Towards a Swahili Universal Dependency Treebank: Leveraging the Annotations of the Helsinki Corpus of Swahili

Kenneth Steimel
Indiana University
ksteimel@iu.edu

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

Dependency annotation can be a laborious process for under-resourced languages. However, in some cases, other resources are available. We investigate whether we can leverage such resources in the case of Swahili: We use the annotations of the Helsinki Corpus of Swahili for creating a Universal Dependency treebank for Swahili. The Helsinki Corpus of Swahili provides word-level annotations for part of speech tags, morphological features, and functional syntactic tags. We train neural taggers for these types of annotations, then use those models to annotate our target corpus, the Swahili portion of the Global Voices Corpus. Based on the word-level annotations, we then manually create constraint grammar rules to annotate the target corpus for Universal Dependencies. In this paper, we describe the process, discuss the annotation decisions we had to make, and we evaluate the approach.

1 Introduction

Swahili is the most-widely spoken Bantu language with an estimated 16 million native speakers (Simons and Fennig, 2018) and 50-100 million L2 speakers. Swahili serves as a national language of Tanzania, Kenya, Uganda, and the DRC, is a working language of the African Union, and serves as lingua franca for the East African Community.

While not a typical under-resourced language in general, there are no syntactic treebanks available for Swahili. Since dependency annotation is a costly endeavor and requires experts in the syntactic framework as well as in the language, we investigate whether we can leverage existing resources to automate the creation of a treebank as much as possible. We investigate an approach where we use the Helsinki Corpus of Swahili (Hurskainen, 2004a) and its annotations as a starting point. We then create automatic taggers for the word-level annotation based on this corpus. After applying these taggers to our target corpus, the Swahili section of the Global Voices Corpus (Tiedemann, 2012), we use a rule-based approach to convert the word-level annotation to Universal Dependency (UD) annotations (de Marneffe et al., 2021). The word-level annotations available in the Helsinki Corpus of Swahili consist of part of speech (POS) tags, based on a POS tagset that has more fine grained information than the Universal Dependency part of speech tagset. Additionally, the corpus contains morphological features and functional syntactic tags, which are based on a constraint grammar framework.

In the process of converting the morphological and syntactic information into the Universal Dependency framework, we encountered challenges based on the fact that little work has been done on UD annotations for Bantu languages. We will discuss some of these cases along with the annotation decisions we have made. The treebank, all rules and code are available at https://git.steimel.info/ksteimel/SWH_UDT. The treebank will be included in the next release of UD.

The remainder of this paper is structured as follows: section 2 reports on related work, section 3 gives an overview of the system used for converting the annotations, and section 4 describes the corpora we use. In section 5, we describe the word-level annotations, and in section 6, we describe the rules for the dependency annotation. Section 7 looks into the quality of our annotations, and section 8 concludes.

2 Related Work: NLP Approaches for Swahili

Political and economic significance promote research on Swahili making it one of the more well-studied Bantu languages in natural language processing. Rule-based systems, data driven approaches, and unsupervised methods have been adopted for Swahili. Hurskainen (1992) presents the first research into morphological analysis of

Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), pages 86–96
May 6, 2023 ©2023 Association for Computational Linguistics

Swahili. The SWATWOL morphological analyzer uses a finite-state two level morphology (Koskenniemi, 1983) to tag words. A constraint grammar system is used to disambiguate when multiple analyses are provided by the finite state-system (Hurskainen, 1996). These components are combined together into the Swahili language manager (SALAMA) (Hurskainen, 2004b). The SALAMA system is extended to include a shallow constraint grammar parser and a deeper grammar-based dependency parser. Though the dependency parser in SALAMA would appear invaluable to the present work, the parser is not freely available and the dependency structure is not provided as part of the Helsinki Corpus of Swahili (Hurskainen, 2004a). Littell et al. (2014) develop a similar finite state morphological analyzer for Swahili. However, their approach involves using an online crowd-sourced dictionary (`kamusi.org`) to ensure wide lexical coverage.

