Swahili in UD

- UD QA session guidelines 2-5-2020
- On the issue of clitic vs morpheme
- Annotation Issues

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