Swahili in the Universal Dependencies Framework

Kenneth Steimel
Indiana University

Outline for Section 1

1. Background
   1.1 Who am I?
   1.2 Overview
   1.3 Swahili
   1.4 The Universal Dependencies Project
2. Swahili in UD
   2.1 Wordhood
   2.2 U-POS tags
   2.3 Morpheme features
   2.4 Syntactic relations

About me

- St. Louisan
- PhD candidate in computational linguistics at Indiana University.
- Interested in research on non-traditional languages in NLP.
- Moved this week.

Larger context

- Part of my thesis research involves creation of a treebank for Swahili in UD with a permissive license.
- This treebank is created by training neural taggers on a large corpus of Swahili (Helsinki Corpus of Swahili).
  - This corpus has a restrictive license: only accessible to CLARIN members, cannot be downloaded in bulk, derivative works must be placed behind the same barriers.
- Those trained taggers then tag the Swahili portion of the OPUS global voices corpus.
- Errors made by taggers are fixed using rules and manual inspection.
- Dependency relations are annotated by hand and augmented using CCG rules.
- This talk concerns how to fit Swahili into the UD project guidelines and precedent.

Why does this matter?

- Strong language biases exist in NLP research.
- 69% of ACL publications in 2016 evaluated on only English `https://sjmielke.com/acl-language-diversity.htm`.
- African languages in general are very under-researched in NLP.
- Often economic motivators are to blame.
- Feedback loop between creation of corpora and NLP research.
- Researching other languages = better understanding of NLP as a whole.

Swahili

- Bantu language spoken in East Africa by 50-100 million.
- National language of Tanzania, Uganda, and the DRC.
- Agglutinative morphology
  - Affix sandwich
    - Prefix - Prefix - Prefix - **Root** - Suffix - Suffix - Suffix
    - ni - na - ku - **pend** - a
- Syntax is fairly consistent, discourse can cause inversions of typical order
- Noun classes and noun class agreement
  - Like grammatical gender systems from German, Spanish, French but with 18 distinctions.
  - Ubiquitous aspect of Swahili grammar.

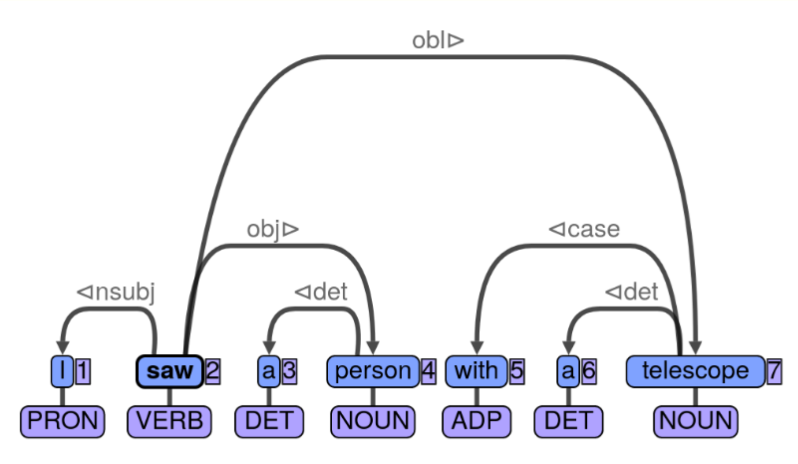
Bantu languages in UD

- Bantu languages are not represented in current UD treebanks.
    - Another researcher, Mariel Aquino, is working on a treebank for Ndengeleko.
- Only Niger-Congo languages are Yoruba, Wolof, and Bambara.
- Yoruba is the closest related to Swahili (Volta-Congo).
    - Yoruba is very different typologically from Swahili.
- Some small considerations were taken to accommodate Bantu languages in UD guidelines.

Universal Dependencies

- Project aiming to provide consistent annotation guidelines for all languages
- Lexicalist dependency grammar
  - Wordhood
  - U-POS tags
  - Morphological features
    - Features describing properties of words
  - Syntactic relations
    - Head -> Dependent directed relations
    - Primacy of content words

Sample English sentence



Meta Universal Dependencies Guidelines

- Conform to official guidelines as much as possible.
- If your language does something the official guidelines have not planned for, try to find precedent in other treebanks.
- Departures from precedent and official guidelines should be noted in documentation for treebank. They may be incorporated in future guideline releases.

