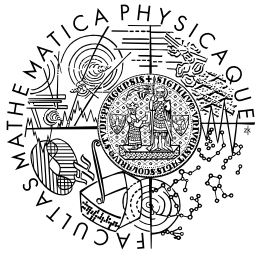

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Adedayo Oluokun

Creation of a Dependency Treebank for Yoruba using Parallel Data

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	RNDr. Daniel Zeman, Ph.D.
Study programme:	Computer Science
Study branch:	Computational Linguistics

Prague 2018

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

I would especially like to thank Dr. Tunde Adegbola, Tobi Ojo who provided their expertise of the subject to me. I give my special gratitude to my local coordinator Prof. Marketa Lopatkova, supervisor RNDr. Daniel Zeman, Ph.D. and the LCT program for giving me this opportunity.

Most of all, I would like to thank my husband, Bolutife Ogunsola, family and friends for support and patience throughout the countless hours spent on this work.

I dedicate this thesis to them.

Title: Creation of a Dependency Treebank for Yoruba using Parallel Data

Author: Adedayo Oluokun

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Daniel Zeman, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The goal of this thesis is to create a dependency treebank for Yorùbá, a language with very little pre-existing machine-readable resources. The treebank follows the Universal Dependencies (UD) annotation standard, certain language-specific guidelines for Yorùbá were specified. Known techniques for porting resources from resource-rich languages were tested, in particular projection of annotation across parallel bilingual data.
Manual annotation is not the main focus of this thesis; nevertheless, a small portion of the data was verified manually in order to evaluate the annotation quality. Also, a model was trained on the manual annotation using UDPipe.

Keywords: dependency parsing, annotation, paralle data, projection, UDPipe, part-of-speech tagging, low-resource

Contents

Introduction

A fundamental hinderance to developing statistical parsers and taggers for low resource languages is the shortage or absence of annotated data. A reasonably sized manually annotated corpus is required. This is costly to build and labour intensive as this requires qualified manpower, treebank design, annotation guidelines, format specification.

How can we build a parser and tagger for Yorùbá without a treebank? We could leverage annotated data for resource rich languages such as English and French in conjunction with other languages such as Vietnamese, Ancient Greek, Gothic, Latin, Old Church Slavonic that have parallel data with Yorùbá to overcome the annotated resource shortage of Yorùbá. Ancient Greek, Gothic, Latin, Old Church Slavonic were chosen because there exist thousands of parallel manually annotated data between them and Yorùbá. English, French and Vietnamese were chosen because of their low morphology in conjunction with huge availability of parallel data between them and Yorùbá.

This thesis investigates a very promising approach which involves leveraging annotated data for resource rich languages to overcome the annotated resource shortage of Yorùbá. We use automatically word-aligned bilingual corpora to project annotations from resource-rich languages to Yorùbá.
We focus on two tasks, part of speech projection and dependency relation projection across word-aligned bilingual corpora. The corpora in these other languages were annotated in the Universal Dependencies format.
"Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective" Nivre et al. [2016]. The annotation scheme is based on an evolution of (universal) Stanford dependencies de Marneffe et al. [2014], Google universal part-of-speech tags Petrov et al. [2011], and the Interset interlingua for morphosyntactic tagsets Zeman [2008]. The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. [1]

We projected dependency relations from English to Yorùbá using the direct projection algorithm described by Hwa et al. [2005] after which the results were trained and tested on manually annotated data.

The results gotten from the cross lingual part of speech tagging and parsing were manually verified after which a manually annotated treebank was created for Yorùbá.

Chapter 1 discusses the structure of Yorùbá, Yorùbá parts-of-speech and morphology. Chapter 2 provides a description of related work that has been done in the field of cross-lingual part-of-speech tagging and parsing among resource rich and low resource languages.
Chapter 3 reports on our work, we use automatically word-aligned bilingual corpora to project annotations from English, French, Vietnamese, Ancient Greek, Old Church Slavic and Gothic to Yorùbá. Our approach is based on the general

[1] http://universaldependencies.org/

idea of *annotation projection* Yarowsky et al. [2001].

Chapter 4 describes how we projected dependency relations between English and Yorùbá. It also discusses the processes involved in the manual annotation of a dependency treebank for Yorùbá. Chapter 5 discusses the evaluation of the different models trained using UDPipe on the manually annotated data and compares it to the results gotten from projecting dependency relations to Yorùbá.

1. Yorùbá Language

1.1 Yorùbá Language

Yorùbá is a 'dialect continuum' spoken in West Africa with over 40 million native speakers . It is spoken majorly in Nigeria and Benin Republic and other parts of Africa, America and Europe. Yorùbá is one of the most widely spoken African languages outside Africa. Yorùbá is classified among the Edekiri languages, which together with Itsekiri and the isolate Igala form the Yoruboid group of languages within the Volta–Niger branch of the Niger–Congo family [1].

It comprises about 12 dialects such as Ìjẹ̀bú, Ẹ̀gbá, Ìjẹ̀ṣà, Ọ̀yọ́, Ọ̀wọ́, each of which differs considerably from the other phonologically and lexically, and, to some extent, grammatically Bamgbose [1966].

1.1.1 Standard Yorùbá

Standard Yorùbá also known as Yorùbá Koine Bamgbose [1966], Literary Yorùbá, Common Yorùbá can be defined as the type of Yorùbá based on the Ọ̀yọ́ dialect. This is the type of Yorùbá learnt at school and spoken or written in formal settings.
The parallel Yorùbá texts used in this thesis are based on Standard Yorùbá.

1.1.2 Writing System

Yorùbá was initially written in Ajami script, a form of Arabic script. The Modern Yorùbá Orthography which used the Latin alphabet originated in the early work of the Church Mission Society(CMS) by Bishop Ajayi Crowther. Early versions of the bible were translated by Bishop Ajayi Crowther using the Latin alphabet but without sufficient tone markings. The only diacritic used was the under dot.

The current orthography of Yorùbá brings Yorùbá orthography in line with actual speech and contains all the diacritics needed to represent the tones.
Table 1.1 shows the Yorùbá alphabets. Table 1.2 shows the Yorùbá vowels with different diacritics representing different tones. In Yorùbá, there are also nasalised vowels such as ***an, in, on, ọn, un***.

Uppercase	A	B	D	E	Ẹ	F	G	Gb	I	H	J	K	L	M	N	O	Ọ	P	R	S	T	Ṣ	U	W	Y
Lowercase	a	b	d	e	ẹ	f	g	gb	i	h	j	k	l	m	n	o	ọ	p	r	s	ṣ	t	u	w	y

Table 1.1: Yorùbá alphabet

Uppercase	A	E	È	É	Ẹ	Ẹ̀	Ẹ́	I	O	Ò	Ó	Ọ	Ọ̀	Ọ́	U	Ù	Ú
Lowercase	a	e	è	é	ẹ	ẹ̀	ẹ́	I	o	ò	ó	ọ	ọ̀	ọ́	u	ù	ú

Table 1.2: Yorùbá vowels with diacritics representing different tones (*high, low and mid*)

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

1.1.3 Yorùbá Tonal System

Yorùbá is a tonal language with three tones, high, mid and low. The high tone is represented by the acute accent '◌́', the low tone is represented by the grave accent '◌̀' and the mid tone is unmarked.