Using corpora developed by rule-based methods, De Pauw et al. (2006) adopt a data driven approach instead. They train part of speech taggers on the Helsinki Corpus of Swahili (Hurskainen, 2004a), which was built using the SALAMA language manager system described above. They use a variety of different off the shelf taggers including a Hidden Markov Model (TnT), SVMs (SVMtool), memory-based taggers (MBT) and a maximum entropy model tagger (MXPOST). The best performance is achieved with the maximum entropy model, though all models reach an accuracy of more than 90%. The maximum entropy model does particularly well with out-of-vocabulary words.

Swahili has also been the focus of research on unsupervised morpheme discovery. Hu et al. (2005) use a string-edit-distance (SED) heuristic to learn the morphology of Swahili. This heuristic goes through all pairs of words in a corpus and creates an alignment between the pairs using edit distance with specific penalties for substitution, addition and deletion. Then, finite-state automata describing pairs of related strings are collapsed together with aligned sequences of characters mapping onto the same state. This process iterates; the finite state automata produced combine with other automata. In the end, the places where the FSA diverges are morpheme boundaries. The authors report high performance when compared to other unsupervised morpheme discovery heuristics such as successor frequency (Harris, 1955).

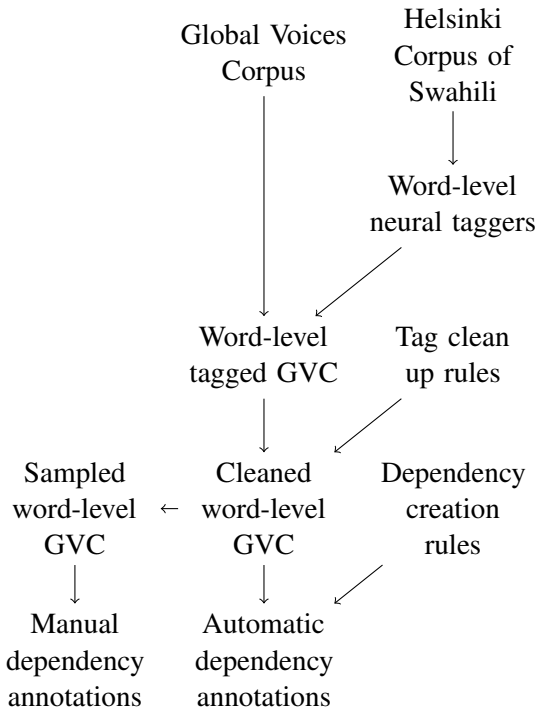

Figure 1: Diagram of the steps involved in creating the Universal Dependency Treebank for Swahili.

Recently, researchers have developed corpora for question answering and emotion classification in Swahili (Martin et al., 2022). In addition, a transformer model for Swahili has been developed (Martin et al., 2022).

3 System Overview

Since no dependency treebank is available, we create the dependency trees automatically, using the annotations in the Helsinki Corpus of Swahili. This corpus contains syntactic information in the form of word-level constraint grammar labels, which need to be converted into Universal Dependency annotations. For this conversion, we use a number of neural tagging models, post-processing rules, and dependency arc creation rules.

The smaller higher-quality, manually annotated, portion of the treebank undergoes a similar series of steps but before dependency creation rules are applied, a portion of the treebank is sampled. This sample is then manually annotated. Figure 1 displays the different steps and resources involved in the process.

word	gloss	POS	syntax	lemma	msd
Afisa	officer	N	@SUBJ	afisa	9/6-SG
Elimu	Education	N	@<P	elimu	9/10-SG
alisimama	stand	V	@FMAINVintr	simama	SUB-PREF=1-SG3 TAM=PAST
kwa	for	PREP	@ADVL	kwa	_
amri	command	N	@<P	amri	9/10-SG
ya	of	GEN-CON	@GCON	ya	9-SG
Mwenyekiti	Chairman	N	@<NH	mwenyekiti	1/2-SG

Table 1: Tags for a single sentence in the HCS.

id	word	lemma	UD POS	HCS POS	Morph. features
1	Afisa	afisa	NOUN	N	NounCl=9|Num=Sing
2	Elimu	elimu	NOUN	N	NounCl=9|Num=Sing
3	alisimama	simama	VERB	V	Mood=Ind|NounCl[subj]=1|Num[subj]=Sing|Pers[subj]=1|Pol=Pos|Subcat=Intr|Tense=Past|Voice=Act
4	kwa	kwa	ADP	PREP	_
5	amri	amri	NOUN	N	NounCl=9|Num=Sing
6	ya	ya	ADP	GEN-CON	NounCl=9|Num=Sing
7	Mwenyekiti	mwenyekiti	NOUN	N	NounCl=1|Num=Sing

Table 2: CONLL-U representation of the sentence in Table 1, focusing on the relevant columns only.

