Yorùbá Dependency Treebank (YTB)

Ọlájídé Ishola, Daniel Zeman
Charles University, Faculty of Mathematics and Physics, ÚFAL
Malostranské náměstí 25, CZ-11800 Praha
lajish@tasmag.com, zeman@ufal.mff.cuni.cz

Abstract

Low-resource languages present enormous NLP opportunities as well as varying degrees of difficulties. The newly released treebank of hand-annotated parts of the Yorùbá Bible provides an avenue for dependency analysis of the Yorùbá language; the application of a new grammar formalism to the language. In this paper, we discuss our choice of Universal Dependencies (UD), important annotation decisions, and results of our parsing experiments. We also highlight future directions for a rapid expansion of the treebank.

Keywords: Universal Dependencies, Yorùbá, annotation, treebank, grammar, parsing

1. Introduction

Yorùbá is one of the most spoken indigenous African languages with estimated 40 million speakers,[1] living mostly in Nigeria, Benin, Togo, and across the Atlantic in Brazil, Cuba, Jamaica and Trinidad. However, there is a distinction between the African and diaspora Yorùbá.

Yorùbá is classified as a Niger-Congo language which shares close relations with Itsekiri and Igala (Akinlabí and Adéníyì, 2017). It is one of the three major Nigerian languages recognised in the National Policy on Education (NPE) document published in 1977 to be used as a medium of instruction for the first 3 years of a child's primary school education. Yorùbá is also offered as a course of study in some Nigerian tertiary institutions. Yorùbá also has Braille notations; a part of the Nigerian Braille writing system for the visually impaired, coordinated by Braille Advancement Association of Nigeria (BRAAN) (UNESCO, 2013).

The earliest work on Yorùbá appeared in 1819 century when the German linguist Bowdich published a vocabulary primer containing the numerals 1-10 (Ogunbiyi, 2003). There were several individual and collaborative efforts thereafter to arrive at the contemporary Yorùbá orthography and scholarship. Today, Yorùbá is one of the most documented West African languages (Akinlabí and Adéníyì, 2017).

Despite the extensive language resources available in the mass media, film industry, books and rich undocumented oral literature, they remain untapped for open-source annotated data. Hence, Yorùbá is only low-resource from a technological point of view due to lack of readily-available corpora for computational analysis. We describe our work on the creation of the first Yorùbá Treebank (YTB) using the Universal Dependencies (UD) framework on data sourced from the Yorùbá Bible.

[1] `https://www.ethnologue.com/language/yor`

2. Universal Dependencies

Despite vast dissimilarities in the languages of the world, there are attempts at formalisms that can be applied to all languages. However, minority languages, especially African languages, do not receive enough attention due to lack of annotated data. Beaming search light on "low-resource" languages will help us make better and more universal grammar rules.

The Universal Dependencies (UD) project (Nivre et al., 2020) was started to provide a universal inventory of categories/tagsets (allowing language-specific extensions where necessary) and guidelines for consistent annotation across languages of the world by providing a transparent and accessible framework for experts and non-specialists alike. The annotation scheme for representing dependency structure is based on Stanford Dependencies (SD) (De Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2011) and the Interset interlingua for morphosyntactic features (Zeman, 2008). The UD initiative harmonised these projects into a single coherent framework.

UD is based on dependency relations that exist between lexical units in a construction. Words are connected by directed relations known as dependencies, the word at the start of the relation is called head (parent), and the word at the end of the relation is called dependent (child). The head-dependent approach can be traced back to (Tesnière, 1959). Unlike some other dependency grammars, in UD the heads are normally content words, while function words and punctuation symbols are normally leaves (dependents that do not have children of their own).

UD is a great opportunity for minority languages like Yorùbá to build publicly available annotated corpora usable in a wide range of NLP applications. UD provides universal tools and guidelines, which makes it easier to start an annotation project in a new language. Morphological properties of words and their dependency relations in UD are encoded in the standard file format called CoNLL-U.[2] Within that format, YTB

[2] See `https://universaldependencies.org/format.`

utilises the MISC (miscellaneous) column to provide English glosses.

