Swahili Dependency Treebank Creation

This describes my thesis research and related research surrounding the creation of a universal dependencies treebank of Swahili.

    - 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
- Errors made by neural models
    - Interrogative adjectives
    - Hashtags need to be rejoined
    - -ote
- Reduplicated words
- Neural models
- Publications to cite
    - A multilabel approach to morphosyntactic probing
- Swahili transformer model bake off
- Meeting with Sandra 1/6/2022
- Low resource dependency parsing in Swahili
    - Experiment list
    - Implementation log

Swahili in UD

UD QA session guidelines 2-5-2020

https://padlite.spline.de/p/clingdingud

Questions:

1. Is Obj used for central arguments in terms of subcategorization frames? For example, 'put' requires a prepositional phrase location, would this be an obj or obl?
   - Essentially no, @obj is used for unmarked/core dependents of predicates, it corresponds to "second core argument" or "most patient-like argument"
   - https://universaldependencies.org/u/dep/all.html#al-u-dep/obj
   - iobj for Bantu languages with applicative extension is okay even though it expresses non-core arguments like benfeciaries and instrumentals as this is indicated by the verb's morphology (this example is specifically called out in the UD documentation)
   - https://universaldependencies.org/u/dep/all.html#al-u-dep/iobj

2. What should you do with things that are not really full sentences? (e.g. newspaper headlines or photo captions)
   - annotate them as if annotating fragments
   - try to go to the highest level of structure possible

3. Can you have multiple case arcs leaving a noun? "The ball rolled from under the chair" . Would that be a compound?
   - Look it up in the English treebank and see
     - Looks like the English example does case to the closest prep and then dep from that preposition to the next preposition

    - english GUM and english lines have examples "from over" / "from under"
  - Probably going to be flat with two case arcs
  - Add that to the UD github issues page

4. In case of polypersonal agreement, the basque treebank used Number[nom], Number[dat] etc for different cases. This seems to be a case driven approach but what if you have a language with no case system?
   - Number[obj] / Number[subj]

5. The distinction between fixed and compound seems fuzzy. Is it basically that compound is used for matching pos tags?
   - If the syntactic relationship between two words is unclear then using fixed is likely a good solution
   - compound is almost always only used for noun noun compounds

Xibe

1). How to calculate the annotate agreement between annotators?

- annotate the same sentences

2. Auxilaries: ombi (to become), sembi (to call), bimbi(to have) . The current annotation: no matter what words are in front of those auxilaries, we all annotate them as AUX.

ex. terei tacin tesei banse de, uju waka oci geli jai ombi. His study their class DAT, first is-not AUX also two AUX. (root of this sentence is 'jai', and 'ombi' depends on 'jai')/

```
  Do we need to annotate them differently?   a. when there is another VERB before these auxilaries, we annotate them as AUX.  b. when there are ADJ, NOUN before the auxilaries, we annotate them as VERB.
```

3). pospositions

Since the case markers are annotated as ADP, there are lots of pospositions, they usually need to collocate with certain case to convey a meaning. Ex: aimaka inenggi šun i adali eldešembi. like

day light ADP ADP shine.

4).

mini gebu be Mutešan sembi. my name ACC Mutešan call. My name is Mutešan.

after 'my name' there is ACC marker, so 'obj' should be object of 'sembi', what is 'Mutešan' then? nsubj?

5). We have several words, the POS in both the dictionary and grammar book are not persuasive. Can we decide by our own linguistic knowledge? For example;

ilanofi ('ilan-nofi', three people, three things), it looks like a noun, in the grammar book, it is a NUM.

akv (is not + ADJ), waka(is not + NOUN), in dictionary, it is NOUN, but now we annoate them as VERB, and has a relation of 'cop' with the words in front of it.

On the issue of clitic vs morpheme

Sandra suggests contacting Dr. Botne to ask about whether the pre-verbal markers in Bantu languages are morphemes or separate clitics.

I can also prepare a questionaire to give to someone like Dr. Omar to see if certain things are possible with clitics in Swahili.

Still need to read 'on clitics'.

Sandra says this will be a good paper for TLT next summer.