4 Corpora

The corpus situation for Swahili is somewhat difficult. We use both the Helsinki Corpus of Swahili (HCS) (Hurskainen, 2004a) and the Swahili section of the Global Voices corpus (GVC) (Tiedemann, 2012) in this work. Because of licensing limitations, we are unable to annotate the Helsinki Corpus of Swahili with dependencies and redistribute it. For this reason, we leverage the word-level annotations of this corpus to create a POS tagger using the Helsinki Corpus of Swahili POS tagset, a POS tagger for UD POS tags, a morphological analyzer, and a word-level constraint-grammar syntactic tagger. We then use these taggers to annotate the Swahili portion of the Global Voices corpus on the word level. In a final step, we create Universal Dependency annotations based on all of these annotations via a rule-based approach.

The texts contained in the Helsinki Corpus of Swahili are primarily from legislative assemblies, a collection of stories, and news articles. In terms of annotations, the Helsinki Corpus of Swahili is a silver standard corpus created by the SALAMA finite state transducer and disambiguator (Hurskainen, 1999). In the corpus, each word is annotated with its lemma, POS tag, a series of morphological tags, and a functional syntactic tag (constraint grammar) describing the word's role in the sentence are included for each word. Table 1 shows the word-level annotation for the sentence *Afisa Elimu alisimama kwa amri ya Mwenyekiti* (Eng.: The education officer stood for the Chairman's command) from the HCS.

Table 2 shows the conversion of this sentence to CONLL-U format[1]. The HCS uses multiple ways to express the noun class or noun class agreement of a word. For example, nouns (which inherently have a noun class) indicate the noun class of the singular form, the noun class of the plural form and whether the given noun is singular or plural (i.e. 9/10-SG). Noun class agreement prefixes on a verb follow a different convention where noun class, number, and person are indicated (e.g., 1-SG3). Adpositions such as *ya*, meanwhile, indicate noun class agreement as simply a noun class followed by the number (i.e., 9-SG). Additionally, for some POS tags, certain morphological features are assumed as the default by the HCS, and only

[1]For more details on CONLL-U format, see https://universaldependencies.org/format.html

Orthographic representation	Wa-Mauritania	M-Jordan	Ki-China
Morpheme analysis	2-Mauritania	1-Jordan	7-China
English Gloss	Mauritanians	Jordanian	China

Table 3: Tokenization examples of Swahili demonyms.

features which diverge from this default are included. For example, active voice and positive polarity are frequently not explicitly annotated in the HCS annotations, but passive voice and negative polarity are.

The Global Voices Corpus, in contrast, is a large massively multilingual corpus with parallel texts in 46 languages. The GVC consists of news articles from around the world. Because the corpus features articles by citizen journalists, social media text is included as well. Unlike the HCS, the GVC consists of plain text without any word-level annotations. For our treebank, we use the section of the GVC with parallel texts in Swahili and English. This section of the corpus consists of 29 698 Swahili sentences, 546 000 words.

5 Creating Word-Level Annotations

Our ultimate goal is the creation of a Universal Dependency Treebank for Swahili. As a first step, we need to annotate the Global Voices Corpus (GVC) for word-level information, based on which we can then apply the rules that will construct the dependency annotations.

We trained word-level neural tagging models using a sample of 789 691 words, corresponding to 35 925 sentences, from the Helsinki Corpus of Swahili (HCS). A common neural architecture was used for the following word-level tagging models: UD POS tags, language specific POS tags from the HCS, morphological features, and functional syntactic tags. This architecture consists of a two layer bidirectional GRU-LSTM using learned character and word embeddings. Because of the large number of morphological tags, we developed a second architecture that models individual morphological features as single tags. This model is very similar in design to the other neural tagging architecture. However, instead of using argmax to predict a single tag, this model predicts all tags that exceed a threshold (0.5) after a sigmoid activation.