1.1.4 Yorùbá Syllables

The syllable in Yorùbá is the smallest tone bearing unit. The three basic syllable types in Yorùbá are Vowel, Consonant-Vowel and Noun, consonant clusters are not allowed to occur in Yorùbá syllables. In Yorùbá, there are no consonant-final words.

1.1.5 Yorùbá as Low-Resource Language

Yorùbá is a low-resource language despite the fact that it has a large population of more than 40 million speakers. Yorùbá was classified as a seriously endangered language Bamgbose [1993]. Although Yorùbá is taught in schools, it is dominated by English. A lot of schools in Yorùbá speaking regions of Nigeria prefer to use English. This is as a result of the language policy in Nigeria which largely favours English over all indigenous languages.

English came to Nigeria as a result of colonialism and has seen been adopted as the official language in education, government and all official business. Fluency in English is seen as a sign of good quality education. Due to this, majority of parents emphasize that their children are taught in English at schools and are required to speak only English at home. Hence, many children therefore can neither speak, read nor write in Yorùbá and many do not even understand the language at all. In most educational settings, speaking indigenous languages is referred to as vernacular and highly prohibited. Failure to comply attract fines and in some cases corporal punishment.

There are little or no human language technologies for Yorùbá due to the lack of data. In Nigeria where Yorùbá has the highest number of speakers, there exist very few websites in Yorùbá and other indigenous languages, most websites are in English. On social media, the language of communication is majorly English. In the media, Yorùbá is largely used but by the older generation since there always exist the English which receive a greater patronage.

Another factor that might be a contribution to the low-resource status of Yorùbá is the great linguistic diversity of Nigeria. In Nigeria there are over 450 languages therefore necessitating the need for English as the lingua franca.

1.2 Part of Speech

1.2.1 Noun

Nouns in Yorùbá can be divided into the following classes Bamgbose [2010]. This division is based on the function or behaviour of nouns.

- **Concrete and Abstract Nouns**: Concrete nouns refer to things that can be seen. Abstract nouns refer to things that cannot be seen. Both types of

nouns differ in the type of verbs they agree with. Table 1.3 shows examples of concrete and abstract nouns in Yorùbá.

Concrete	**Abstract**
ẹja (fish)	ìfẹ́ (love)
àpótí (stool)	àláfíà (peace)
iwè (book)	ìrònú (worries)
igi (tree)	ikú (death)
ìyàwó (wife)	èrò (thought)

Table 1.3: Concrete and Abstract Nouns

- **Countable and Uncountable Nouns**: In Yorùbá, countable nouns are nouns that can be used with numerals while uncountable nouns are nouns that cannot be used with numerals. Table 1.4 shows examples of countable and uncountable nouns in Yorùbá.

Countable	**Uncountable**
ẹja (fish)	omi (water)
àpótí (stool)	iyanrin (sand)
ìwé (book)	irun (hair)
igi (tree)	òtútò (cold)
ìyàwó (wife)	ẹmu (palm wine)

Table 1.4: Countable vs Uncountable Nouns

- **Human or Non-Human**: A human noun is a valid answer to a '**ta**' (who) question, whereas a non-human noun answers a '**kí**' (what) question. An example is illustrated below:
  Wọ́n rí olùkọ́ : **Ta** ni wọ́n rí?
  They saw the teacher : Who did they see?

  Wọ́n rí ẹṣin : **Kí** ni wọ́n rí?
  They saw a horse : What did they see?
  Table 1.5 shows examples of Yorùbá Human and Non-Human nouns.

Countable	**Uncountable**
àbúrò (younger one)	ajá (dog)
ṣójà (soldier)	iyanrin (sand)
ọkùnrin (man)	irun (hair)
onílé (house owner)	òtútò (cold)
ìyàwó (wife)	ẹmu (palm wine)

Table 1.5: Human vs Non-Human Nouns

- **Nouns describing location**: In Yorùbá, a noun describing a location is a valid answer to a **'ibo'** (where) question and verbs **'ti'** or **'gbé'** can be used with the noun. For example:

Ó wà ní Èkó - Níbo l'ó wà?
He is in Lagos - Where is he?

Ó rí mi ní òkè - òkè ni ò **ti/gbé** rí mi
He saw me upstairs - Upstairs was where he saw me
Examples of this category of nouns are : orí (head), ìta (outside), ìsàlè (downstairs), ilé-ìwé (school)

- **Nouns describing time**: In Yorùbá a noun describing time is a valid answer to a **'ìgbà'** (when) question. For example:
Ó lọ ní àná : **ìgbà** wo ní ó lo?
He went yesterday : When did he go?

A dìde ni àáro : **ìgbà** wo ni a dì'de?
We got up in the morning : When did we get up?

- **Numbering Nouns**: These are nouns used for counting. Examples are ení(first), èjì(second), ẹ̀ta(third), ẹ̀rin(fourth), oókan(one).

- **Quantifying Nouns**: These are nouns that quantify other nouns, they cannot stand alone. Example of this category of nouns include ọ̀pọ̀ (plenty), sàsà (rare), ìdàjì (half).
ọ̀pọ̀ ènìyàn
Plenty people

- **Monetary Nouns**: A monetary noun is a valid answer to a **'élò'** (how much) question. For example,
Mo pa Naira meta : **'Élò'** ni mo pa?
I earned three Naira : How much did I earn?

O ku kobo merin : **'Élò'** ni o ku?
It is remaining four kobo : How much is remaining?

- **Descriptive Nouns**: These includes nouns used for description such as eyi (this), iyen (that), iwonyi (these), iwonyen (those) e.t.c.

- **Question Nouns**: These are nouns used instead of another noun when asking questions. Examples include ki (what), ta (who), èwo (which), èló (how much), ibo (where), èkélòó (how many times).

1.2.2 Verbs

Bamgbose [2010] divides Yorùbá verbs into the following categories. This division is based on the function or behaviour of verbs in a sentence.

- **Verbs that go with 'do'**: This category of verbs focus on the action of the subject. These are verbs that can be replaced with **'se'**(do) when forming a question based on the sentence. Also if a verb and its object can be replaced with **'se'** when forming a question based on the sentence, the verb belongs to this class. Examples include:
Olu lo : Ki ni Olu **'se'** ?

Olu went : What did Olu do?

Aja naa n to : Ki ni aja naa n **'se'** ?
The dog is urinating : What is the dog doing

This category of verbs has the highest number of verbs in Yorùbá, verbs such as je (eat), ta (sell), rin (walk), dide (stand), jade (go out), jokoo (sit), sun (sleep).

- **Explanatory or Informative Verbs**: This category of verbs focus on the action of the subject or object. These are verbs that can be replaced with **'se'**(happened) and the subject becomes the object when forming a question based on the sentence. For example,
  Okunrin naa ku : Ki ni o **'se'** okunrin naa?
  The man died : What happened to the man?

  Ebi pa okunrin naa : Ki ni o **'se'** okunrin naa?
  The man was hungry : What happened to the man?

  Verbs in this category include ku (die), fo (break), jo (leak), pa (kill) e.t.c.

- **Adjectival Verbs**: This category of verbs is used for qualifying the subject of a sentence. These are verbs that can be replaced with **'ri'**(look) when forming a question based on the sentence. For example,
  Okunrin naa gaa : Bawo ni okunrin naa ti **'ri'**?
  The man is tall : How does the man look?