3. Text Source

The Yorùbá Bible is our primary source of data. This choice is opportunistic: the Bible does not pose copyright issues, and it is a massively parallel text, allowing for cross-lingual transfer techniques. (Oluokun, 2018) provided a preliminary automatic annotation of the data through projection from Bible text in other languages, annotated according to the UD guidelines. A subset of 100 sentences was manually checked and released. We re-checked and corrected this initial dataset, and doubled the size of the corpus by adding new annotated sentences. These were first processed using a parser (Straka et al., 2016) trained on the first 100 sentences, then a fair amount of post-editing was performed using the tree editor TrEd (Pajas and Fabian, 2000). Well-formedness of the data is checked with the Python validation script maintained by the Universal Dependencies Consortium.[3]
To facilitate cross-lingual analysis, we focus on those Bible chapters that are also available in other languages in UD, viz. Ancient Greek, Latin, Gothic and Old Church Slavonic.

4. Yorùbá-specific Decisions

A critical part of any corpus annotation project, regardless of its type and scale, is the annotation scheme (Lu, 2014). No Yorùbá-specific guidelines were available at the time when (Oluokun, 2018) did her work on annotation projection, and her annotation decisions are not documented. (Ishola, 2019) drafted the first annotation guidelines[4] for the Yorùbá Dependency Treebank (YTB) within the frame of the general UD annotation guidelines. The guidelines were drawn as a result of the issues encountered during new manual check of the sentences originally released by (Oluokun, 2018), and during manual annotation of new text.
We now proceed to summarize interesting Yorùbá-specific issues in this section.

Yorùbá has a strict subject-verb-object (SVO) word order. The subject position is filled by a noun, a nominal phrase, or a subject clause while the predicate position is filled by a verb or a series of verbs; verb serialization is prominent in Yorùbá. Noun phrases, verb phrases and prepositional phrases in the language are head-initial (Àjàní, 2001). The example below shows a simple sentence structure.

(1) Jídé jẹ ìrẹsì
Jide eat rice
'Jide ate rice'

`html` for the specification of the format.

[3] `https://universaldependencies.org/release_checklist.html#validation`

[4] `https://universaldependencies.org/yo/index.html`

4.1. Tokenization

Tokenization can be language-dependent and character-set dependent. There are 25 letters in the Yorùbá alphabet comprising consonants, vowels (some of them with diacritics) and a digraph. Moreover, there are three tone levels in Yorùbá: high (´), mid (-) and low (`); the mid tone is not explicitly marked by diacritics in the orthography. Only vowels and syllabic nasals ***n*** and ***m*** are tone-marked. Some Yorùbá letters must be encoded as strings of several characters, viz. the base letter and a combining character for the tone mark. Consequently, not all software tools can render Yorùbá text properly.
Hyphenated words are left as one token if the parts cannot be correctly annotated individually. For example, lengthened nasals are common in Yorùbá and splitting these words would result in word classification problems.

(2) **níhìn-ín** 'here'
kìn-ín-ní 'first'
karùn-ún 'fifth'

Having said that, some words derived through morphological processes required splitting to be assigned POS tags. Their mapping to the original texts is preserved in accordance with the UD principles; the resultant words are assigned their rightful categories:

(3) **lórúkọ** 'in name' → **ní** 'in' + **orúkọ** 'name'
lókúta 'with stone' → **ní** 'with' + **òkúta** 'stone'
gbàágbọ́ 'believe' → **gbà** 'accept' + **á** 'him' + **gbọ́** 'hear'

4.2. Part-of-speech Annotation

(Yusuf, 2010) posits four sustainable parts of speech for Yorùbá, more particularly for teaching purposes. They are: Noun, Verb, Preposition and Conjunction. Any other 'suspected' category could find sufficient resemblance in these four to be called a type of one or the other. Nevertheless, we have examined other categories defined in UD for their relevance in Yorùbá, providing for more fine-grained distinctions and increasing parallelism with other languages annotated in UD. Detailed information about the other parts of speech are discussed in (Ishola, 2019). YTB uses 15 of the 17 universal parts-of-speech tags, **SYM** and **INTJ** are not used in the corpus at present.

ADP: adposition – a roof term for prepositions and postpositions, however, Yorùbá only has prepositions. There is no morphological case in the language; instead, prepositions are used as case markers and specify the role or direction of a noun in a phrase. (Adékẹ́yè, 2016) argues that preposition is not a lexical class in the standard Yorùbá but it is "part of the functional support for the noun in the language." Nevertheless, the words that we tag ADP invariably function as prepositions in our data.