On Clitics:

Diagnostics for whether you have a clitic:

- Ordering relative to other markers can change (p. 2)
- Phonological rules that apply within word boundaries do not apply
- If the marker is bound to the root, it is a morpheme
- If a morpheme is in construction with an affix it is either a base or an affix
- '' Proper parts of words do not undsergo rules of deletion under identity''
  - Means that yellowish or grayish happens not yellow or grayish
  - E.g. there is no deletion when coordinated with a similarly inflected component
- Markers that are are not accented are affixes

''Binding, Construction with affixes and accent are the most suceptible to attack''

Typology of clitics:

First class:

The ''unaccented bound form acts as a variant of a astressed free form with the same cognitive meaning and with similar phonological makeup''

- The unaccented bound unit is said to be ''conjoint'' while the other is said to be ''independent'' or ''strong''.

Second Class:

''Cases where a free morpheme, when unaccented may be phonologically reduced, the resultant form being phonologically subordinated to a neighboring word'' (p. 5)

Third Class:

'' Cases where a morpheme that is always bound and always unaccented show considerable syntactic freedom in the sense that they can be associated with words of a variety of morphosyntactic categories.''

''Frequently, such a 'bound word' is semantically associated with an entire constituent while being phonologically attached to one word of this constituent and ordinarily the bound word is located at the very margins of the word, standing outside even inflectional affixes'' (p.6)

How are verbal markers discussed in the literature?

The Bantu Languages 1st ed ()

> “

> Bantu languages are agglutinating. Verbs have an elaborate set of affixes. Most Bantu languages have non-derived and derived nouns, the latter having an inflectional prefix and a derivational suffix. For verbs and nouns, the conventional analysis (Ch. 5) starts with an (abstract) root/radical, most often of the shape -(i)CV(C)-. For nouns a stem is formed by the addition of a derivational suffix (mostly consisting of a single vowel). For verbs, an (abstract) base may be derived from the root, via the suffixation of an extension, and the addition of a final inflectional suffix then provides a stem, to which pre-stem inflection is added. The set of suffixes is limited, for nouns and verbs. For nouns a class prefix is then added, and in some languages, a pre-prefix. All nouns are assigned to a class. Over twenty such classes are reconstructed for PB, although most of today's languages have between twelve and twenty (a few languages reduced or even eliminated classes, see e.g. Chs 15, 16, and especially 23). A class is characterized by: a distinct prefix, a specific (and characteristic) singular/plural pairing (a 'gender'), and agreement with other constituents. During most of the twentieth century the semantic arbitrariness of noun classes was emphasized but recent years have seen attempts to find semantic generalizations. Bantu languages have been described as verby. The verb is pivotal in the sentence, it incorporates much information, and may stand alone as a sentence. Nearly all Bantu lan- guages are prodrop. In many languages verbs have six possible pre-stem positions, and since some of these may be filled by more than one morpheme, it is often possible to get a string of a dozen or more morphemes in one verbal word. -- p. 8

> "

> However, this diversity must not be allowed to obscure the fundamental principle of dou- ble representation which is common to most Bantu languages, albeit to greatly varying degrees: information coded in the syntax of the sentence may be cumulatively or alter- natively coded in the verb morphology. Any sentence can be reduced to its verb without ceasing to function as a sentence --p 125

The Bantu Languages 2nd ed ()

> “ Preceding a verbal stem, we may find prefixes specifying, amongst other things, the person or noun class of the subject as well as inflectional categories such as time, aspect, negation, etc. (p 175)

> “ Slots to the left and right of the Root and Extension(s) in (1) involve inflection. (p 205).

> “ Bantu languages with grammatical agreement-like object marking (Stage II) include, for example, Swahili G42. Swahili object markers can serve a pronominal function, resuming objects mentioned earlier in the discourse, but they also serve a grammatical agreement function: object marking is (almost) obligatory with overt human objects (especially definite ones), as in (21a), and common with definite non-human objects, as in (21c). (p 275)

> “ Riedel (2009: 42, 46) affirms that the object marker in these examples occurs even though the co-referential object noun phrase is in its base position, not

> dislocated. This means that class 1/2 object markers in Swahili can be analysed as agreement markers (22). (p 275)

> “However, there are also arguments against confounding subjects and topics. Most generative analyses have, in fact, rejected the Bresnan and Mchombo (1987) proposal that the subject marker may function as an incorporated pronoun. For one thing, subject marking is typically obligatory in inflected verb forms, and obligatoriness is more con- sistent with morphological agreement than with syntactic incorporation. Furthermore, in complex verb phrases, several subject markers can occur in agreement with one subject. (p 281)

A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages (Taljard & Bosch 2006)

> “In this article it is argued that the different orthographic systems obscure the morphological similarities and that these systems impact directly on word class tagging for the two languages.

