

# Annotation decisions

- Dealing with non-sentences
- Possessive pronouns
- Uninflected "modals"
- Are infinitival verbs, verbs or nouns?
- Copulas?
- Relative pronouns
- Multiple agreement?
- CCOMP vs XCOMP
- Juu can be used as a noun?
- -enye
- Reduced relative clauses
- kuwa na
- List of PARTicles
- List of fixed expressions
- tu
- Auxiliaries
- Verbal nouns with auxiliaries
- Tense?
- Verbal interrogatives

# Dealing with non-sentences

This page will feature the decisions we make during the annotation process

Non-sentence constructions will be annotated to their highest level of structure

E.g. if the given text is only a noun phrase and not a complete sentence, it will be annotated as a noun phrase and the root will be the noun.

The only exception to this is when there are two constituents in the structure but they have no relation to each other. In this case, annotation is skipped.

# Possessive pronouns

## Possessive pronouns

The UD annotation guidelines state:

“ In some of the datasets, a possessive determiner like [en] my is currently given the POS tag DET but the relation nmod, so that it is parallel with other possessive constructions. This is not yet completely parallel across languages; in some languages, it is much more clear than in English how possessive determiners relate to adjectives, and the nmod relation is out of question

All possessives should be changed to have nmod arcs coming in but DET as the part of speech.

# Uninflected "modals"

Lazima and other modal adverbs will be marked as adverbs and connected with an advmod relation to the verb of the clause they are in.

# Are infinitival verbs, verbs or nouns?

## Infinitive verbs should always be treated as verbs, even when they have nominal modifiers

The UD documentation states

“Note that some verb forms such as gerunds and infinitives may share properties and usage of nouns and verbs. Depending on language and context, they may be classified as either VERB or NOUN.

We, therefore, need to determine when these will be treated as verbs and when they will be nouns. Originally, we were operating under the principle that if an infinitive verb was modified by nominal modifiers, it should be marked "NOUN" with "VERB" being the default. However, there are numerous cases where the infinitival has both nominal modifiers and elements in the verbal subcategorization frame (e.g. objects; subjects are illicit).

Because the verbal morphological features cannot be (validly) represented on nouns, all infinitives/gerunds are verbs.

# Copulas?

## Inflected copulas will be marked AUX and given a cop arc

copulas, will be marked as AUX and have an incoming cop arc, just as the simpler copula *ni* would. Though the UD documentation states:

“The cop relation should only be used for pure copulas that add at most TAME categories to the meaning of the predicate, which means that most languages have at most one copula, and only when the nonverbal predicate is treated as the head of the clause.

According to Mohammed, (2001), there are several forms of copulars.

Ni

---

predication without a verb

used to indicate present time references

Ni can also be at the beginning of the sentence to indicate emphasis.

ni yeye anayepika sasa

be 3S 3S-PRES-3S.REL-COOK now It is she who it cooking now

In these constructions, the verb cooking has a relative marker 'ye' indicating that it is

# Si

Functions very similar to Ni but indicates a negated copula

# Ndi-

an emphatic version of the ni copula with noun class/person/number agreement with the subject of the copula.

# Si-

An emphatic version of the Si copula with noun class/person/number agreement with the subject of the copula.

# -ko

a locative copula indicating location without a specific reference.

## -po

a locative copula indicating location with a specified reference.

## -mo

a locative copula indicating location inside or around something.

## -na

these indicate location in combination with a locative. (kuna, pana, mna).

## yu

A rare copula that can be used when the subject is 3rd person animate.

## u

A rare copula that can be used when the subject is 3rd person human.

# Is kuwa a copula?

kuwa is a weird one. It semantically is similar to a copula but it can have so many different morphemes added like a regular verb. It's also not the reduced forms that we see for the other



copulas.

I will still treat kuwa as a copula because in other languages with inflected copulas like German, they still use the cop relation. (though German doesn't have both inflected copulas and relatively bare copulas).

# Relative pronouns

## Relative pronouns

In cases where relative pronouns like *ambayo* are used, they should be assigned a PRON part of speech tag and use the PronType feature to indicate they are relative pronouns. I currently have these marked as SCONJ and they should not be.

The relative pronouns should not be attached with a **mark** relation as the documentation points out

“ it is a normally uninflected word, which simply introduces a relative clause, such as [he] *še*. (In this last use, one needs to distinguish between relative clause markers, which are **mark**, from relative pronouns such as [en] *who* or *that*, which fill a regular verbal argument or modifier grammatical relation.).