  Eja naa tobi : Kini ni eja naa ti ri?
  The fish is big : How does the fish look?

  Verbs in this category include, ga (tall), po (plenty), tobi (fat), dun(sweet), gun (long), kun (full) e.t.c.

- **Narrative Verbs**: This category of verbs focus on the experience of a subject in a sentence. There is no way to form a question based on this sentence using **'ri'** or **'se'**. For example,
  O mo ise
  He is experienced

  O feran owo
  He loves money

  Verbs in this category includes mo (know), feran (love), gbadun(enjoy), koriira(hate)

- **Serial Verbs**: In this category, two words function as a verb. This category can be broken down into four sub-categories.

1. Serial verb in which one of the words cannot stand as a verb. Examples include ba je (spoil), fi si (put)
   Won **ba** fere **je**
   They spoilt the flute

   O **fi** ata **si** oju
   He put pepper in his face
2. Serial verb in which one of the words can stand as a verb. Examples include ba wi (correct), ba mu (match), fi han (show), tun se (repair)
   O **ba** won **wi**
   He corrected them

   Aso yii **ba** mi **mu**
   This cloth matches me
3. Serial verb in which both words can act as a verb but have different meanings when they stand alone and when they are together. Examples include be wo (visit), gba gbo (visit), bu ku (shorten), ko lu (hit)
   O **be** mi **wo**
   He visited me

   O **gba** mi **gbo**
   He believed me
4. Serial verb in which both words can act as a verb but the meaning of the verb is based on just one of the two words. Examples include ja bo (reply), fa ya (tear), tan je (deceive), gbe ta (sell).
   Aso naa **fa ya** ⇔ Aso naa **ya**
   The cloth tore ⇔ The cloth tore

   O **gbe** ile baba re **ta** ⇔ O **ta** ile babe re
   He sold his father's house ⇔ He sold his father's house

- **Repetitive Verbs**: This category of verbs work in a similar way as serial verbs. Here, the same verb is repeated twice. Examples include
  Kobo lo **ku** mi **ku**
  I have just kobo remaining

  O le **da** mi **da** oro naa
  He can leave me solely with the words

- **Nominal Verbs**: These are verbs from concatenation of a verb and Noun. Examples are illustrated in Table 1.6

Verb	**Noun**	**Verb**
gba (collect)	ina (light)	gbina (spark)
gbe (carry)	ese (leg)	gbese (to be in debt)
la (open)	oju (eye)	laju (enlighten)
wo (enter)	ile (house)	wole (to enter)

Table 1.6: Nominal Verb formed from Verb+Noun

. Other categories of verbs in Yorùbá are Null verbs, Elongated verbs, Causative verbs, Interrogative verbs, Passive verbs, Commands and Adverbials.

1.2.3 Pronouns

Pronouns in Yorùbá can be divided into 3 categories

1. Subject: These are pronouns that act as subject. Table 1.7 shows these pronouns

	Singular	**Plural**
First Person	mo (I)	a (we)
Second Person	o (you)	ẹ (you‘)
Third Person	ó (he/she)	wọ́n (they)

Table 1.7: Yorùbá Pronouns in Subject position

2. Object: These are pronouns that can act as an object in a sentence. Table 1.8 shows these pronouns

	Singular	**Plural**
First Person	mi (me)	wa (us)
Second Person	ọ (you)	ẹ (you)
Third Person	i (him)	wọ́n (them)

Table 1.8: Yorùbá Pronouns in Object position

3. Possessive Pronouns: This is illustrated in 1.9.

	Singular	**Plural**
First Person	mi (my)	wa (our)
Second Person	rẹ (your)	yín (your)
Third Person	rẹ̀ (his)	wọ́n (their)

Table 1.9: Yorùbá Possessive Pronouns

1.2.4 Adverbs

According to Bamgbose [2010], there are two types of adverbs in Yorùbá.

1. **Noun qualifiers** : This is similar to ‘adjectives’ in English. This category of adverbs qualify a noun or noun phrase.

2. **Sentence based adverbs**: These are adverbs that qualify verbs or verb phrases in a sentence such as pátápátá (totally), púpọ̀ (many, a lot), gidigidi (very much), fòò.
   Mo je e **pátápátá**
   I finished eating it

O dun **púpọ̀**
It was very sweet

Sentence based adverbs have the following characteristics:

- The low tone of a verb before an adverb doesn't change.
- They can stand after an object.
- They cannot be used to replace the object in a sentence.
- Their position after an object cannot be changed.

1.3 Morphology

Morphology in Yorùbá is mostly derivational, not inflectional. Grammatical relations are expressed with little or no inflection. This means that it does not complicate lemmatization and assignment of morphological features when we built our treebank.
Yorùbá is an analytical language. Words are formed in Yorùbá in three major ways, affixation, compounding and reduplication.

1.3.1 Affixation

Prefixes

Prefixation is very dominant in the word formation process of Yorùbá. Adewole [1995] describes the morphological structure of prefixation in Yorùbá and divides them into categories based on combination of semantic, syntactic an phonological factors.
This is a modification of the work done by Awoyale [1981]. Some examples are from Awobuluyi [1978]. Below are ways by which nominal compounds are formed in Yorùbá. These prefixes are attached to predicative phrases.

1. **à- factive nominal** : This denotes agentive when it attaches to a verb to give the meaning "doer of an action" or "one who is in a state of verb". This prefix takes a midtone.
   Examples of when it attaches to a verb to give the meaning "doer of an action" can be seen below

   (a) a + pa ẹran → apẹja
   PREFIX + kill animal → hunter

   (b) a + fọ́ ju → afọ́ju
   PREFIX + break eye → blind person

2. **à- Consequential**: This prefix attaches mostly to intransitive verbs to yield words meaning the result of the action or state expressed by the verb. The transitive verbs in this class do not require noun complements but can be accompanied by other verbal elements constituting a serial sequence.

(a) à + lo → àlo
PREFIX + to go → departure

(b) à + se → àse
PREFIX + be fulfilled → order

3. **é- Consequential**: This prefix takes monosyllabic verbal stems to form nouns which also appear to be mostly result noun. The prefix has alternates e and ẹ as required by Vowel Harmony.

(a) é + tò → étò
PREFIX + arrange → arrangement

(b) ẹ́ + gàn → ẹ́ gàn
PREFIX + deride → derision

4. **ì- Consequential**: This prefix is similar to **à- Consequential** as it mostly attaches to intransitive verbs to yield words meaning the result of the action or state expressed by the verb.

(a) ì+ tò → itò
PREFIX + to urinate → urine

(b) ì + fé → ìfé
PREFIX + to love → love

5. **ì- Action**: This category forms action names from transitive and intransitive verbs. The verbs are either transitive or compounded from the combination of a simple transitive verb and its argument.The prefix tone is either middle or low.

(a) ì + jà → ìjà
PREFIX + to fight → a fight

(b) ì + ya + ẹnu → ìyanu
PREFIX + open + mouth → surprise

6. **ì- Implement**: This prefix attaches to the verb it implements. In a lot of cases, the verbs consist of a transitive head and the patient that receives the action of the implement.

(a) ì + kò + ilẹ̀ → ìkòlẹ̀
PREFIX + collect + dust → dust pan

(b) ì + gbà + ilẹ̀ → ìgbàlẹ̀
PREFIX + kick + dust → broom

7. **ì- Thematic**: The nouns in this category are similar to the consequential nouns but they contain mostly human entities.