POS Tag	Description	Frequency
ADJ	Adjective	57
ADP	Adposition	175
ADV	Adverb	89
AUX	Auxiliary	132
CCONJ	Coordinating Conjunction	147
DET	Determiner	72
NOUN	Noun	360
NUM	Number	10
PART	Particle	83
PRON	Pronoun	486
PROPN	Proper Noun	72
PUNCT	Punctuation	449
SCONJ	Subordinating Conjunction	138
VERB	Verb	399
X	Other	2

Table 1: Universal part-of-speech tag statistics in YTB.

(4) Jẹ́ kí omi **abẹ́** ọ̀run wọ́
let be water under heaven gather
VERB PART NOUN ADP NOUN VERB
papọ̀ **sí** ojúkan
together to place
ADV ADP NOUN
'let the waters under the heaven be gathered together unto one place'

Conjunction: Compound sentences are joined with the aid of phrasal conjunctions such as: **àti** 'and **tàbí/àbí** 'or' and sentential conjunctions: **sugbón**: 'but', **yálà...tàbí**: 'either...or' and **sì** 'and'/'then' (Yusuf, 2010). A complex sentence is also possible by embedding a sentence under another; this involves the use of the keyword **'pé'** (Adesola, 2005). UD annotation gives room to make a distinction between two conjunction types:

CCONJ: coordinating conjunction – these are words that link constituents in a construction together without syntactically subordinating one to the other.

(5) Ọ̀kan ni èmi **àti** Baba mi
one FOCUS I and father me
NUM PART PRON CCONJ NOUN PRON
'I and my Father are one.'

SCONJ: subordinating conjunction – these are conjunctions that mark subordinating relations between clauses or compound constructions.

(6) Jésù wí fún un **pé**,
Jesus say for her that
PROPN VERB ADP PRON SCONJ
"Arákùnrin rẹ yóò jíǹde."
brother her shall rise
NOUN PRON AUX VERB
'Jesus saith unto her, Thy brother shall rise again.'

See Figure 1 for the dependency tree of example (6).

NOUN: noun – Yorùbá nouns are easy to identify, (Yusuf, 2010) puts forward that a good guide to the identification of a noun in the language is generally a vowel-initial word and consistently of more than one syllable. Nouns give information about people, places and things and they can be in the subject or object position. The **PROPN** tag is reserved for proper names in UD.

(7) **Òwe** yìí ni **Jésù** pa fún
parable this FOCUS Jesus tell for
NOUN DET PART PROPN VERB ADP
wọn
them
PRON
'This parable spake Jesus unto them'

VERB: verb – Yorùbá has constraints prohibiting deletions which are recoverable, such as deletion of subjects and of verbs; in essence, a sentence can not be verbless (Lawal, 1987). Yorùbá verbs are mostly monosyllabic; bisyllabic verbs are a closed class.

(8) Ẹnìkan kò **gbà** á lọ́wọ́ mi
nobody not take it from me
PRON PART VERB PRON ADP PRON
'No man taketh it from me'

See Figure 2 for the dependency tree of example (8).

AUX: auxiliary – there are various forms of auxiliaries in Yorùbá. For instance, **ti** 'have' usually occurs before the main verb in a perfective construction.

(9) Ìwọ ó **ti** mú wa ṣe
you he have take us do
PRON PRON AUX VERB PRON VERB
iyèméjì pẹ́ tó
doubt long how
VERB ADV ADV
'how long dost thou make us to doubt?'

Modal auxiliary: yóò 'shall' marks the future tense and it is also followed by a main verb.

(10) Arákùnrin rẹ **yóò** jíǹde.
brother her will resurrect
NOUN PRON AUX VERB
'Thy brother shall rise again.'

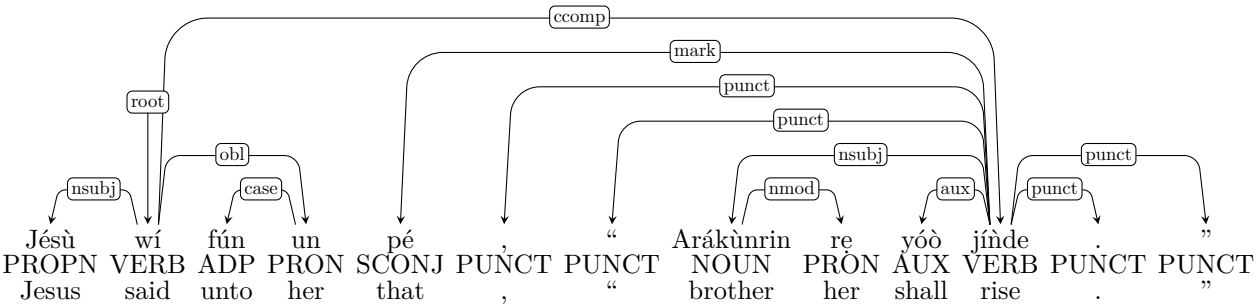

Figure 1: **SCONJ** example for (6)