We then automatically tagged the Global Voices Corpus (GVC) with these neural models. From the automatically tagged GVC data, 150 short sentences (length 8–30 words) and 30 longer sentences (length 20–50 words) were randomly sampled for manual annotation. Then, the rule system was applied to the remainder of the GVC to create the UD annotations.

Integrating Swahili into the Universal Dependency framework required us to make annotation decisions regarding tokenization, conversion of part of speech tags, handling of particular constructions in Swahili. We detail these decisions below.

Tokenization When tokenizing, we took special consideration to ensure that demonyms with hyphens were not separated, but compounds were split apart. Demonyms such as those shown in Table 3 are frequently present in the Global Voices corpus. The regular expressions used for tokenization ensured that hyphens after noun class prefixes were not separated from the rest of the word.

Following UD guidelines, compounds of coordinating conjunctions and pronouns are separated into their component tokens. For example, *naye* is separated into *na yeye* (Eng.: and he/she) and *nasi* is separate into *na sisi* (Eng: and you all).

Converting POS Tags to UD The POS tags for the Helsinki Corpus of Swahili are relatively close to Universal Dependency POS tags. In general, the conversion is a many-to-one mapping: the Helsinki corpus tags annotate a number of distinctions that are not featured in Universal Dependencies. Additionally, all POS tags reserved for punctuation in the Helsinki Corpus of Swahili tagset (i.e., COLON, HYPHEN, etc.) are mapped to PUNCT. Table 4 shows the correspondence between the two POS tags annotations.

In some cases, additional information was required to determine the appropriate Universal Dependency tag. For example, in HCS, EXCLAM is typically used for interjections such as *jamani* (Eng.: hey there). However, it is also used for exclamation marks. For those cases, we assign PUNCT.

Both relative pronouns and verbs with relative markers received the Helsinki POS tags REL-LI

Helsinki Corpus	Universal Dependency
A-UNINFL	ADJ
ABBR	SYM
ADJ	ADJ
ADV	ADV
AG-PART	ADP
CC	CCONJ
CONJ	SCONJ
CONJ/CC	CCONJ
DEM	DET
EXCLAM	INTJ
GEN-CON	ADP
GEN-CON-KWA	ADP
INTERROG	PRON
N	NOUN
NUM	NUM
NUM-ROM	NUM
POSS-PROM	SCONJ
PREP	ADP
PREP/ADV	ADV
PRON	PRON
PROPNAME	PROPN
REL-LI	PRON
REL-LI-VYO	PRON
REL-SI	PRON
REL-SI-VYO	PRON
TITLE	PROPN
V	VERB
V-BE	AUX
V-DEF	VERB

Table 4: Correspondence between Helsinki Corpus of Swahili and Universal Dependencies POS tags.

or REL-LI-VYO. In such cases, we use the functional syntactic tag assigned to the word for disambiguation: if the word has a verbal functional tag, then we assign the UD tag V. In all other cases, we assign PRON.

Morphological Features Unlike the conversion of POS tags, which is a relatively straightforward process, morphological tag extraction is considerably more complex.

Morphological features in Universal Dependency consist of attribute-value pairs. These are represented in the CONLL-U format with attributes separated from their associated values by '=' and each pair is separated by '|'. The Universal Dependency annotation guidelines (Nivre et al., 2017) provide morphological features to accommodate noun classes in Bantu languages. These features are indicated using different values for the NounClass attribute. However, these features are not sufficient since verbs can have multiple noun class markers indicating the noun class of the subject, object, and relative head.

For example, the word *aliyoitumia* has markers indicating noun class, person, and number for the subject, object, and relative head, see the analysis in example (1).