(a) ì + rán + isé → ìránsé
PREFIX + send + work → servant

(b) ì + jẹ + oyè → ìjòyè
PREFIX + eat + title → chief

8. **ò- Agentive**: This prefix takes stems of no more than two syllables to form nouns with reference to human entities. This prefix is ò or ọ̀ by the requirements of vowel harmony. The tone is low and does not vary.

   (a) ò + kú → okú
   PREFIX + die → corpse

   (b) ọ̀ + dà + ọ̀ràn → ọ̀dàràn
   PREFIX + create + trouble → trouble maker

Other processes are **ì- non-factive nominal**, **a-/o- agentive nominal**, **ò-/ọ̀- agentive nominal**, **e-/ẹ - consequential**,

Another predicative rule in Yorùbá is the 'N = N + VP' where a noun is derived from the affixation of a noun to a verb phrase.

1.3.2 Compounding

A compound noun in Yorùbá can be a combination of two or three nouns, or a noun and a verb.

Noun+Noun Compound

In Yorùbá, many compounds are formed by the concatenation of two or three nouns. A lot of these combinations are of modifier-modified type where one noun modifies the other noun. This is similar to possessive expressions where one noun modifies the other but in compounding in Yorùbá this is quite different as the modified noun almost always excludes the ability to possess and is mostly inanimate. Also, compounding in Yorùbá involves phonological features such as vowel deletion and tone swapping which is determined by a tonal hierachy which gives the highest priority to the *high* tone.

1. **Second vowel deletion**: In compounding two nouns, if the second noun begins with a vowel, the vowel is deleted and the tone is altered in some cases as seen below.

   (a) omi + oje→ omije
   water + sap → tears

   (b) erin + omi → erinmi
   elephant + water → hippopotamus

   (c) ojú + ò → ojúde
   eye + outside→ open space

2. **First vowel deletion**: In compounding two nouns, if the first noun begins with a vowel, the vowel is deleted and the tone is altered in some cases as seen below.

   (a) ògbó+ ẹni→ ògbẹ́ni
   elder+ someone → mister

   (b) omo + idan → omidan
   child + virgin→ miss

Noun+Verb+Noun Compound

Some compounds in Yorùbá are composed from three formatives, Noun Verb and Noun where the Verb is consistently identical to the verb 'ní' which means 'to have'. Examples are:

1. ìyá + ní + ọjà → ìyálọ́jà
   mother + has + market → trader (female)

2. baba + ní + ọjà → babalọ́jà
   father + has + market → trader (male)

à+NEG+ì Compounding

Certain Yorùbá nouns are formed when two derived nouns come together, one from **à- Consequential** class and the other from **ì- Consequential** class. Examples are:

1. à + bí + ì + kọ́ → àbíìkọ́
   PREFIX + to beget + PREFIX + to teach → untutored person(one who is born but not taught)

2. à + wí + ì + gbọ́ → àwíìgbọ́
   PREFIX+ to say + PREFIX + understand → one who never takes advice

1.3.3 Reduplication

Reduplication in Yorùbá involves one basic operation which is the copying of the first basic word in the stem. The basic word, which is usually the first minimal form within the stem that can stand by itself may fall under any of the following categories:

1. Basic verb, noun or adverb

2. Verb phrases

3. Derived nouns of the form 'PREFIX + Verb Phrase'

The result of the duplication process in Yorùbá could belong to one of three different syntactic categories Folarin [1989]

1. **Agentive nouns from phrasal verbs**: A verb phrase consisting of a transitive verb and its noun object is copied to form an agentive noun. Almost any transitive verb in combination with its complement can undergo this process in Yorùbá. Intransitive verbs are not involved in this type of reduplication.
   Also, this type of reduplication involves the regular processes of vowel deletion and tonal displacement. Some examples include:

   (a) jà ogun → jagunjagun
   fight war → warrior

(b) kó ilé (rob a house)→ kólékólé
gather house → thief

(c) mo ọ́ràn → mọ́rànmọ́ràn
know issues → savant

(d) gbé ọmọ→ gbọ́mọgbọ́mọ
carry child → kidnapper

2. **Quantified noun**: The process of noun quantification in Yorùbá involves the reduplication of nouns and the addition of a morpheme '**kì**' which stands between two nouns in a structure **Noun-kì-Noun** to yield the interpretation 'any Noun' or 'some disappointing Noun' Adewole [1995]. Almost any noun can ndergo this process.
Without the word '**kì**' between them, the resulting word is ungrammatical or read as something entirely different. Examples include:

(a) ilé → ilé-kì-ilé → ilékìlé
house → house-kì-house → any house

(b) ẹyẹ → ẹyẹ-kì-ẹyẹ → ẹyẹkẹyẹ
bird → bird-kì-bird→ any bird

(c) èyàn → èyàn-kì-èyàn → èyànkeyàn
person → person-kì-person→ any person

In each case the vowel in '**kì**' is deleted with the following tonal alignment taking place.

(a) H + M → H

(b) H + L → M

3. **Adverb of time/manner**: Yorùbá adverbs of time/manner can be divided into compositional and non-compositional adverbs.

(a) **Compositional adverbs**: Examples of compositional adverbs are:

i. ní ọwọ́ → lọ́wọ́lọ́wọ́
in hand → currently

ii. dá ojú → dájúdájú
clear eye → surely

iii. dí ẹ̀ → dí ẹ̀dí ẹ̀
little → slowly

(b) **Non-Compositional adverbs**: These types of adverbs are known as ideophones Awoyale [1981]. They are usually of the form 'CVCV-CVCV or CVV-CVV' Examples of non-compositional adverbs are:

i. bámúbámú - totally

ii. pátápátá - entirely

iii. wéréwéré - immediately

iv. mósámósá - promptly

2. Related Work

This chapter discusses work of other authors that is related to the task of projecting part of speech and dependency relations. We begin with a brief section on early approaches to part of speech projection. We then describe existing research in part of speech projection from resource rich languages to low resource languages and also dependency/syntactic projection across languages.

2.1 Early Part-of-Speech Projection approaches

One of the earliest approaches of Projection of Part of Speech is by Yarowsky and Ngai [2001] where they investigated the potential of projecting linguistic annotations such as part-of-speech tags and base noun phrase bracketings from one language to another via automatically word-aligned parallel corpora.
They carried out experiments to evaluate the accuracy of unmodified direct transfer of tags and brackets from the source language English to the target languages French and Chinese, both for noisy machine aligned sentences and clean hand aligned sentences. They boosted performance gotten from this experiment over both of these baselines by using training techniques optimized for very noisy data. They obtained 94-96% core French part-of-speech tag accuracy and 90% French bracketing F-measure for stand-alone monolingual tools. These monolingual tools were trained without the need for any manually annotated data in the given language.

Yarowsky et al. [2001] developed a set of methods for automatically inducing stand-alone monolingual part-of-speech taggers, base noun-phrase bracketers, named-entitiy taggers and morphological analysers for an arbitrary foreign language. 96% core part-of-speech accuracy was achieved when the induced stand-alone part-of-speech tagger was applied to French. Their experiments were done with no manually annotated data in the given language and with no linguistic knowledge or resources apart from raw text. Their performance exceeded that obtained via direct annotation projection.