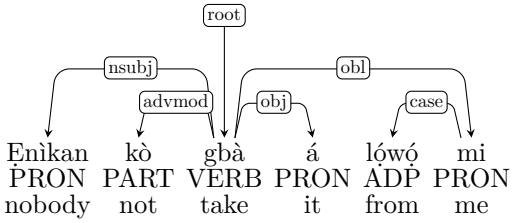

Figure 2: **VERB** example for (8)

Also when **kí** follows **jẹ́** (‘be’), jẹ́ is tagged VERB while kí is tagged PART instead of AUX. **Jẹ́ kí** (‘let’) is a multi-word expression and both words are connected with the dependency relation *compound:prt*; tagging **kí** as AUX would contravene the provision that an auxiliary should receive a corresponding *aux* relation.

(11) Ọlọ́run sì wí pé, “**jẹ́ kí**
god then say that, “let let
NOUN CCONJ VERB SCONJ VERB PART
ìmọ́lẹ̀ kí ó wà,” ìmọ́lẹ̀ sì
light let it be,” light then
NOUN AUX PRON VERB NOUN CCONJ
wà.
be
VERB
‘and God said, “let there be light”: and there was light.’

However, there are instances of **kí** without **jẹ́**, and these are less problematic.

(12) **Kí** wọn ó jẹ́ ìmọ́lẹ̀ ní ojú
let them will be light in eye
AUX PRON AUX AUX NOUN ADP NOUN
ọ̀run
heaven
NOUN
‘And let them be for lights in the firmament of the heaven’

When **ó** (‘will/should’) is preceded by **wọn** (‘they’) like in the example above, it is tagged AUX as it is neither coreferential nor for reinforcing the third person plural **wọn**; it is a future tense marker in this instance.

The imperfective marker **ń** is also tagged AUX in the corpus.

(13) Jésù sì **ń** rìn ní tẹ́mpílì,
Jesus then is walk in temple
PROPN CCONJ AUX VERB ADP NOUN
ní ìloro Sólómónì
in porch Solomon
ADP NOUN PROPN
‘And Jesus walked in the temple in Solomon’s porch.’

See Figure 3 for the dependency tree of example (13).

4.3. Other Considerations

Polysemy and homonymy are important phenomena in Yorùbá. To correctly categorize a word, the context where it is used is the determining factor for disambiguation. Tone can distinguish meaning:

(14) **Ẹ̀gbẹ̀** – ‘name of town in Kogi state’
ẹgbẹ́ – ‘mate, colleague, association’
ẹ̀gbẹ́ – ‘side’
ẹ̀gbẹ – ‘dried’

Nevertheless, even a word with the same tone can mean different things based on the position and context, e.g. **bí** can be ‘procreate’, ‘if’, ‘not’.
The negation marker **kò** is tagged PART, the frequently-occurring focus-marker **ni** is also tagged as a particle as there is no dedicated tag for focus markers in UD.
Determiners can specify the grammatical plural of nouns while words indicating quantity follow the words they quantify. A word like **àwọn** (‘the’/‘them’) can be tagged DET (pluralizer) or PRON, consequently, the actual function has to be determined based on context.
Pronouns do not distinguish gender: for instance, **ó** can be ‘he’, ‘she’ or ‘it’.
Adjectives specify the attribute of nouns and they only have the positive degree form, comparison is done with the introduction of **jù lọ, jù...lọ**.
Verb serialization is prominent in Yorùbá, the first verb marks the tense while the second indicates the direction of an action (see section 4.4.).

