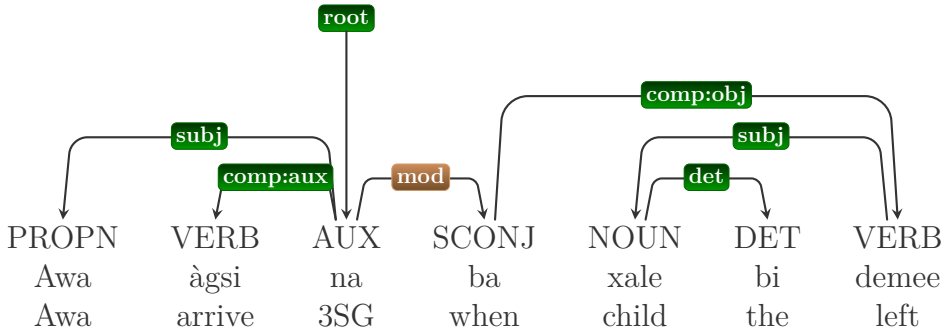
(19) *Xale bi moom dem na*
child the as for leave 3SG
'As for the child, he left'
Lit. 'child the concerning leave has'



Clausal adverbial modifier

The **mod** relation is also used when an **adverbial clause** modifies a verb or an auxiliary. This often includes temporal clauses, conditional clauses, etc. In the example (20), the clause *ba xale bi demee* 'when the child left' is a time adverbial phrase that modifies the auxiliary *na*.

(20) *Awa àgsi na ba xale bi demee*
Awa arrive has when child the leave.Perf
'Awa (has) arrived when the child left'



With respect to this example, note the following

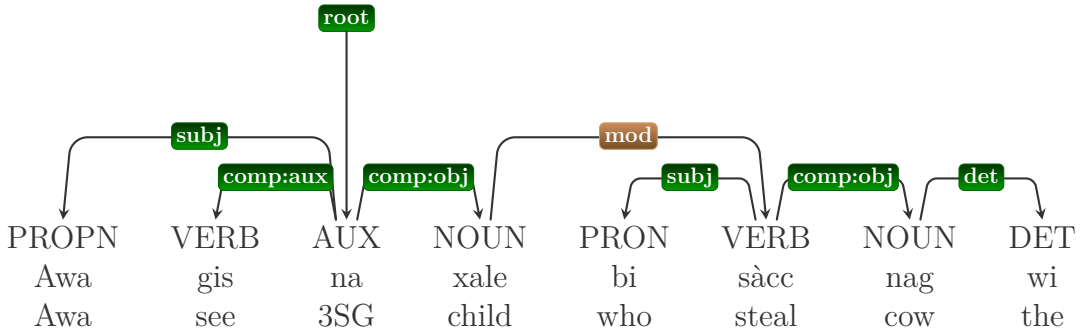
- SUD considers the SCONJ (e.g. *ba* "when") as the head of the subordinate adverbial clause.
- In turn, the SCONJ takes as dependent the main verb of the subordinate clause (e.g. *demee* "left") through the comp:obj relation.

Clausal modifier of nouns

The **mod** relation is also used when a clause modifies a noun (i.e. **relative clause**). For example, in (21), the underlined clause *bi sàcc nag wi* ('who has stolen the cow') is a relative clause that modifies the noun *xale* (child). In this case, the relation is *mod*, with *xale* 'child' as the head and the verb *sàcc* 'steal' as its dependent.[5]

(21) *Awa gis na xale bi sàcc nag wi*
Awa see 3SG child who steal cow the
'Awa has seen the child who has stolen the cow.'

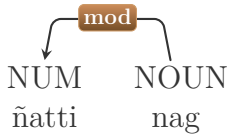
[5]Note that the word *bi* in Wolof can be a determiner and a pronoun. In example (21), *bi* is a pronoun (and not a determiner). As a relative pronoun, it is playing here the role of subject within the relative clause.



Numerical expressions

The **mod** relationship is also used for a numerical expression that serves to modify the meaning of the noun by a quantity. For instance, in (22), the word *ñatti* 'three' modifies the noun *nag*.