2.2 Part-of-Speech Projection for Low Resource Languages

Agić et al. [2015] presents a method for learning part-of-speech taggers for low-resource languages like Akawaio, Aukan, Cakchiqquel for which only aggregation of parts of the bible exists. This is done via word alignment and aggregation of tags from few annotated languages. Their approach combines annotation projection, bootstrapping, and label propagation to learn POS taggers. They learnt POS taggers for 100 languages using the languages to bootstrap each other.
Their cross-lingual models was evaluated on the twenty five languages where they had test sets, and also on another ten for which they had dictionaries. Their results on the 25 test languages were better than the unsupervised baselines.

They obtain token-level accuracies of 80-90% for Afrikaans, Lithuanian and Russian. For Latin, Maori, Albanian and Ewe, they obtain token-level accuracies of 35-50%

Das and Petrov [2011] explored an approach similar to that of Yarowsky and Ngai [2001], they describe a novel approach for inducing unsupervised part-of-speech taggers for low-resource languages that have only aggregation of translated texts in resource-rich languages. Their method can be applied to other low-resource languages because they do not assume any knowledge of the target language.
They use a graph-based label propagation for cross-lingual knowledge transfer and use the projected labels as features in an unsupervised model. Their approach results in an average absolute improvement of 10.4% over the state-of-the-art baseline, and 16.7% over vanilla hidden Markov models induced with the Expectation Maximization algorithm.

Zeman and Resnik [2008] describes an approach for adapting a parser to a new language where one of the languages is low-resource. The technique was tested on two closely related European languages, Swedish and Danish.
The performance of their adaptation technique which involves using annotations in the source language achieves is equivalent to that obtained by training on 1546 trees in the target.

Agić et al. [2016] proposed a novel approach to cross-lingual part-of-speech tagging and dependency parsing for ow-resource languages.
They assume only linguistic resources that are available for most of the world's written languages, such as Bible excerpts and translations of the Watchtower. They extend annotation projection of dependencies relations across parallel text using a multi source approach. They introduce a new projection algorithm, this algorithm projects weight matrices from multiple sources rather than dependency trees or individual dependencies from a single source as seen in Hwa et al. [2005]. This method performs significantly better than commonly used annotation projection methods and delexicalized transfer baselines. This method performs well on low-resource non-Indo-European languages.

Agić [2017] introduced an unbiased approach for cross-lingual transfer of delexicalized parsers. They solve the problem of selecting the single best parser for any target language by proposing a lean method for parser selection. They propose a set of methods for matching texts to source parsers. Their methods rely on character based language identification and typological similarity.
Their best system exceeds the performance of single-best oracle source parsers without disadvantaging the truly low-resource languages.

2.2.1 Evaluating Part-of-Speech Projection for Low Resource Languages

One of the challenges of part-of-speech projection for low-resource languages is lack of testing data. In cross-lingual learning work, it is common to evaluate POS taggers for accuracy by using test data annotated by human experts. But what happens when there is no manually annotated data?
Agić et al. [2017] addresses this challenge by describing two dictionary-based metrics. They perform two sets of experiments numerical score prediction and

rank prediction. They compare the POS tagger rankings induced by evaluation against dictionaries to those induced by evaluation on manually annotated gold standards across 25 languages.
They introduce a novel metric that presumes nothing but an English tag dictionary and a small bilingual dictionary for the target language. They also found this metric to be a relatively robust estimator for tagging accuracy as this methods discovers the best tagger for 11 out of 20 languages. They discovered that translating a small list of frequent words from English is sufficient and reliable for evaluating crosslingual taggers for target languages.

2.2.2 Unsupervised Part-of-Speech Projection for Low Resource Languages

Purely unsupervised techniques for part-of-speech (POS) tagging are yet to achieve useful accuracies required by many language processing tasks.
Li et al. [2012] show how POS-taggers exceeding state-of-the-art bilingual methods can be built by using simple Hidden Markov Models (HMMs) and Wiktionary. They labelled data to evaluate results across eight languages and achieve an accuracy that significantly exceeds best unsupervised and parallel text methods

Fossum and Abney [2005] implement a variant of the algorithm described by Yarowsky and Ngai [2001] to induce an HMM POS tagger for an arbitrary target language using only an existing POS tagger for a source language and an unannotated parallel corpus between the source and target languages. They also project from multiple source languages onto a single target language and show that this method significantly improves the performance of automatically induced POS taggers on a target language

Which is more important, annotation of word types or tokens? when building a part-of-speech tagger for low-resource languages using semi-supervised learning. Garrette et al. [2013] perform various experiments to explore how the amount of time spent on manual annotation or gathering of resources affect performance. They use four types of data for Kinyarwanda and Malagasy, two low-resource languages.

Their results show that the combination of type supervision and an effective semi supervised learning approach is a very important source of linguistic information. They also showed that for a morphologically rich language such as Kinyarwanda, a morphological transducer can give good results when there is lack of manually annotated data.

Wisniewski et al. [2014] introduces an approach for projecting part-of-speech tags via ambiguous learning. They use a history based model Black et al. [1993] with Laso-like training Daumé and Marcu [2009]
They evaluated their approach on ten different languages and used English as the source language. Their method achieves an error rate of 10.4% for Czech, 8.8% for German, 10.2 % for French and 9.1% for Italian.

Duong et al. [2013] presents an interesting unsupervised method which is similar to annotation projection by Yarowsky and Ngai [2001] but different in that they developed a method to automatically filter good training sentences from their parallel data after which they apply self training. This was done by using a seed tagger from the directly-projected labels. Using self revision and

training they obtain part-of-speech tagging accuracy of 85.6% for Danish, 84.0% for Dutch, 85.4% for German and 81% for Swedish.

2.2.3 Part-of-Speech tagging for Low Resource Languages using Neural Networks

Zennaki et al. [2015] proposes an interesting approach to induce automatically a Part-Of-Speech tagger for resource-poor languages. This approach is also based on cross-lingual annotation projection from parallel text based on sentence alignment not word alignment. They make use of Recurrent Neural Networks as multilingual analysis tools.
Common words representation based only on sentence level alignment are extracted from a parallel corpus between a resource-rich language and a low-resource language. They achieved comparable results to the state-of-the-art by combining their approach with a basic crosslingual projection method.
They evaluated their approach by using only parallel corpora for four languages: French, German, Greek and Spanish. For English–German–Greek–Spanish multilingual part-of-speech tagger, they obtain a close to state-of-the-art result with only a subset (65,000) of Europarl corpus used.

2.3 Projecting Syntactic Relations

Hwa et al. [2005] explored using parallel text to solve the problem of creating syntactic annotation in more languages. They annotated the English side of a parallel corpus, projected the analysis to a second language and then trained a stochastic analyser on the resulting noisy annotations.
They train Collins's (1997) Model 2 parser on the Penn Treebank WSJ data and use it to parse the English side of a parallel corpus. The resulting parses are transformed into dependencies, these dependencies are projected to the second language using automatically obtained word alignments and the resulting dependencies cleaned up using a limited set of language-specific post-projection transformation rules.
A dependency parser for the target language is trained on this projected dependency treebank, and the accuracy compared with gold standard. They report dependency accuracy of 72.1% for Spanish, comparable to rule-based commercial parser; accuracy on Chinese is 53.9%.