The authors argue that different approaches are needed for word-level tagging when working with disjunctive and conjunctive languages as a result. In particular for the disjunctive language (Northern Sotho), no separation of morphemes has to happen, the tagger is doing both part of speech tagging and something resembling morphological analysis, then some word-formation

rules would need to be applied to determine how the small disjunctive pieces fuse together.

> “In the case of Zulu, the morphological analyser plays a significant role on levels I and II where constituent roots and affixes are separated and identified by means of the modelling of two general linguistic components. The morphotactics component contains the word formation rules, which determine the construction of words from the inventory of morphemes (roots and affixes). This component includes the classification of morpheme sequences. The morphophonological alternations component describes the morphophonological changes between lexical and surface levels (cf. Pretorius & Bosch, 2003: 273–274).

> “Finally, Northern Sotho and Zulu are on a par on level III, where the identification of word classes, associated with the assigning of tags, takes place.

Annotation Issues

Non-copyright encumbered corpora

SketchEngine

Sandra is contacting the owner of the SketchEngine Swahili data to see if we can get a license that allows us to release our annotated data.

Global voices corpus

- available in opus
- non-copyrighted

Unannotated version of helsinki corpus is under CC by 4

Research questions for thesis

- How do Bantu languages, and Swahili in particular, integrate into the Universal Dependencies framework?
  - Being as there are no existing Bantu treebanks at all, what are the relevant syntactic phenomenon to consider when making a treebank?
    - Discuss relevant morphological phenomenon in Swahili (for the morphological features in UD)
    - Discuss some relevant syntactic features (for dependency relations)
    - Provide some suggestions for how other Bantu languages could be handled in UD as part of an effort to encourage future treebanks for Bantu languages.
    - Discuss additional Bantu syntactic phenomenon using resources on Bantu syntactic typology.
- How do these phenomena fit into the universal dependencies frame- work?
  - How is the specific corpus developed for this work created?
    - Automatic POS tagging, morphological tagging, lemmsiatization and functional tagging using the Helsinki corpus
    - Annotation procedure (e.g. sampling and sentence selection procedure)
    - Automatic labelling procedure
- What considerations need to be taken to allow for accurate parsing using a relatively small dataset of mixed origin?
  - What strategies can be used to enable parsing on a small dataset?
  - Given that the treebank produced is partially created by rules and partially hand annotated, what is the best way to integrate the hand annotated trees with the trees created using rules?
    - How should the hand annotated data be integrated into train, eval, and test data splits?

Annotation Issues

Annotation Issues

relative clauses without overt modifiers

Consider this noun phrase: Idadi ya waliofariki (from sentence #6799)

the issue is that waliofariki is a relative clause meaning something like "who have died".

The issue is that there seems to be an elided "watu" which serves as the thing that waliofariki modifies. An alternative, is that waliofariki is a noun derived from fariki. E.g. it's those who have died not who have died.

Currently, I am treating these as nouns derived from verbs. (see # 5260)

case with no noun?

What the heck is up with `mpaka hivi sasa` ?

Why do I have an adposition modifying an adverb???

Things to go back and fix in manually annotated corpora

- check that iobj is used correctly. E.g. check that verbs which could have iobj but weren't marked with one, don't have one and check that all use of iobj is appropriate.
  - Follow up: see if HCS has iobj indicated somewhere on nouns. this doesn't seem to be making it through the tagger. NO rules are currently leveraging iobj.
- mark should be used when "prepositions" precede a clause. E.g. after he went outside, after should be connected to went with a mark
- ni is always a copula. kuwa can be sconj. verbal kuwa is a verb. Go back and fix uses of kuwa.
- any cases where there's a hyphenated demonym, use the version generated by the rules and correct it, the tokenizer was fixed to treat these correctly.