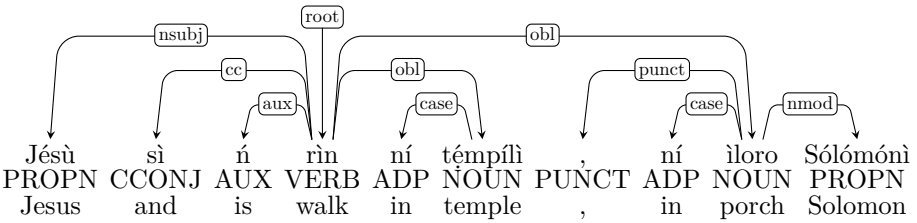

Figure 3: **AUX** example for (13)

If a verb ends with a nasal vowel, the following object must also be written nasalized. Examples below:

(15) **Mo fún un.** 'I gave him/her.'
A rán an. 'We sent him/her.'

Finally, because of the classical diction used in the Yorùbá Bible, there are many redundant words used for literary effects; they have little to no contribution to the meaning of a sentence, classifying them can be tricky. This is similar to what (Roorda, 2017) observed in the annotation of the Hebrew Bible.

4.3.1. Morphological Features

Yorùbá is an isolating language with virtually no inflection; therefore we only use features that serve to finer partitioning of word categories. **'NumType'**, **'PronType'** and **'Typo'** are the only features added to the treebank presently. ***NumType*** indicates whether a word is cardinal or ordinal number; ***PronType*** encodes the type of pronominal forms (personal, demonstrative, interrogative etc.); ***Typo*** is used to mark misspelled words. We do not correct spelling at the word form level so that statistical models (including parsers) can be trained that will be robust enough and applicable to unnormalized text. However, we do normalization in word lemmas. Examples of features from the corpus are given below:

	méjì 'two'	→	`NumType=Card`
	ó 'he'	→	`Case=Nom|Number=Sing|Person=3|PronType=Prs`
(16)	**Bétanì** (correct: **Bẹ́tẹ́nì**) 'Bethany'	→	`Typo=Yes`

4.4. Dependency Annotation

Here we discuss the second level of annotation inherent in YTB; these are the dependency relations that exist between lexical units in a sentence. As stated earlier, Yorùbá has an SVO word order structure where nominal subjects are in initial position followed by adjectives, demonstratives and relative clauses. YTB currently utilizes 29 of the 37 UD dependency relations shown in Table 2; the most crucial ones are described below.

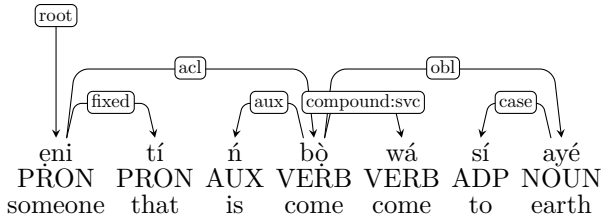

Figure 4: **compound** example for (18)

root: root – the root relation points to the root of the sentence. The main verb is normally the root. If it is not present (for instance, if the annotated segment is not a full sentence), one of the orphaned dependents is promoted to be the root. Auxiliaries and copulas are not treated as full verbs in UD.

compound: compound – there is a non-exhaustive list of **multi-word** compounds in Yorùbá in (Bamgbọṣe, 1983; Yoruba Cross-Border Language Commission, 2017). These words are connected with the **compound** tag. Each word of the multi-word expression is assigned a POS tag and connected with the aforementioned dependency tag. The most frequent co-occurrences include:

(17) **wí pé** 'say that'
nítorí náà 'for that reason'
nígbà tí 'when'

The **compound** tag is used for any kind of compounding. Since UD allows language-specific extensions, Yorùbá uses 2 relation subtypes: **compound:prt** to attach verbal particles to verbs and **compound:svc** to connect verbs in a serial verb construction. It is worthy of note that both these extensions have already been used in other languages in UD.

(18) *ẹni tí ń* ***bọ̀*** ***wá*** *sí ayé*
someone that is come come to earth
'someone coming to the world'

See Figure 4 for the dependency tree of example (18).

Furthermore, there are many instances of **wí pé** ('say that') which is a multi-token expression and the words are tagged individually. The verb **wí** is tagged according to its function with respect to its parent, while **pé**

is connected to **wí** via a **compound** relation. Other instances of **pé** are tagged SCONJ.

fixed: fixed multiword expression – is used for certain fixed expressions that behave like function words or short adverbials.

(19) *ìwọ a máa gbọ́ ti èmi* ***nígbà gbogbo***
you usually will hear of I always

'thou hearest me always'

See Figure 5 for the dependency tree of example (19).