(1) *a-li-yo-i-tum-ia*
3SG-PST-4.REL-9.OBJ-use-APPL

those that he/she used for it

To address this issue, we use layered morphological features. For `Person`, `Number`, and `NounClass`, additional subtypes such as `rel`, `subj`, and `obj` are added. Until recently, this was not a documented possibility in Universal Dependencies, however other treebanks have established these layered features as a precedent. One example is the Basque Universal Dependency treebank (Aranzabe et al., 2015). Though subtypes are not described in the documentation of the conversion process (Aranzabe et al., 2015), this treebank includes verbs with multiple number features using subtypes to indicate the type. Unlike the Basque Universal Dependency treebank, we do not use subtypes with the *case* of the agreeing element indicated for Swahili. Rather, the subtypes simply specify the function of the agreeing element. For example, where the Basque corpus uses `Number[nom]`, the Swahili corpus uses `Number[subj]`. The Basque option is not usable for Swahili as Swahili does not have overt case, and adopting a covert case analysis for all languages like Swahili does not follow the principles of Universal Dependency.

6 Creating Dependency Annotations

6.1 Dependency Guidelines

We adhere to the guidelines laid out by Universal Dependency (Nivre et al., 2017). This section outlines some specifics for how we applied these guidelines to Swahili. The Universal Dependency guidelines state that "[t]he copula be is not treated as the head of a clause, but rather the nonverbal predicate" (de Marneffe et al., 2020). The guidelines also advise that "[t]he cop relation should only be used for pure copulas that add at most TAME categories to the meaning of the predicate".

In the Swahili treebank, we make a distinction between the "verbal" copula *kuwa* and other copulas like *ni*, the negated form *si*, emphatic forms like *ndiyo* and locative copulas like *uko*[2]. We analyze *kuwa* as a verb while the other copulas are given the POS tag COP and are not considered the head of their clause.

6.2 Rule application

To create rules for correcting common issues with the neural taggers and generate dependency arcs, CG3 was used (Bick and Didriksen, 2015). CG3 is an extended variant of constraint grammar with implementations for compiling and applying constraint grammar rules, allowing us to develop and apply complex rules.

Addressing errors in word-level tags Initially, rules were written to remedy errors with the word-level tags produced by the neural POS models. To correct errors produced by the automatic tagger, `SUBSTITUTE` and `ADD` commands were used to change one tag to another, add a missing tag, or remove an errant tag. These word-level tag correction rules were applied before all rules creating dependency arcs.

In many cases, tag rewrite rules were leveraged to rewrite a word-level tag if three or more of the other word-level tags indicated that an error had occurred. The first example in Table 5 displays a rule that replaces `NOUN` with `VERB` in cases where other taggers indicate that the word in question is actually a verb. More specifically, the language specific POS tagger must assign this word a `V` tag, and it must have the morphological feature specifying polarity and one of a number of functional syntactic tags indicating that this word is serving as a verb in the sentence for the rule to apply.

Dependency arc creation To create dependency arcs from sentences plus the word-level annotations, we created ordered regular expression rules, such that highly specific rules were followed by more generic versions using more lax restrictions; and occasionally we used fallback rules. Some phenomena were addressed using a single generic rule without more specific or lax versions. For example, the bold phrase in the sentence in example (2) shows an example of multiple noun phrases linked together in an associative noun phrase chain.

(2) *Biti ni katibu mkuu w-a MDC*
Biti COP secretary major ASSOC.1 MDC
, *ki-na-cho-ongoz-wa* ***na***
, 7.SUBJ-PRES-7.REL-lead with/by
Wa-ziri Mkuu w-a zamani
minister major ASSOC.1 past
w-a nchi hiyo , *Morgan*
ASSOC.1 country DEM , Morgan
Tsvangirai .
Tsvangirai .

Biti is the secretary general for Movement for Democratic Change, led by former Prime Minister Morgan Tsvangirai.

Without taking noun class agreement into consideration, the noun phrase in bold is ambiguous, i.e., both dependency analyses in Figure 2 would be possible.

As the interlinear analysis in example (2) indicates, *zamani* is a class 9 noun, the associative adposition *wa* is class 1, and *Waziri* is class 1. The associative agrees in noun class with the noun that its parent noun modifies. Thus the syntactic analysis in the dependency analysis on the right of Figure 2 is the correct one. The specific version of the rule, shown as the second rule in Table 5, leverages this agreement.