Additional things to go back and fix.

gani and ngapi should be changed from ADJ to DET, have their arcs changed to be det and nummod respectively and have PronType=Int added as a morph feature.

ka should be assigned continuative aspect?

a- indefinite tense marker should be assigned some special features.

check that hu- is assigned habitual aspect.

Check that ki- TAM marker is conditional mood.

All infinitive verbs should be given VerbForm=Inf and assigned verbal dependencies. Other morphological features will also need to be adjusted.

When reduplication happens, the second reduplicant is the head source.

Progress

- Longer sentence sample:
  - ka has been assigned continuative aspect (1 examples).
  - hu has been assigned habitual aspect (no examples).
  - ki has been assigned conditional mood. (5 examples).
  - ni has been assigned copula.
  - Verbal kuwa has been assigned verb status.
  - no hyphenated demonyms left to change.
  - Prepositions with verbal heads are now using the deprel mark (1 change)
  - gani fixed (3 changes)
  - ngapi fixed (no changes)
- Shorter sentence sample:

SWH in UD questions

should a derived noun like waliokusanyika be a Vnoun?

should statives be

what is kuna/hakuna? is it a compound (kuwa + na) or is it related to kuna from arabic? should it be a pleonastic or a copula?

Annotation decisions

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 constitutents in the structure but they have no relation to eachother. 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 becuase in other languages with inflected copulas like German, they stil 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 wkijiandaa 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.

Annotation decisions

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.

word	gloss	part of speech	synt	lemma	msd
yaliyoko	which is there	REL-LI loc	@FMAINVintr-	-LI LOC 17	SUB-PREF=6-PLSG
ulioimo	which is therein	REL-LI loc	@FMAINVintr-	-LI LOC 18	SUB-PREF=11-SG

wo	gloss	par spee	synt	de	msd
walio	who are	REL-LI	@SUBJ		SUB-PREF=2-PL
iliyo	which is	REL-LI	@SUBJ		SUB-PREF=9-SG
iliyopo	which is there	REL-CL-INVintr-LI loc	@-FLOC	16	SUB-PREF=9-SG
kilichomo	which is therein	REL-CL-INVintr-LI loc	@-FLOC	18	SUB-PREF=7-SG
yaliyo	which is	REL-LI	@SUBJ		SUB-PREF=6-PLSG
alipo	where he is	REL-LI	@SUBJ		SUB-PREF=1-SG 16-LOC
aliyenayo	which he/she has	REL-LI	@SUBJ	CC PRON CC-PL 4-PL	SUB-PREF=1-SG3
uliye	who are	REL-LI	@SUBJ		SUB-PREF=SG2
ulio	which is	REL-LI	@SUBJ		SUB-PREF=3-SG
tulionao	which we have	REL-LI	@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 potentiallity 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 occurance 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 thime 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.

Errors made by neural models

Interrogative adjectives

It may seem strange but **gani** is treated as an interrogative adjective in the Helsinki corpus of Swahili and by Mohammed (2001). This is probably due to analogy with **-pi** and **-ngapi** which are both inflected with adjectival concord.

Hashtags need to be rejoined

The tokenizer used split pound signs from the rest of the hash tag: #NairobiBlast -> # NairobiBlast.

These should all be rejoined.

-ote

-ote meaning 'all' is nearly always labeled as adv instead of det like it should be.

Reduplicated words

Any["hivyohivyo", "yuleyule", "shamrashamra", "kwelikweli", "hiyohiyo", "hatihati", "kimyakimya", "wasiwasi", "zilezile", "ndogondogo", "vilevile", "takataka", "mdogomdogo", "fupifupi", "humuhumu", "vuguvugu", "pilipili", "hawahawa", "pilikapilika", "kwambakwamba", "mmojammoja", "sawasawa", "fulanifulani", "hichohicho", "hizohizo", "mbalimbali", "kandokando", "tofautitofauti", "motomoto", "katakata", "waziwazi", "kunakuna", "madogomadogo", "majimaji", "palepale", "mbogamboga", "chembechembe", "polepole", "ombaomba", "katikati", "vidogovidogo", "hilohilo", "barabara", "VuguVugu", "hekaheka", "sanasana", "washawasha", "mafupimafupi"]