5. Parsing Experiment with UDPipe

We have briefly explained Universal Dependencies framework, how it has been applied in YTB and how the information is encoded in the CoNLL-U file format, suitable for a parser to process. The end goal is to automatically annotate new Yorùbá texts and to examine the accuracy and quality of the dependency annotation using UDPipe.

UDPipe is an easy-to-use and open-source trainable pipeline for tokenization, POS tagging, lemmatization and dependency parsing (Straka et al., 2016). It generates a model for tokenization, tagging and parsing based on training data in the CoNLL-U format. UDPipe can then use the model to process new raw text and annotate it automatically. UDPipe itself is language-independent. It can be considered as a universal tool for working with UD and it has been evaluated in the CoNLL 2017 and 2018 shared tasks on Universal Dependencies (Straka and Straková, 2017).

We wanted to use the parser as soon as possible, hoping that it will reduce the amount of work when annotating additional data. That means that the experiment was done with the first batch of manually annotated sentences, arguably very small to train a decent model: there were 200 sentences (about 5K tokens). We split the data into different sets for the experiment.

The standard evaluation metric in dependency parsing is labeled attachment score (LAS), that is, the percentage of words that have both their parent and the relation type (label) assigned correctly. If c_w denotes the number of words in the test set, and c_{ok} denotes the number of words whose parser-predicted parent id and relation label match those in the gold standard data, LAS is defined by the formula

$$LAS = 100 \times \frac{c_{ok}}{c_w}$$

A less strict metric is unlabeled attachment score (UAS), where only the parent word is considered but not the relation label.

5.1. The First 100 Sentences

(Oluokun, 2018) reports on several experiments evaluated on her 100 manually checked sentences. Even though our corpus is now larger, we first focus on comparison with the previous results, taking only the 100 previously published sentences into account. We assume that the first public release of YTB, which was part of UD release 2.2, is very close (although probably not identical) to the data used in (Oluokun, 2018)'s experiments. We then compare it to YTB in UD release 2.5, which is still only 100 sentences; nevertheless, we thoroughly revised the annotation between the two releases, it now reflects our improved guidelines and follows them more consistently. We thus expect better

Dependency Relation	Description	Frequency
acl	adnominal clause	50
advcl	adverbial clause	49
advmod	adverbial modifier	102
amod	adjectival modifier	34
appos	appositional modifier	18
aux	auxiliary	107
case	case marking	204
cc	coordinating conjunction	147
ccomp	clausal complement	70
compound	compound	36
compound:prt	compound particle	17
compound:svc	compound serial verb construction	58
conj	conjunct	109
cop	copula	19
det	determiner	72
discourse	discourse element	4
expl	expletive	51
fixed	fixed multiword expression	54
goeswith	wrongly split token	1
mark	subordinating marker	121
nmod	nominal modifier	110
nsubj	nominal subject	321
nummod	numeric modifier	9
obj	object	163
obl	oblique nominal	128
orphan	orphan	2
parataxis	parataxis	31
punct	punctuation	449
root	root	100
vocative	vocative	9
xcomp	open clausal complement	23

Table 2: Universal dependency relations statistics

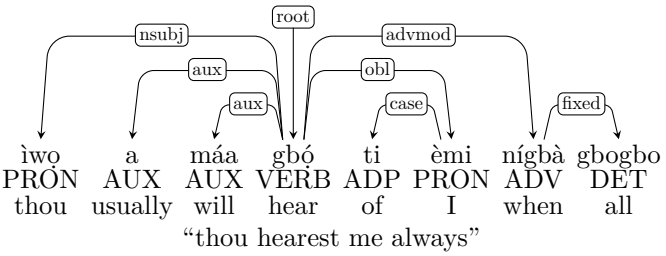

Figure 5: **fixed** example for (19)

parsing accuracy on the new version.
The results are summarized in Table 3 where we also repeat the findings of (Oluokun, 2018). 50:50 refers to an experiment where 50 sentences were used as training data and 50 sentences as test data. 90:10 refers to experiments where the dataset is split to 10 parts of 10 sentences each, then 10 different models are trained, always skipping a different part, which is reserved for evaluation. The average score from the 10 runs is reported in the table. These scores are higher than in the 50:50 experiment because more data is used for training; they should be also more reliable than a result on one randomly picked subset. Our own results on UD 2.2 are significantly lower than those reported by Oluokun; since we use the same version of UDPipe (1.2) and the same hyperparameters, this could mean that our assumption was wrong and the annotation published in UD 2.2 actually differs from the data used in Oluokun's experiments. On the other hand, on UD 2.5 we obtain scores that outperform all previously reported results.