Outline for Section 2

1. Background
   - 1.1 Who am I?
   - 1.2 Overview
   - 1.3 Swahili
   - 1.4 The Universal Dependencies Project

2. Swahili in UD
   - 2.1 Wordhood
   - 2.2 U-POS tags
   - 2.3 Morpheme features
   - 2.4 Syntactic relations

What is a word?

- “Dependency relations hold between words” (
  - “Split off clitics…undo contractions”
- Clitics
  - Little pieces of language that are between a separate word and an affix.
  - Is X a clitic? (Zwicky 1977)
    - Can its order relative to other markers change?
    - Do sound changes that apply inside words, fail to apply?
    - If the marker is bound to the root, it is a morpheme.
    - No deletion of marker when coordinated with a similarly inflected component, you have a morpheme. (*dance and singing)
  - Small number of clitics in English
    - *n’t* as in *didn’t*.
  - Agreement markers as clitics in Swahili are debated

Agreement markers in Swahili

- ni - na - ku - **pend** - a
  - ni = I (1st person singular non-negative)
  - na = present tense
  - ku = you (2nd person singular)
  - pend = love
  - a = Indicative (Bantu final vowel)
- Bresnan & Mchombo (1987)
  - Concluded that the subject marker in Chichewa is both a pronoun and agreement marker, Object marker is a pronoun.
- Zwart (1997)
  - Swahili tense marker is an auxiliary verb
  - The subject marker is a clitic that attaches to the auxiliary tense marker.
  - Overt subject (Juma) is topicalized/left dislocated
  - Juma, a.na ku-pend-a
  - Juma, he.is loving.you

Agreement markers in Swahili

- Ud Deen (2006)
  - Nairobi Swahili's subject marker is agreement, not a clitic.
  - Cross linguistic evidence that topics cannot be quantified.
    - Very possible in Nairobi Swahili
  - Subject markers may historically be pronouns that cliticized onto verb and eventually became affixes.

What to do?

- Some dialects of Swahili may require a different analysis from others.
  - Global voices corpus doesn't have information about the dialect of the author/translator.
- Cannot have two subject relations in UD.
- Zwart's analysis of subjects as topics would have to be used.
- For the sake of parallelism with other UD treebanks, verbal markers are prefixes.

U-POS tags

- Largely a matter of conversion from Helsinki tags to U-POS tags.
- Genitive connectors are used frequently in place of compounds in Swahili.
  - Internet Protocol = Itifaki **ya** Wavuti
  - Indiana University = Chuo kikuu **cha** Indiana
  - Should these be adpositions? Is this an English-centric analysis?
- Are infinitive verbs verbs or nouns?
  - Infinitival verbs can have nominal modifiers and verbal arguments
  - **vi**-atu hu-**vy**-o (those shoes)
  - **ku**-tamani hu-**k**-o (that dreaming)
  - **ku**-penda ndizi hu-**k**-o (that love of bananas)
  - *Verbs* because the features needed to describe the infinitives are specific verbal features.

Morpheme features

- UD specifications include Bantu noun classes as features for nouns and features for agreement with other parts of speech.
- If wordhood is established as above, we have agreement with both subject and object (polypersonal agreement).
  - ni - na - ku - **pend** - a
    - ni = I (1st person singular non-negative)
    - na = present tense
    - ku = you (2nd person singular)
    - pend = love
    - a = Indicative (Bantu final vowel)
  - The basque treebank establishes a precedent for this.
    - Used *Number[nom]=3* to indicate number feature of subject.
    - Swahili does not have an overt case system.
    - Likely, both languages should use *Number[subj]* and *Number[obj]*.

Syntactic relations

- In cases of multiple agreement, what is the relationship between the two inflected verbs?
  - Juma **a**-li-kuwa **a**-me-pika chakula (Carstens, 2002)
  - Appears to only be applicable to auxiliaries in examples in Carstens (2002).
  - Currently, I'm ignoring that multiple agreement is happening.
- Copulas
  - "The cop relation should be used for pure copulas that add at most tense-aspect-mood categories to the meaning of the predicate"
  - "Most languages have at most one copula"
  - Swahili has uninflected copulas like *ni* but also inflected copulas like *kuwa*, *kuna* and *hapana*

Conclusion

- A number of analytical decisions must be made prior to annotating Swahili in UD.
- Decisions made can set precedent for future Bantu treebanks in UD.
- Sorting these issues out and creating a treebank of Swahili should foster future NLP research into Swahili and other Bantu languages.

Thank you!

Kenneth Steimel

@ksteimel@scholar.social