More generic rules are applied if no match for a specific rule is found. A generic rule may apply because of errors in the morphological tags produced by the tagging model or because of missing agreement between the two tokens[3]. In this particular case, associative adpositions in different noun classes are often polysemous. For example, *wa* can indicate that the noun modified by its parent noun is class 1, as in the example above, or classes 2, 3, or 11. The neural taggers can leverage the surrounding context; however errors still occur with some frequency. A generic rule is thus used to combat errors in the morphological annotations. This rule, shown as the third rule in Table 5, has no agreement constraints.

Fallback rules are applied in cases where criteria using morphological and functional tags both

[2]Locative copulas like *uko* could perhaps be annotated as an adverbial. However, this only affects the label. The non-verbal predicate following *uko* would still be the head of the clause.

[3]For example, some adpositions like *wa* agree with the noun class of the preceding noun and would match a specific rule form. However, other adpositions like *katika* do not indicate the noun class of the preceding noun in any way.

```
SUBSTITUTE NOUN VERB TARGET V (0 (/FMAINV/r) LINK 0 (/Polarity=/r) );
```

```
# Go from a nominal to another nominal with a genative connector in between,
# the first nominal and the genative connector have to agree in noun class
ADDRELATION nmod EXTENDED_NOMINAL TO
  (0 $$NOUN_CLASS
    LINK 1* GCON BARRIER mainv
    LINK 0 $$NOUN_CLASS
    LINK 1 EXTENDED_NOMINAL) ;
```

```
ADDRELATION REVERSE nmod EXTENDED_NOMINAL (T:no_parent) TO
  (-1 ADP
    LINK -1* EXTENDED_NOMINAL BARRIER mainv);
```

Table 5: Example rule for error correction in word-level tags.

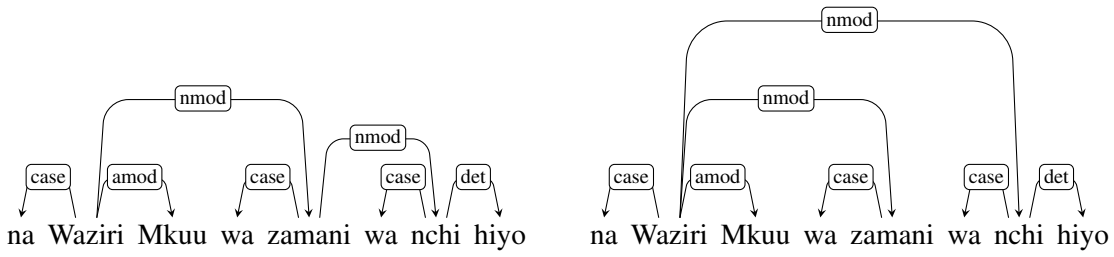

Figure 2: Two possible dependency analyses for the noun phrase chain from example (2).

fail. Instead, simple linear ordering patterns conditioned on POS are used in fallback rules.

Table 6 shows how often specific, generic, and fallback rule types are used to address each type of phenomenon. While the highly specific rules are able to disambiguate between different possible arcs more effectively, these specific rules are triggered less frequently.

7 Quality of Annotations

7.1 Word-Level Tagging Results

Before we trained the neural tagging models, we set aside 2 000 sentences from the Helsinki Corpus of Swahili, 1 000 sentences for validation and 1 000 sentences for testing the model's performance. Table 7 shows the results of this evaluation. All models exceed 94% accuracy, and all but the multilabel morphological tagger reach an accuracy of 97% and higher; corresponding F-scores for POS tagging, and functional role are in the same range. As discussed in section 5, the multilabel tagger outputs a tag if it exceeds a threshold of 0.5. While this allows the tagger to be more flexible and consider combinations of morphological features that were not present in the training data, it also does not enforce co-occurrence restrictions between morphemes. The multilabel model can predict both NounClass=9 and NounClass=7 for the same noun, though this should not be permitted. These restrictions are automatically followed when using the monolithic morphological tagger.

Note that the test data are derived from the source corpus, the Helsinki Corpus of Swahili, and therefore the exact performance of the models on the target corpus (GVC) cannot be determined. During the manual corpus creation process, we corrected UD POS tags and added depedency arcs but did not correct other word level tags. In the future, other word level tags will also be corrected.