3. Cross-lingual POS tagging and POS Voting

3.1 Cross-lingual POS tagging

We investigate a very promising approach which involves leveraging annotated data for resource rich languages to overcome the annotated resource shortage of Yorùbá. We use automatically word-aligned bilingual corpora to project annotations from resource-rich languages to Yorùbá. Our approach is based on the general idea of *annotation projection* Yarowsky et al. [2001].

3.2 Universal Part-of-Speech tags

Part-of-Speech projection across languages relies on the assumption that morpho-syntactic categories in the source and target language are the same.
This assumption might not always hold. The universal part-of-speech tags contains labels which are stable among languages. The tags in table 3.1 mark the core part-of-speech categories.

ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxilliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb
X	Other

Table 3.1: UD Part of Speech tags

3.2.1 Manually annotated data in Latin, Gothic, Ancient Greek, Old Church Slavonic

For the Cross-lingual Part of Speech tagging, parallel bible data was obtained between Yorùbá and the following languages Latin, Gothic, Ancient Greek, Old Church Slavonic.

These four languages are linguistically very different from Yorùbá but we were interested in exploiting the fact that manual syntactic annotation in UD already exists for these languages and has been released as part of the UD collection.

Tables 3.2, 3.3, 3.4, 3.5 show the number of sentences, tokens, word types between Yorùbá and the other languages. The parallel data in these four languages were manually annotated in the UD POS format. Originally, they were manually annotated using other tagsets and were later automatically converted to the UD tagset.

The manual annotation of data in these four languages consisted of sentences which were equivalent to different verses of the bible. These sentences contained bible verse references as part of the annotations. However, there were cases were one verse was split into multiple CoNLL-U sentences. In such case, we rearranged and merged these sentences into one.

	Yorùbá	**Latin**
Tokens	186556	93322
Word types	5877	12400
Sentences	5985	5985

Table 3.2: Yorùbá and Latin parallel data statistics

	Yorùbá	**Gothic**
Tokens	91301	46639
Word types	3956	7579
Sentences	2969	2969

Table 3.3: Yorùbá and Gothic parallel data statistics

	Yorùbá	**Ancient Greek**
Tokens	194927	107411
Word types	6127	15475
Sentences	6224	6224

Table 3.4: Yorùbá and Ancient Greek parallel data statistics

	Yorùbá	**Old Church Slavonic**
Tokens	100842	51483
Word types	3596	9114
Sentences	3251	3251

Table 3.5: Yorùbá and Old Church Slavonic parallel data statistics

3.2.2 Alignment of Parallel Corpus

Yorùbá sentences and the equivalent sentences in the other languages were word-aligned using Fast Align. Fast Align is a Simple, Fast, and Effective Reparameterization of IBM Model 2 Dyer et al. [2013]. Fast align generates asymmetric

alignments (i.e., by treating either the left or right language in the parallel corpus as primary language being modelled, slightly different alignments will be generated). There are different types of alignments such as

1. Intersection Alignment
2. Forward Alignment
3. Backward Alignment

For our Cross-lingual POS tagging we used the intersection of the forward and reverse alignments, i.e one to one alignment. Figure 3.1 shows the result of word alignment on the Yorùbá and Latin sentence in Table 3.6. As seen in Figure 3.1 only six Yorùbá words were aligned to Latin. Figure 3.2 shows the alignment between Yorùbá and Ancient Greek. Figure 3.3 shows the alignment between Yorùbá and Gothic. The visualisations were done using CoNLL-U viewer [1].

Yorùbá	Ẹ̀yin ará mi àwọn kan láti ilé kiloe sọ di mímọ́ fún mi pé ìjà ń bẹ́ láàrin yín
Latin	significatum est enim mihi de vobis fratres mei ab his qui sunt chloes quia contentiones inter vos sunt

Table 3.6: Yorùbá and Latin sentence
English Translation: *For it hath been declared unto me of you, my brethren, by them which are of the house of Chloe, that there are contentions among you.*

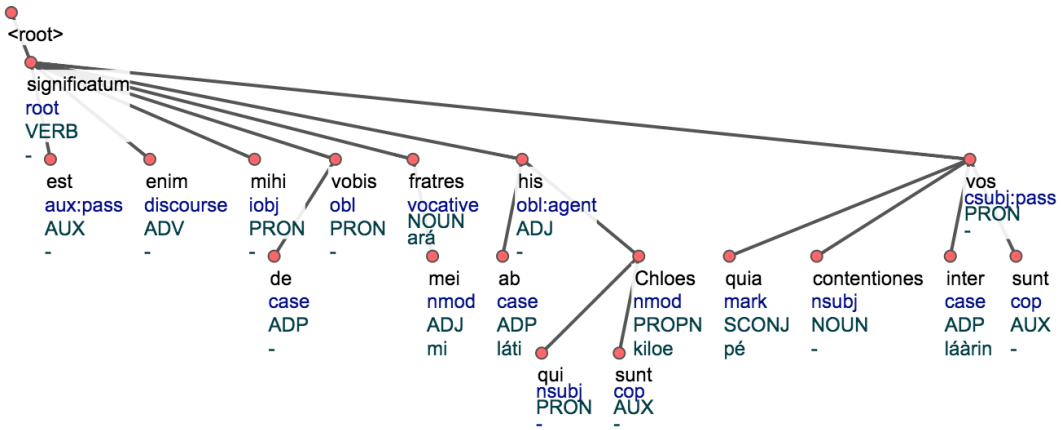

Figure 3.1: Example alignment between Yorùbá and Latin sentence. Analyses visualized using CoNLL-U file viewer

[1] http://universaldependencies.org/conllu_viewer.html

Yorùbá	Èyin ará mi àwọn kan láti ilé kiloe sọ di mímọ́ fún mi pé ìjà ń bẹ láàrin yín
Ancient Greek	ἐδηλώθη γάρ μοι περὶ ὑμῶν ἀδελφοί μου ὑπὸ τῶν χλόης ὅτι ἔριδες ἐν ὑμῖν εἰσιν

Table 3.7: Yorùbá and Ancient Greek sentence
English Translation: *For it hath been declared unto me of you, my brethren, by them which are of the house of Chloe, that there are contentions among you.*

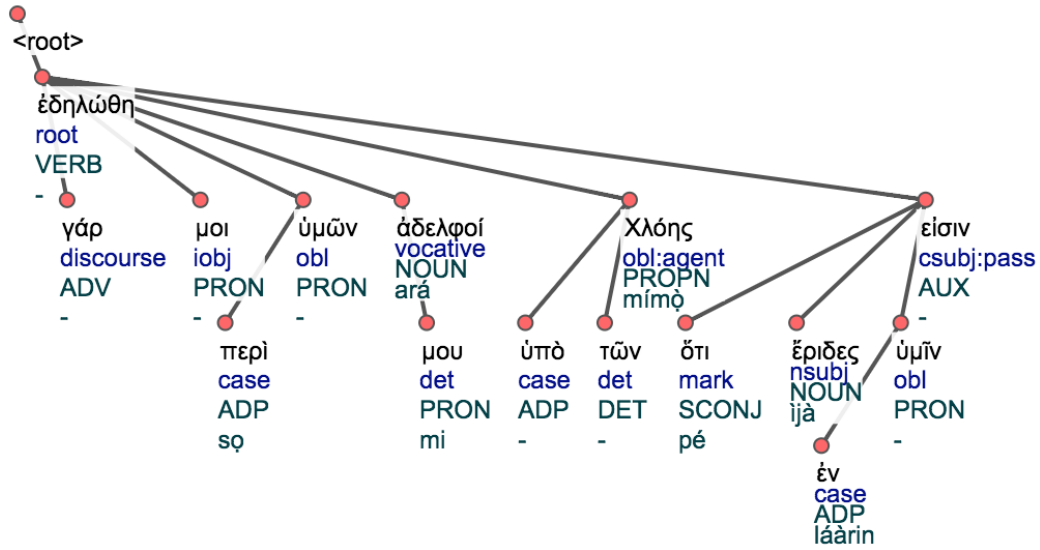

Figure 3.2: Example alignment between Yorùbá and Ancient Greek. Analyses visualized using CoNLL-U file viewer