(22) *ñatti nag*
three cows
'three cows'

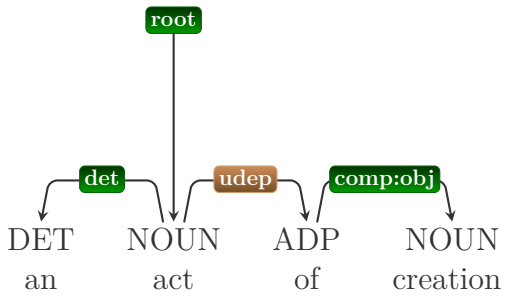


3.10 Relation: udep

The **udep** relation is used for complements when one does not wish to or is unable to distinguish between arguments and modifiers. It can notably be used in ambiguous cases where the head-dependent relation is not clear. However, when there is little debate about the nature of the relationship, the more specific label (e,g. comp:obl, mod, ...) is preferred. The head and the dependent of a udep relation can have any POS.

An example in English is given in (23).

(23) an act of creation

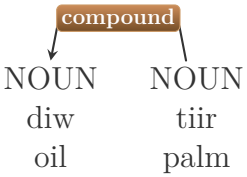


3.11 Relation: compound

The **compound** relation is used to analyze compounds, that is, combinations of lexemes that morphosyntactically behave as single words. There are various kinds of compounds, including (1) nominal compounds (e.g. *apple juice*), (2) numerical compounds, (3) particle verbs and (4) serial verbs.

Nominal compounds

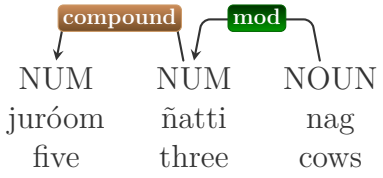
An example of nominal compound in Wolof is *diw tiir* 'palm oil' (Lit. 'oil palm').



Numerical compounds

The Wolof numeral system is based on the numbers 5 and 10. For instance, the number eight (8) is formed by combining five and three: *juróom ñett* 'five three'. In such a numeric expression, the first term (here *juróom*) is considered a **compound** which modifies the second one, i.e. *ñatt* 'three', as illustrated in (24).

(24) *juróom ñatti nag*
five three cows
'8 cows'



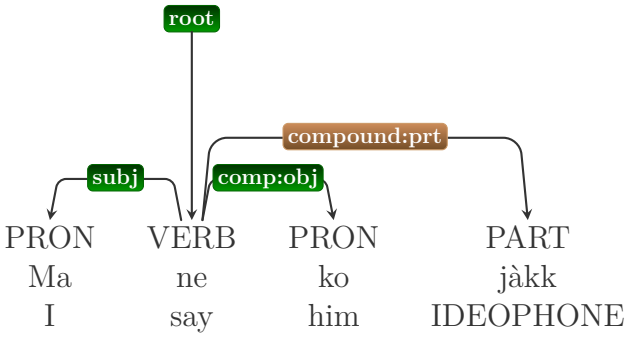
3.12 Relation: compound:prt

The phrasal verb particle relation identifies an idiomatic phrasal verb, and holds between the verb and its particle. It is a subtype of the compound relation. An example in English is the particle *down* as in *shut down* in (25).

(25) a. They **shut down** the station.
b. They **shut** the station **down**.

For Wolof, there are ideophones that behave like particles that are selected by the verb. An example is given in (26). The expression *ne jàkk* "to stare at someone" constitutes an idiomatic phrasal verb. In fact, the particle *jàkk* is an ideophone (meaning staring at someone) that only appears with the verb *ne*.[6]

(26) *Ma ne ko* ***jàkk***
I say him IDEOPHONE
'I was staring at him'



[6]The individual gloss is a bit misleading, because *ne* is translated as 'say', but actually, the translation only makes sense if we consider the combination of *ne* and *jàkk* as a single unit. Also note that we do not treat this expression as a serial verb construction (see below), because jàkk is not a verb in Wolof, but rather an adverb.