7.2 Strengths and Limitations in Generated Trees

Out of 29 698 sentences, the dependency creation rules rules generate spanning trees for 4 994 sentences. 3 499 of these trees are projective dependency trees. When examining particular linguistic phenomena, the rules do well at some and

Phenomenon	Rule type	Number of rule applications	Percentage of rule applications
amod	Specific	487	37.12%
	Generic	825	62.88%
case	Specific	2438	25.06%
	Generic	7292	74.94%
det	Specific	1030	91.15%
	Generic	100	8.85%
det	Specific	292	82.95%
	Generic	60	17.05%
case	Generic	140	10.26%
	Fallback	1224	89.74%
nummod	Specific	247	18.14%
	Generic	1115	81.86%
nmod	Specific	8152	72.28%
	Generic	3126	27.72%
nsubj	Specific	155	0.36%
	Generic	13237	30.57%
	Fallback	29903	69.07%
obj	Specific	535	1.36%
	Generic	19355	49.34%
	Fallback	19336	49.29%

Table 6: Rule type frequency for dependency creation rules on Global Voices Corpus.

Task	Number of tags	Accuracy	Macro-average F-score	Weighted-average F-score
Functional role tagging	39	97.53	92.01	97.51
Helsinki POS tagging	55	98.99	93.13	98.96
UD POS tagging	15	98.98	97.81	98.98
Multilabel morphological tagging	99	94.74	42.35	95.04
Monolithic morphological tagging	7 397	97.63	72.87	97.52

Table 7: Model accuracy in relation to tagset size for bidirectional GRU-based models.

produce incorrect or incomplete trees when confronted with others.

Associative chains In Swahili, complex noun phrases are often constructed using chains of associative adpositions. Our system of rules does well with these constructions. Figure 3 shows the tree generated for the associative chain at the beginning of the sentence in example (3). While the associative chain is handled correctly, the locative noun phrase *nchini Cambodia* should be modifying *Watumiaji* (users), at the beginning of the text. However, there are no morphological features that can help the rules disambiguate the attachment of this locative noun phrase.

(3) *Wa-tumiaji* *w-a* *mitandao* *y-a*
1-users 1-ASSOC network 9-ASSOC
jamii *nchini* *Cambodia* *pia*
society country Cambodia also
wa-me-hamas-ish-wa *ku-weka*
3PL-PERF-motivate-CAUS-PASS 15-set
picha *z-a* *alama* *y-a*
pictures 10-ASSOC sign 9-ASSOC
kampeni
campaign

Social media users in Cambodia are also encouraged to replace their profile photos with icons of the campaign.

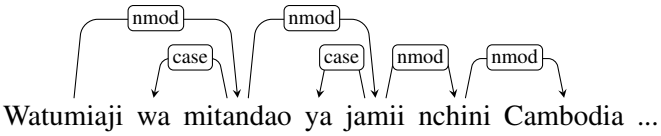

Figure 3: Initial associate chain in example 3

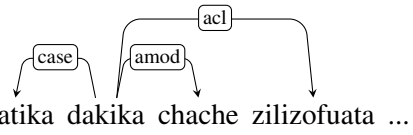

Figure 4: A tree for a short headless relative clause

Relative clauses Both headless and headed relative clauses are often handled correctly by the system of rules. Only headless relative clauses are shown here due to space constraints. Example (4) displays a phrase with a headless relative clause. The noun *dakika* (Eng.: minutes) is modified by the relative clause *zilizofuata* (Eng.: following). The intervening adjective *chache* does not interfere with the rules and is also connected to *dakika*, as appropriate.

(4) *Ikiwa* *i-ta-tok-ea* ,
if 9-FUT-come.from-PASS ,
basi *siku* *z-a* *ahadi*
then day 10-ASSOC vow
z-a *ndoa* *ku-pewa*
10-ASSOC marriage 15-give
u-muhimu *u-na-o-stahili*
14-importance 14-PRES-14.REL-deserve
“ *Mpaka* *ki-fo*
“ until 7-death
ki-ta-ka-po-tu-tengan-isha
7-FUT-CONT-16-1PL-be.separated-CAUSE
” *zi-me-pita* *?*
” 10-PERF-pass ?