Yorùbá	ohun tí mo ń sọ ni pé olúkúlùku yí ń ń wí pé èmi tẹ̀lé pọ́ọ̀lù èmi tẹ̀lé àpólò òmíràn èmi tẹ̀lé kéfà ìtúmọ̀ pétérù àti ẹlòmìíràn èmi tẹ̀lé kírísítì
Gothic	ik im pawlus ip ik apaullons ip ik kefins ip ik xristaus

Table 3.8: Yorùbá and Gothic sentence
English Translation: *Now this I say, that every one of you saith, I am of Paul; and I of Apollos; and I of Cephas; and I of Christ*

Alignment Results

For the alignment between Yorùbá and the other four languages, punctuations were removed from Yorùbá because these other languages do not contain punc-

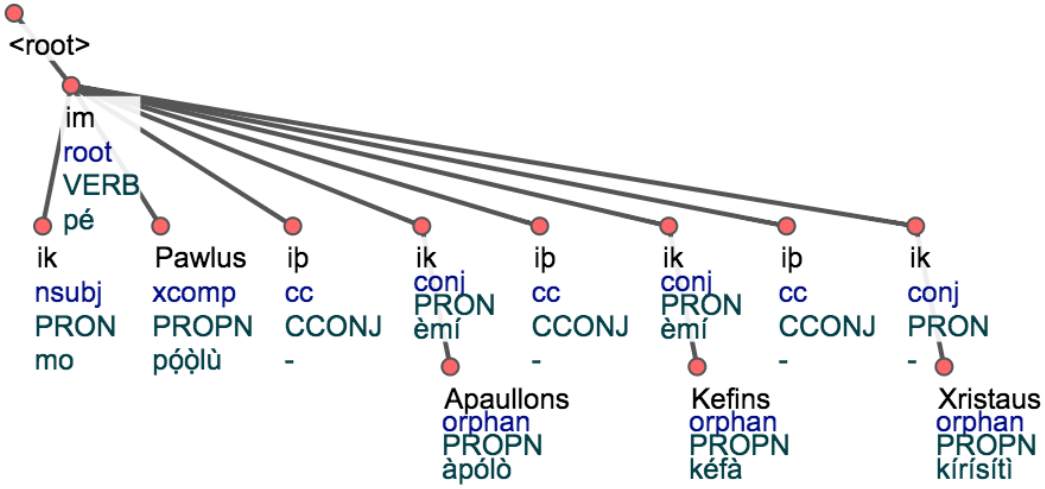

Figure 3.3: Example of Yorùbá and Gothic alignment. Analyses visualized using CoNLL-U file viewer

tuations. Tables 3.9, 3.10, 3.11, 3.12 show the result from these alignment.
30-35 % of Yorùbá words were not aligned.
This is due Yorùbá being an analytical language, with almost no morphology and a lot of function words while the other four languages are all morphologically rich. Thus the function words in Yorùbá correspond to morphological features in the other languages, and there are no independent words that could be aligned with the Yorùbá function words.

Aligned Tokens	57452
Unaligned Tokens	100211
Aligned Word types	4051
Unaligned Word types	1403

Table 3.9: Yorùbá and Latin Alignment result

Aligned Tokens	59488
Unaligned Tokens	105493
Aligned Word types	4359
Unaligned Word types	1333

Table 3.10: Yorùbá and Ancient Greek Alignment result

Aligned Tokens	27692
Unaligned Tokens	49581
Aligned Word types	2701
Unaligned Word types	932

Table 3.11: Yorùbá and Gothic Alignment result

Aligned Tokens	28779
Unaligned Tokens	55246
Aligned Word types	2547
Unaligned Word types	711

Table 3.12: Yorùbá and Old Church Slavic Alignment result

3.2.3 UDPipe

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/ OS X, as a library for C++, Python, Perl, Java and as a web service Straka and Straková [2017].

For the following experiments, trained models in English, Vietnamese and French were used. These models were trained on Universal Dependencies 2.0 treebank.

MorphoDiTa

Morphological Dictionary and Tagger is an open-source tool for morphological analysis of natural language texts. It performs morphological analysis, morphological generation, tagging and tokenization and is distributed as a standalone tool or a library, along with trained linguistic models Straková et al. [2014]

The UDPipe tagger consists of possibly several MorphoDiTa models, each tagging some of the POS tags and/or lemmas.

3.2.4 UDPipe annotated data in English, French and Vietnamese

We have seen earlier that Yorùbá isn't related to Ancient Greek, Latin, Gothic and Old Church Slavonic in terms of morphology. We are interested in augmenting our projection by using English, French and Vietnamese.
Vietnamese is an analytical language like Yorùbá. English is also a language which is not morphologically rich. Apart from these languages being less rich morphologically, there are thousands of parallel data from Watchtower Corpus [2] between them and Yorùbá and also parallel bible data between English and Yorùbá.

We obtain parallel bible data from the Edinburgh Bible Corpus (EBC) collected by [3] Christodouloupoulos and Steedman [2015] between Yorùbá and English.
Also Parallel data from the Watchtower Corpus between Yorùbá, English, French and Vietnamese was also obtained. The Watchtower Corpus and Edinburgh Bible Corpus (EBC) both consist of religious texts, but they are very different in terms

[2] http://wol.jw.org/
[3] http://homepages.inf.ed.ac.uk/s0787820/bible/

of domain, style and content. The data from Watchtower Corpus was already sentence-aligned.

Table 3.15, 3.16, 3.17 show the number of sentences, tokens, word tokens between Yorùbá and the other languages. The parallel data in these other languages were automatically tagged with UD POS using UDPipe.

Yorùbá	ìrètí ìgbàlà lè jẹ́ kéèyàn mọ̀kàn le kódà nígbà tínú bá jó dóríi kókó pàápàá.
English	the hope of being saved can help a person to hold on even in the direst of circumstances.