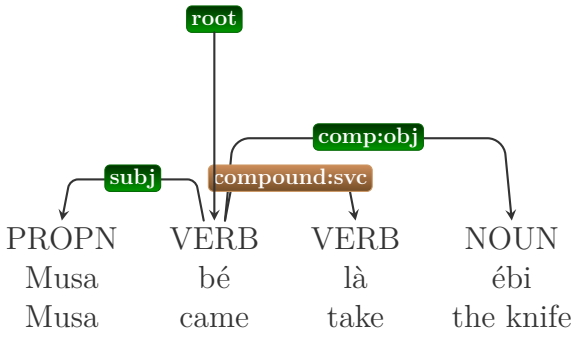
Note that the expression is not fixed (unlike those described in section 3.17). As this example shows, there may be some intervening elements (e.g. *ko* 'him') between the verb (*ne*) and the particle (*jàkk*).

3.13 Relation: compound:svc

The relation **compound:svc** is used for **serial verb constructions** (SVC). In this type of construction, several verbs are combined to describe the same action. An example of an SVC from the Nupe language (Nigeria) is given in (27).[7]

(27) *Musa* ***bé*** ***lá*** *èbi.*
Musa came took knife
'Musa came to take the knife.'

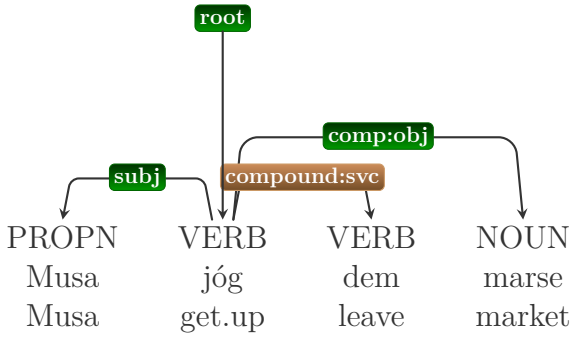
This sentence contains two verbs (*bé* 'come' and *là* 'take') that appear consecutively without any linking word between them that would indicate that one verb is subordinate to the other. The syntactic analysis of this sentence looks as follows:



Wolof does contain few of these kinds of constructions, as given in (28).

(28) *Musa* ***jóg*** ***dem*** *marse.*
Musa get.up leave market
'Musa got up and left for the market'

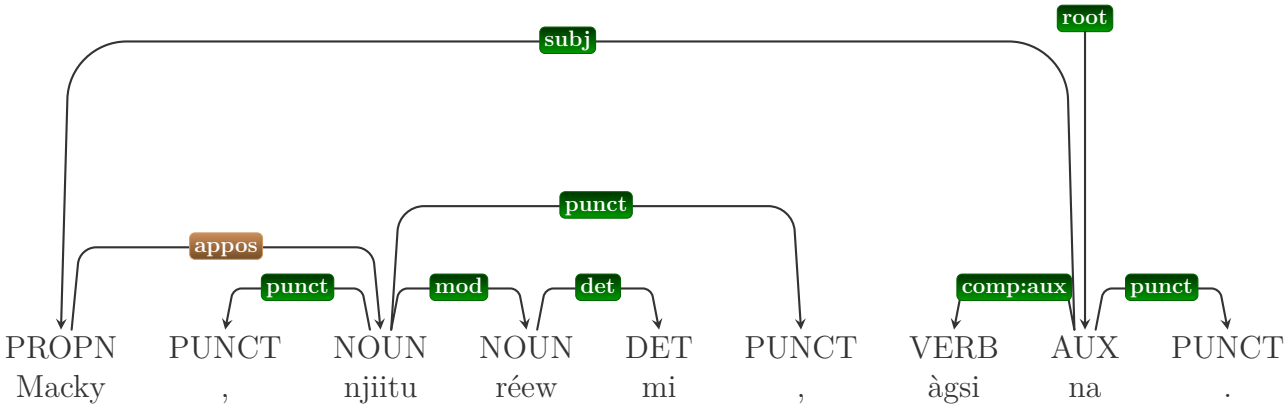
[7]This example is taken from Wikipedia.



3.14 Relation: appos

The relationship **appos** (appositional modifier) is used for an expression that immediately follows the first name and is used to define, modify, name or describe that name. It includes examples in parentheses, as well as the definition of abbreviations. For example, in the sentence (29), the expression *ñjiitu réewm mi* 'president of the country' is an appositional modifier of the name Macky.