Then the root of the relative clause is connected to the noun the relative clause modifies by an **acl** relation.

## Wasio/walio etc.

These are not actually relative pronouns. Instead, they are pronouns derived from the verb "to be". They are not part of a relative clause, instead they are simply nouns so the dependents of this are the same as any nominal dependent.

# Multiple agreement?

In cases of multiple agreement, the first verb is connected to the second with a ccomp relation.

- For a thorough discussion of this, the beginning of Carstens (2002) is useful

Here's an example of where this happens:

```
watoto hao waliokuwa wakijiandaa kulitumikia taifa lao (sentence 28861)
```

Notice that both waliokuwa and wakijiandaa are inflected (though note that in this case it's more appropriate to connect watoto to wakijiandaa with a acl relation and then connect wakijiandaa to waliokuwa since this is a copula. (Though note that I'm not sure if this is an example of translationese as I don't think this would normally require an auxiliary but I could be wrong).

# CCOMP vs XCOMP

The UD documentation states

“ Clausal complements (objects), divided into those with obligatory control (xcomp) and those without (ccomp).

I assume obligatory control in cases where there is no subject (no overt subject or subject marker on verb), these should all be xcomp instead of ccomp.

# Juu can be used as a noun?

Madaktari Wasiokuwa na Mipaka waliitisha mkutano wa waandishi wa habari baada ya habari hiyo, lakini José Antonio Bastos, Rais wa DWB Hispania, alisema hawangetoa taarifa juu ya mchakato wa ukombozi "ili kutokusababisha matatizo kwa watu wengine wanaojitolea nchini Somalia pamoja na watoa habari."

Juu in this sentence seems to be saying like official statement (taarifa juu). I connected taarifa to juu using nmod and changed juu from ADV to N and added features for

**NounClass=Bantu9|Number=Sing**

# -enye

“ ukurasa wenye zaidi ya wafuasi 4,5000 (sentence 8521)

wenye was labeled as SCONJ but I think i made that judgment based on parallels with English not for good reasons. Looking at it, it seems similar to 'wa' but with the meaning that the thing is possessed by the first thing not the second.

I should go back through everything and make 'enye' an ADP and give it a case connection.

# Reduced relative clauses

Vitale identifies three types of relative clauses, "full relatives" which use *amba-*, "reduced relatives" which use a verbal prefix to indicate realtiveness (e.g. the *ye* in *a-li-ye-mw-ona*).

The third type is reduced relative clauses which have a relative marker suffix.

wanafunzi wa-soma-o  
students they-study-REL  
'students who study'

sentensi zi-fuata-zo  
sentences they-follow-REL  
'sentences which follow'

Note that these have no tense information.

The words that were labeled with REL-LI for their HCS part of speech are actually the copula version of these reduced relatives.

e.g.

par	gloss	system	sch
spee			
which	REL	SUB-	
yalio	LOC	INTR-	
there	PLSG	PREF=6-	
which	REL	SUB-	
ulimo	LOC	INTR-	
therein	SG	PREF=11-	
where	REL	SUB-	
walio	LOC	INTR-	
are	PL	PREF=2-	

par  
gloss  
system  
spee

where  
iliyo  
is  
REL-  
@SUBJ  
SUB-  
REF=9-  
SG

which  
iliyo  
there  
REL-  
@SUBJ  
SUB-  
REF=9-  
LOC  
IN  
Vintr-  
PREF=9-  
SG

which  
kili  
therein  
REL-  
@SUBJ  
SUB-  
REF=7-  
LOC  
IN  
Vintr-  
PREF=7-  
SG

where  
yaliyo  
is  
REL-  
@SUBJ  
SUB-  
REF=6-  
PLSG

where  
alipe  
is  
REL-  
@SUBJ  
SUB-  
REF=1-  
SG  
16-  
LOC

which  
alipe  
has  
REL-  
@SUBJ  
SUB-  
REF=1-  
PL  
CC  
PRON  
SUB-  
REF=1-  
PL  
SG3  
4-  
PL

where  
uliye  
are  
REL-  
@SUBJ  
SUB-  
REF=SG2

where  
ulio  
is  
REL-  
@SUBJ  
SUB-  
REF=3-  
SG

which  
tuma  
have  
REL-  
@SUBJ



# kuwa na

"have" is expressed using a copula "kuwa" with a preposition "na" (with).