	UPOS	**UAS**	**LAS**
Oluokun 50:50	85.23	58.34	50.00
Oluokun 90:10	89.56	68.00	58.10
UD 2.2 90:10	88.38	65.98	55.68
UD 2.5 90:10	90.65	69.95	60.92

Table 3: Results of universal part-of-speech (UPOS) tagging, unlabeled (UAS) and labeled attachment score (LAS) with 100 sentences.

The improvement can be traced to the hyphenated words which are common in many of the sentences in the corpus. We have also taken care of some of the ambiguous and frequently-occurring function words in the corpus as highlighted by (Oluokun, 2018).

5.2. All 200 Sentences

We now take the entire corpus of 200 sentences and divide it in the same 90:10 fashion as we did previously with 100 sentences. The training and test set sizes of the ten test runs are summarized in Table 4. Due to the small size of our data, UDPipe parameters for parser training were taken from (Oluokun, 2018) and not fine-tuned for this experiment.
The UAS and LAS scores are shown in Table 5. As expected, bigger data leads to higher scores. The scores are promising also in absolute terms, given that the model was trained only on 180 manually annotated sentences.

Run	**TrSent**	**TrWord**	**TsSent**	**TsWord**
1	180	5102	20	481
2	180	4975	20	608
3	180	5068	20	515
4	180	5108	20	475
5	180	4991	20	592
6	180	4995	20	588
7	180	4961	20	622
8	180	5082	20	501
9	180	4884	20	699
10	180	5081	20	502

Table 4: Data splits of the 200-sentence corpus for ten-fold cross-validation. Number of training sentences (TrSent), training words (TrWord), test sentences (TsSent) and test words (TsWord) for each run. There are 5583 words in total, and 27.9 words per sentence on average.

	UPOS	**UAS**	**LAS**
200 sentences 90:10	92.63	71.77	64.88

Table 5: Average results of universal part-of-speech (UPOS) tagging, unlabeled (UAS) and labeled attachment score (LAS) with 200 sentences.

It is difficult and potentially misleading to compare parsing scores in different languages; nevertheless, we would like to provide some context by looking at a few results from the CoNLL 2018 shared task (Zeman et al., 2018). Table 6 shows shared task results for four treebanks from the same domain, i.e., the Bible. The parser used in our experiments, UDPipe 1.2, served as the baseline parser in the shared task; in addition, we also show the score of the best parser for each treebank. The shared task setting was different because the systems had to process raw text while in our experiments the parser has access to gold-standard tokenization. On the other hand, the other four treebanks contain significantly larger training sets. Interestingly, the scores obtained by UDPipe 1.2 on these treebanks are comparable to our result on Yorùbá.
Of course, a labeled attachment score of 65% is still too low to be useful for downstream applications, even more so if we consider that our training and test data

Lng	TrSent	TrWord	UDPipe	Best
grc	15014	187033	67.57	79.25
la	15917	172133	59.66	73.61
got	3387	35024	62.16	69.55
cu	4124	37432	65.46	75.73
yo	180	ca. 5000	64.88	—

Table 6: LAS of UDPipe 1.2 (baseline) and of the best parser in the CoNLL 2018 shared task on Bible corpora in Ancient Greek (grc), Latin (la), Gothic (got), Old Church Slavonic (cu), and Yorùbá (yo; our work, not in the shared task).

are extremely similar and come from a rather peculiar domain. However, the parser is still very helpful for speeding up further annotation, as manual correction of parsing errors is significantly easier than doing all the annotation manually. For this purpose we trained a model on the entire set of 200 sentences and used it for automatic pre-annotation of 200 newly selected sentences from the unannotated parts of the Yorùbá Bible. We have done this to have a larger dataset for annotating Yorùbá Wikipedia articles; an attempt to incorporate other genres of Yorùbá text into the YTB. We have manually corrected the automatically parsed 200 sentences as part of our drive for rapid annotation and expansion of the treebank.

In the future, we will regularly re-train ("bootstrap") the parser on newly verified data, which will gradually decrease the amount of manual post-editing needed.