If this will happen, gone are the days when the marriage vows are to be taken seriously “Til death do us part”?

Root identification Identifying the root of the clause using rules is difficult. This is particularly true in cases where topicalization of some kind has occurred.

In example (4), the fronted adverbial clause *Ikiwa itatokea* (Eng.: If this will happen) interferes with the current rules for root identification. Instead of assigning *kupewa* root status, the verb in the adverbial clause *itatokea* is erroneously labeled as the root. While *itatokea* is a verbal form and thus a possible root, it is the head of the subordinate clause.

8 Conclusion

Our work is concerned with creating a Universal Dependency treebank for Swahili leveraging the annotations in the Helsinki Corpus of Swahili. We show that we can train neural taggers to annotate UD POS tags, language specific POS tags, morphological tags, as well as functional syntactic tags, and that we can use those annotations on a new corpus to create regular expression rules to derive a dependency annotation. The results are not perfect: for about 30% of the sentences, our methods do not create fully connected trees. But our annotations can be improved in the future. It is, of course, possible to add more rules to our framework to cover more cases. However, any new rule will be very specific and will thus only improve a very small number of cases. We consequently argue that additional improvement should come from training a robust parser and manually correcting the parse trees. We are planning to investigate robust parsing methods that will provide reliable parses when trained on the available annotations.

9 Limitations

It is certain that there are errors in the automatic and manual annotations. Our conversion procedure is limited by the information available in the Helsinki Corpus of Swahili. And while the first author, who manually annotated the portion of the Swahili treebank, has extensive training in Bantu syntax, he is not a native speaker of Swahili. We hope that this initial step inspires future expansion and/or correction of the corpus.

10 Ethics Statement

Working on an under-resourced language is always accompanied by the danger of disenfranchising the language community. However, depen-

dency annotations require a syntactic background, and there are often not enough speakers of the language with such a training. We hope that the Swahili community will adopt and improve our treebank. We have consciously chosen a corpus the can be freely distributed, rather than working with the Helsinki Corpus of Swahili, which comes with restrictive licensing requirements.

References

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2015. Automatic conversion of the Basque Dependency Treebank to Universal Dependencies. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 233–241.

Eckhard Bick and Tino Didriksen. 2015. CG-3 - Beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 305–308.

Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Joakim Nivre, Slav Petrov, Sampo Pyysalo, Sebastian Schuster, Natalia Silveira, Reut Tsarfaty, Francis Tyers, and Dan Zeman. 2020. Universal Dependencies.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Guy De Pauw, Gilles-Maurice De Schryver, and Peter W Wagacha. 2006. Data-driven part-of-speech tagging of Kiswahili. In *International Conference on Text, Speech and Dialogue*, pages 197–204. Springer.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Yu Hu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005. Refining the SED heuristic for morpheme discovery: Another look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 28–35, Ann Arbor, Michigan. Association for Computational Linguistics.

Arvi Hurskainen. 1992. A two-level computer formalism for the analysis of Bantu morphology. an application to Swahili. *Nordic Journal of African Studies*, 1(1).

Arvi Hurskainen. 1996. Disambiguation of morphological analysis in Bantu languages. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 568–573.

Arvi Hurskainen. 1999. Salama: Swahili language manager. *Nordic Journal of African Studies*, 8:139–157.

Arvi Hurskainen. 2004a. Helsinki corpus of Swahili. *Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC*.

Arvi Hurskainen. 2004b. Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, 13(3):363–397.

Kimmo Koskenniemi. 1983. Two-level model for morphological analysis. In *IJCAI*, volume 83, pages 683–685.

Patrick Littell, Kaitlyn Price, and Lori Levin. 2014. Morphological parsing of Swahili using crowdsourced lexical resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3333–3339, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jiyoung Woo. 2022. SwahBERT: Language model of Swahili. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, Seattle, United States. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina,

Kaili Müürisep, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Gary F Simons and Charles D Fennig. 2018. Ethnologue: Languages of the world, twenty. *Dallas: SIL International. Retrieved from www. ethnologue. com. Accessed*, page 2018.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).