In these cases, the object that follows "na" is the root of the clause and is connected to "na" with a "case" relation and "kuwa" with a "cop" relation.

# List of PARTicles

- the question particle **je**

# List of fixed expressions

## Compound prepositions (Mohammed, 2001)

- baada ya: after
- kati ya: between
- ndani ya: inside of
- mbele ya: in front of
- karibu ? : near
- pamoja na: together with
- mahali pa: instead of
- juu ya: on
- kutoka kwa: from
- kwa ajili ya: for the sake of
- shingoni mwa: around the neck of
- ukingoni mwa: along the bank of
- mbali na: far from
- nyuma ya: after
- katikati ya: among
- kwa habari ya: about
- zaidi ya: more than
- nje ya: outside
- kwa sababu ya: because of
- chini ya: below, under

# Adverbial fixed expressions

- hata hivyo: consequently

# tu

"tu" meaning "only" or "just" should be the dependent of the thing it is modifying.

In cases like jambo moja tu (just one problem), tu, should be an advmod dependent of the nummod *moja*.

# Auxiliaries

## kuwa

“ used to refer to the protracted nature of an action or action at a definite moment in the past or in the future.

Has this interpretation when the verb that follows is modulated by markers like '-ki-' or '-na'.

## weza

“ assumes the meaning of potentiality and possibility at some point in the past, present, or future.

the full verb that follows is infinitive.

## pata

“ implies ability or opportunity for a subject to accomplish a particular thing or, on the contrary, suffers a stroke of misfortune by the occurrence of the action denoted by the main verb.

verb stem that follows can either be infinitive or not.

# kuja

“ used to refer to an action that will take place at an implied time in the near or distant future

verb stem that follows is likely infinitive.

# taka

“ suggests the meaning of assurance that a desire or purpose will definitely or most probably be fulfilled.

- Can be followed by bare verb stem or *ku* + verb stem.
- Mgonjwa ataka kunywa maji (the patient wants to drink water)
- Mwizi yule ataka toroka (that thief wants to escape)

# kwenda

1. when used with me- li- tense, it assumes the meaning of an action being carried out at the time indicated in the context (concurrent actions).
- Hamida amekwenda kuleta chakula (Hamida has gone to get food.)

2. When used with or without a subject prefix and with the relative *po*, the verb indicates something like "should it happen", "if by chance"
  - Wendapo hutakuja shuleni kesho, mwalimu wako atahamaki (and if it happens that you do not come to school tomorrow, your teacher will get angry)
3. When followed by -ka- or -me- the word huenda means "perhaps", "Pmaybe"
  - Huenda mvua ikayesha leo. (it may rain today)

# kwisha

“ refers to a state of existing or action completed before the point in time indicated in the context.

- simba amekwish kuuliwa (the lion is already killed)
- Walimu wamekwisha fika mkutaoni (the teachers have already arrived in the meeting)



# Verbal nouns with auxiliaries

If we have a construction like "kuhusu kilichokuwa kikitokea" (sentence 5202 in the global voices data), it becomes thorny. UD wants us to not have auxiliaries as the heads of their clauses, instead the verb that follows or in the case of copulas, the noun that follows should be the head of the clause. However, in a construction like this, kilichokuwa is a verb with relative marking but no overt noun that this simple relative is modifying.

As such, this is a derived noun/nominal relative clause. It is functioning as a noun (the preposition kuhusu is a case dependent of it). Because it is a derived noun, it is contentful and not purely functional. Thus, **kilichokuwa** is the head of a noun phrase. If the verb that follows (**kikitokea**) was another relative, then the verb that follows would be an acl dependent of **kilichokuwa**. However, in this case it is not, so it **kikitokea** is a ccomp dependent of **kilichokuwa** instead.

# Tense?

## -ki-

M.A. Mohammed reports that *ki* has several features that it could be assigned as a tense marker.

*ki* in a single verb indicates conditional use (Mood=Cond).

*ki* as part of a complex verbal construction (as the second verb in that construction) indicate the imperfective, continuous or incomplete. (Aspect=Imp)

## -ka-

Indicates consecutiveness of a sequence of events.

# Verbal interrogatives

Verbs like unawezaje where the -je suffix indicates the verb is an interrogative are expressed by the feature Polarity=Int.