6. Conclusion

This paper contributed to the computational research on Yorùbá and its first dependency treebank. We explained why we chose the Universal Dependencies framework and why it is beneficial for the task and how it also enabled the creation of the annotation guidelines for Yorùbá; following the Universal Dependencies guidelines and highlighting language-specific issues. Moreover, the scheme caters for general language analysis, germane to future development of the treebank.

We also examined the most important language-specific cases where justifiable annotation choices were made. While the treebank is relatively small at present it is growing fast and it is open-source, allowing others the opportunity to contribute.

The parallel nature of the annotated text opens the door for cross-lingual studies featuring Yorùbá. The expanded treebank will be released and freely available as part of Universal Dependencies release 2.6 in May 2020 (a subset was already included in the previous releases).

This work has also laid a foundation for exploring other genres with the use of Wikipedia articles for automatic assignment of POS tags and dependency relations from our trained model using UDPipe.

6.1. Future Work

It is important for the treebank to grow rapidly, there are also potential areas that require attention. In essence, this work has opened up potential multi-faceted research goals for future work which are highlighted below:

1. We have added morphological features to pronouns in the treebank but there is need to assign features to other word categories.
2. Explore other domains of Yorùbá texts to provide diversity, depth and breadth of data that can be used for training, testing and creation of gold standard data. We plan to pay attention to texts that are not tone-marked. When data from other genres are incorporated, there will be a need to review the annotation scheme to cater for new patterns uncovered.
3. A Bible verse is taken as a sentence in the treebank but there are some verses that have more than one sentence. We plan to split these kind of sentences without loosing information about their origins; the *Ref* keyword in the MISC column encodes information about the chapter and verse of each sentence.
4. We need to adjust UDPipe parameters using more data in future experiments to determine if significant improvement is achievable.
5. While the present annotation has been worked on by three annotators in total, lack of manpower prevented us from genuine independent double annotation and cross-validation on the same text. We hope to find other annotators in the future and use them to assess annotation consistency.

Acknowledgements

This work has been supported by LINDAT/CLARIAH-CZ, the grant no. LM2018101 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

Adékẹ́yè, F. B. (2016). Preposition as a non-lexical class in standard Yorùbá. International Journal of Humanities and Cultural Studies (IJHCS) ISSN 2356-5926, 1(4):24–31.

Adesola, O. (2005). Yoruba: A grammar sketch: Version 1.0. Rutgers University, USA.

Àjàní, T. T. (2001). Aspect in Yoruba and Nigerian English. Ph.D. thesis, University of Florida.

Akinlabí, A. and Adéníyì, H. (2017). The language and its dialects. In Tóyìn Fálọlá et al., editors, Culture and Customs of the Yorùbá, pages 31–43. Pan-African University Press.

Bamgboṣe, A. (1983). Orthographies of Nigerian Languages: Manual. Number v. 1 in Orthographies of Nigerian Languages: Manual. National Language Centre Lagos, Federal Ministry of Education.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In LREC, volume 14, pages 4585–4592.

Ishola, O. (2019). Universal Dependencies for Yorùbá. Master's thesis, Eberhard Karls Universität Tübingen & Charles University in Prague.

Lawal, N. (1987). Yoruba relativisation and the continuous segment principle. Studies in African linguistics, 18(1):67–79.

Lu, X. (2014). Computational methods for corpus annotation and analysis. Springer.

Nivre, J., De Marneffe, M.-C., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20).

Ogunbiyi, I. A. (2003). The search for a Yorùbá orthography since the 1840s: obstacles to the choice of the arabic script. Sudanic Africa, 14:77–102.

Oluokun, A. (2018). Creation of a dependency treebank for Yoruba using parallel data. Master's thesis, Charles University in Prague.

Pajas, P. and Fabian, P. (2000). Tree editor TrEd, Prague dependency treebank, Charles University, Prague.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

Roorda, D. (2017). Practical linguistic annotation: The Hebrew Bible. International Journal of Humanities and Arts Computing, 11(2):276–288.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In LREC.

Tesnière, L. (1959). Éléments de syntaxe structurale.

UNESCO. (2013). World Braille usage.

Yoruba Cross-Border Language Commission. (2017). Modern Yorùbá writing manual.

Yusuf, O. (2010). Yorùbá syntax. In Ore Yusuf, editor, Basic Linguistics for Nigerian Languages, pages 262–277. Shebiotimo Publications.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In LREC, volume 2008, pages 28–30.