(29) *Macky, ñjiitu réew mi, àgsi na.*
Macky leader.OF country the arrive 3SG
'Macky, the president of the country, has arrived.'



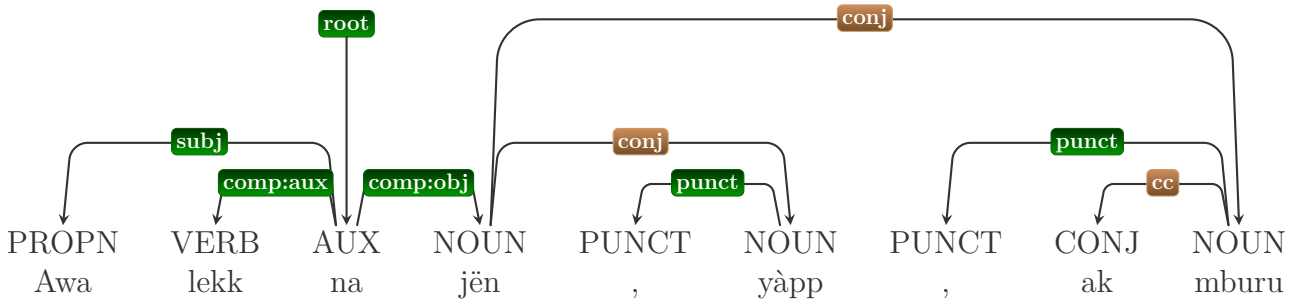
Note the attachment of punctuation here: the commas surrounding the appositional modifier attach directly to the head of this modifier, i,e. *njiitu* 'leader of'.

3.15 Relations conj and cc

The relation **conj** is the dependency relation between two elements linked by a coordinating conjunction, such as *and*, *or*, *but*, etc. The coordinating conjunction plays the role of **cc**. The rule in SUD is that the first conjunction is conventionally treated as the head of all subsequent coordinated clauses via the conj relation. Regarding the direction, the coordination link runs from **left to right**.

In the example (30), *jën* 'fish' is the head of the entire coordination structure. All coordinated elements are direct dependents of *jën*. The coordinating conjunction *ak* is a **cc** that depends on the last "conjunct" (here *mburu*). For punctuation, we generally attach them to the "conjunct" that comes immediately after.

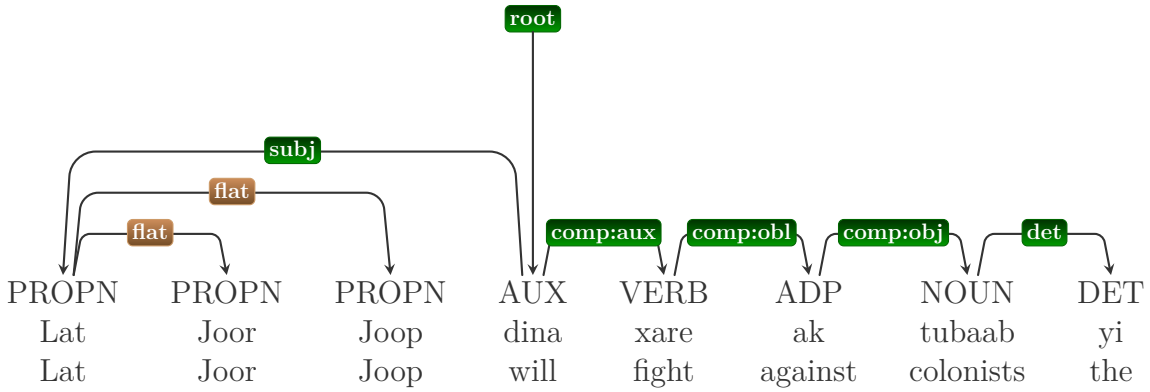
(30) *Awa lekk na jën, yàpp, ak mburu*
Awa eat 3SG fish meat and bread
'Awa has eaten fish, meat and bread.'