Neural models

za3.txt is test.txt na2.txt is val.txt

Publications to cite

A multilabel approach to morphosyntactic probing

Swahili transformer model bake off

pretrained swahili transformer, lr=0.0001, pos tagging


```
2022-06-11 17:40:51,753 - INFO - allennlp.common.util - Metrics: {
  "best_epoch": 4,
  "peak_worker_0_memory_MB": 5448.7890625,
  "peak_gpu_0_memory_MB": 9332.94873046875,
  "training_duration": "1:26:00.349669",
  "epoch": 9,
  "training_accuracy": 0.996923352964769,
  "training_accuracy3": 0.9999548500435081,
  "training_precision": 0.9857814311981201,
  "training_recall": 0.9809743165969849,
  "training_fscore": 0.9833330512046814,
  "training_loss": 0.008818848443899869,
  "training_worker_0_memory_MB": 5448.7890625,  "training_gpu_0_memory_MB":
9332.94873046875,
  "validation_accuracy": 0.9915935820231985,
  "validation_accuracy3": 0.9994341834054076,
  "validation_precision": 0.9857875108718872,
  "validation_recall": 0.9809861779212952,
  "validation_fscore": 0.9833420515060425,
  "validation_loss": 0.06279067079275392,
  "best_validation_accuracy": 0.989734470355252,  "best_validation_accuracy3":
0.9997575071737461,
  "best_validation_precision": 0.9777423143386841,  "best_validation_recall":
0.9695621728897095,
  "best_validation_fscore": 0.9735202789306641,  "best_validation_loss":
0.05510239019787424,
  "test_accuracy": 0.9954022988505747,
  "test_accuracy3": 0.9996934865900383,
  "test_precision": 0.9858008623123169,
  "test_recall": 0.9810028076171875,
```

```
  "test_fscore": 0.9833571910858154,
  "test_loss": 0.0319001576157281
}
```


pretrained swahili transformer, lr=0.00001, pos tagging


```
  "best_epoch": 7,
  "peak_worker_0_memory_MB": 5449.3203125,
  "peak_gpu_0_memory_MB": 9332.94873046875,
  "training_duration": "1:25:54.586565",
  "epoch": 9,
  "training_accuracy": 0.9972294344879995,
  "training_accuracy3": 0.9999730272987192,
  "training_precision": 0.9764115214347839,
  "training_recall": 0.9566909074783325,
  "training_fscore": 0.9659566879272461,
  "training_loss": 0.007170079944642032,
  "training_worker_0_memory_MB": 5449.3203125,  "training_gpu_0_memory_MB":
9332.94873046875,
  "validation_accuracy": 0.9921189831467486,
  "validation_accuracy3": 0.9996766762316615,
  "validation_precision": 0.9764279127120972,
  "validation_recall": 0.9567330479621887,
  "validation_fscore": 0.9659876227378845,
  "validation_loss": 0.0693623513392432,
  "best_validation_accuracy": 0.991916905791537,  "best_validation_accuracy3":
0.9996362607606192,
  "best_validation_precision": 0.9715831279754639,  "best_validation_recall":
0.9473865032196045,
  "best_validation_fscore": 0.9585748910903931,  "best_validation_loss":
0.05559036874520399,
  "test_accuracy": 0.9962835249042146,
  "test_accuracy3": 0.9997318007662835,
  "test_precision": 0.9764541387557983,
  "test_recall": 0.9567817449569702,
  "test_fscore": 0.9660264253616333,
  "test_loss": 0.032172908316944124
}
```