3.16 Relation flat

The **flat** relation is often used for personal names and surnames. In this structure, the first name is considered the head; and all other names are generally dependent on the first name, as illustrated in (31).

(31) *Lat Joor Joop dina xare ak tubaab yi*
Lat Joor Joob will fight against colonists the
'Lat Joor Joop will fight against the colonists.'

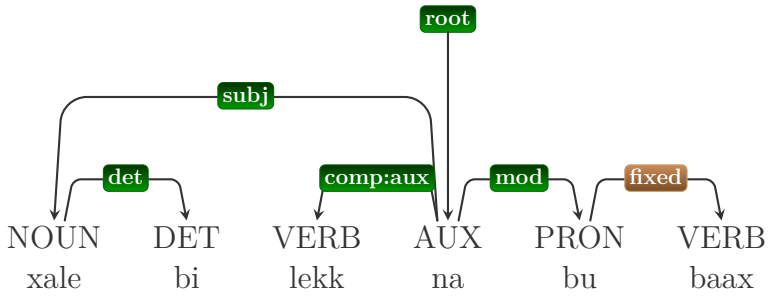


3.17 Relation fixed

The **fixed** relationship is used for certain **fixed** grammatical expressions. For instance, Wolof has adverbial expressions that are fixed in the sense that they constitute on unit and they do not allow an intervening element in-between. Here are some examples:

- *bu baax* 'well'
- *lu dul* 'unless'

(32) *Xale bi lekk na bu baax*
child the eat 3SG well
'The child ate well.'

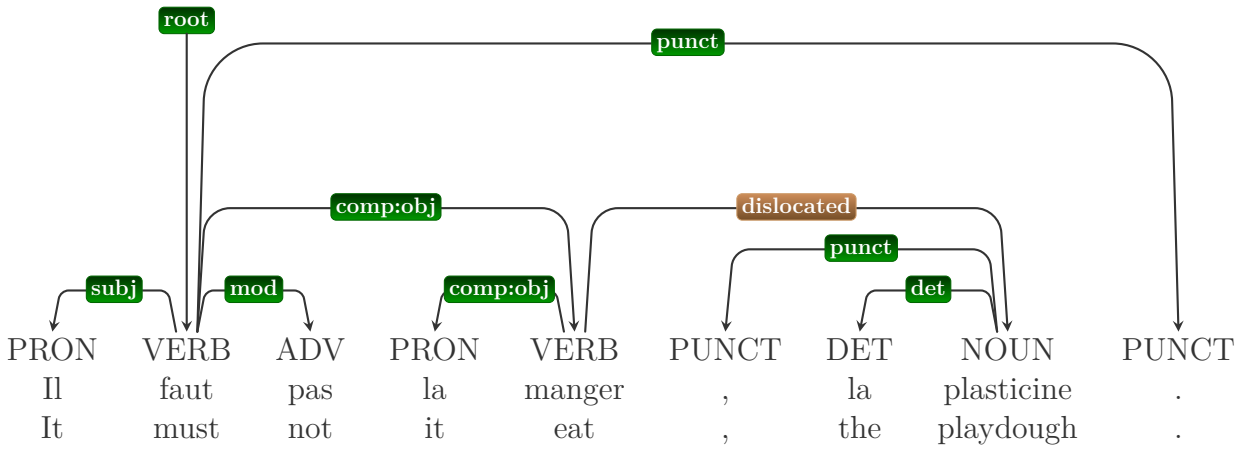


3.18 Relation dislocated

The **dislocated** relation is used for fronted or postposed elements that do not fulfill the usual core grammatical relations of a sentence. These elements

often appear to be in the periphery of the sentence, and may be separated off with a comma or intonation. Example (33) illustrates dislocated elements in spoken French. The dislocated element (here playdough) attach to the same head as the dependent that it doubles for (here *la* 'it'). Generally, the dislocation punctuation symbol (here comma) attaches to the dislocated element (playdough).

(33) *Il faut pas la manger, la plasticine.*
It must not it eat, the playdough
'You must not eat the playdough.'



3.19 Other SUD relations

There are other syntactic relations used in SUD that are not discussed in the guidelines. These include (links to UD related page are given): discourse, parataxis, and orphan.