pretrained swahili transformer, lr=0.0001, crf, pos tagging


```
2022-06-11 22:44:50,984 - INFO - allennlp.common.util - Metrics: {
  "best_epoch": 4,
  "peak_worker_0_memory_MB": 5698.4609375,
  "peak_gpu_0_memory_MB": 9332.60205078125,
  "training_duration": "2:12:03.720062",
  "epoch": 9,
  "training_accuracy": 0.9975718705216521,
  "training_accuracy3": 0.9978339748145334,
  "training_loss": 1.5775271221179867,
  "training_worker_0_memory_MB": 5698.4609375,  "training_gpu_0_memory_MB":
9332.60205078125,
  "validation_accuracy": 0.9919573212625793,
  "validation_accuracy3": 0.9926039687992564,
  "validation_loss": 8.17987885513926,
  "best_validation_accuracy": 0.9915127510811138,  "best_validation_accuracy3":
0.9925231378571717,
  "best_validation_loss": 7.377661480167048,
  "test_accuracy": 0.9957471264367816,
  "test_accuracy3": 0.9962835249042146,
  "test_loss": 3.963569987903942
}
```


pretrained swahili transformer, lr=0.0001, helsinki pos tagging


```
{
  "best_epoch": 5,
  "peak_worker_0_memory_MB": 5583.07421875,
  "peak_gpu_0_memory_MB": 9335.52783203125,
  "training_duration": "1:24:44.563542",
  "epoch": 9,
  "training_accuracy": 0.9965898688553921,
  "training_accuracy3": 0.9999364222373633,
  "training_precision": 0.9467342495918274,
  "training_recall": 0.8755066990852356,
  "training_fscore": 0.8997867107391357,
  "training_loss": 0.01112793169544195,
  "training_worker_0_memory_MB": 5583.07421875,  "training_gpu_0_memory_MB":
```

```
9335.52783203125,
  "validation_accuracy": 0.9915556244189649,
  "validation_accuracy3": 0.999535171986365,
  "validation_precision": 0.9467248916625977,
  "validation_recall": 0.8757268190383911,
  "validation_fscore": 0.8999462723731995,
  "validation_loss": 0.06682281847298145,
  "best_validation_accuracy": 0.9905484970560893,  "best_validation_accuracy3":
0.9994964363185621,
  "best_validation_precision": 0.9082106351852417,  "best_validation_recall":
0.8251690864562988,
  "best_validation_fscore": 0.8477039337158203,  "best_validation_loss":
0.05897250342555344,
  "test_accuracy": 0.9952177563611142,
  "test_accuracy3": 0.9998174716168364,
  "test_precision": 0.9467272162437439,
  "test_recall": 0.8757625222206116,
  "test_fscore": 0.8999730348587036,
  "test_loss": 0.02968891077139415
}
```

Meeting with Sandra 1/6/2022

Explain things more thoroughly, you have as much space as you need. Write the thesis for an intelligent person (2nd year CL student).

Any time you introduce a full sentence in Swahili, provide an interlinear gloss. You can use leipzig glossing or whatever. if you're talking about noun class then include noun class information, otherwise don't.

All dependencies go in figures.

Chapters:

Chapter for how to handle the heterogeneous nature of corpus (high quality and low quality).

Which language model to use (e.g. multilingual language model or swahili transformer I trained).

Can partial dependencies be used (in teh case of rule derived parses) for an additional, very low quality setting?

Low resource dependency parsing in Swahili

Experiment list

I have 174 sentences and 3467 tokens annotated, filtered and checked.

Because of how supar does batching, the number of tokens is actually what's meaningful for splitting up the data into batches. However, note that the batching will not split apart a sentence so if the batch size is set to 8, the result will be one sentence per batch because none of my sentences are less than 8 words in length.

Does augmentation with rule generated data help at all in this scenario?

20 runs total

- Baseline 5 fold cross validation using only the 175 sentences.
- Balanced inclusion of 175, 350, 525, 700 augmented sentences with full trees.

Is curriculum learning useful in a scenario where both quality and difficulty are

considered?

5-fold cross validation between two difficulty functions and three curriculum: 30 total runs.

Difficulty functions:

- Quality is modeled as an offset where some value *p* is added to the difficulty of the sentence. (probably a negative offset because the higher quality sentences should probably come later).
- Quality is modeled as an orthogonal aspect where the standard bucketing is used for length/difficulty clustering but the buckets also encode information about quality

Curriculum:

- homogeneous interleaved batches
- heterogeneous linear introduction
- homogeneous series (one then the other)

Is it better to use rule generated augmented data or augmented data derived from the output of some

parser or a group of parsers?

Implementation log

How does the existing implementation in supar use the buckets of lengths? Does this logic live in the batch sampler?