Table 3.13: Yorùbá and English sentence

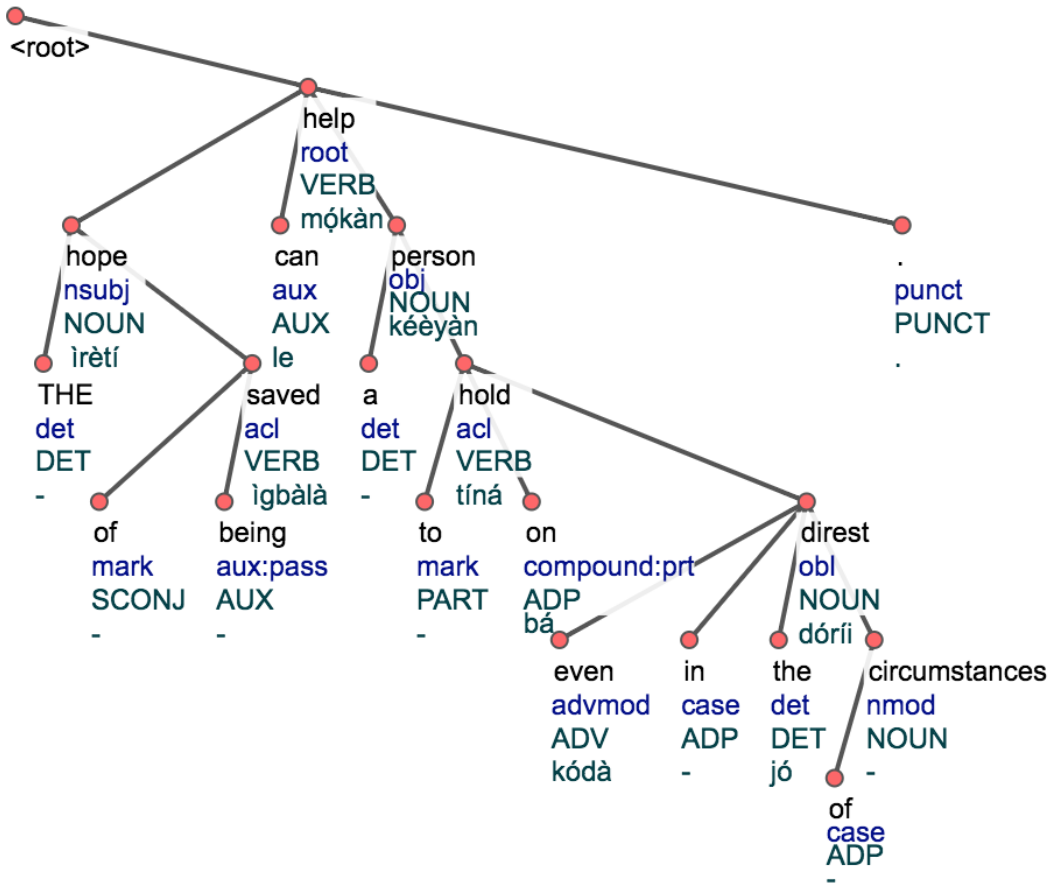

Figure 3.4: Example alignment between Yorùbá and English sentence.
The English tree is not a manual annotation but output of the UDPipe parser.
Analyses visualized using CoNLL-U file viewer

Yorùbá	ìrètí ìgbàlà lè jẹ́ kéèyàn mọ̀kàn le kódà nígbà tínú bá jó dóríi kókó pàápàá.
French	l'espoir d'etre secouru donne la force d'endurer les situations les plus désespérées.

Table 3.14: Yorùbá and French sentence

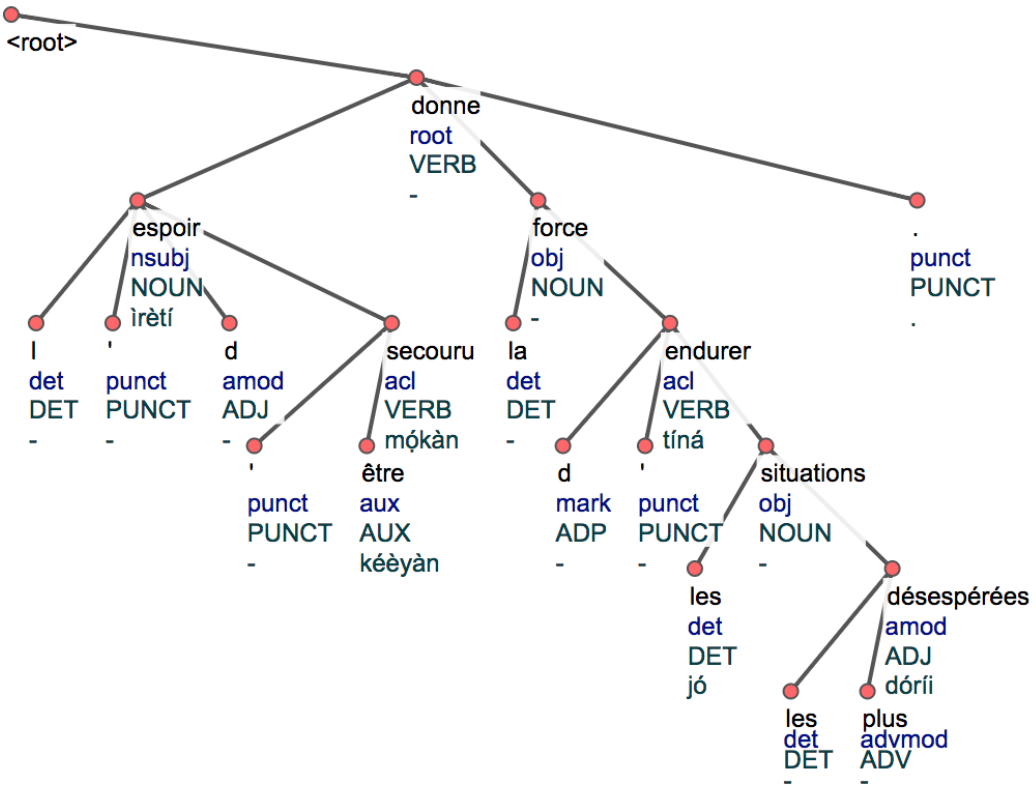

Figure 3.5: Example alignment between Yorùbá and French sentence.
The French tree is not a manual annotation but output of the UDPipe parser.
Analyses visualized using CoNLL-U file viewer

	Yorùbá	**English**
Tokens	945233	909610
Word types	18656	12579
Sentences	30866	30866

Table 3.15: Yorùbá and English parallel data statistics

	Yorùbá	**French**
Tokens	1952431	1786440
Word types	18010	37990
Sentences	92350	92350

Table 3.16: Yorùbá and French parallel data statistics

	Yorùbá	**Vietnamese**
Tokens	2091948	1669834
Word types	18728	49769
Sentences	97857	97857

Table 3.17: Yorùbá and Vietnamese parallel data statistics

3.2.5 Alignment of Parallel Corpus

Yorùbá sentences and the equivalent sentences in the other languages were word-aligned using Fast Align Dyer et al. [2013].

Alignment Results

Tables 3.18, 3.19, 3.20, 3.21 show the result from these alignment.

Aligned Tokens	553325
Unaligned Tokens	391908
Aligned Word types	10819
Unaligned Word types	7837

Table 3.18: Yorùbá and English Bible Alignment result

Aligned Tokens	1212604
Unaligned Tokens	1306594
Aligned Word types	16067
Unaligned Word types	4103

Table 3.19: Yorùbá and English Watchtower Alignment result

Aligned Tokens	903291
Unaligned Tokens	1049140
Aligned Word types	14449
Unaligned Word types	3561

Table 3.20: Yorùbá and French Watchtower Alignment result

Aligned Tokens	877117
Unaligned Tokens	1214831
Aligned Word types	15103
Unaligned Word types	3625

Table 3.21: Yorùbá and Vietnamese Watchtower Alignment result

3.3 POS Voting

3.3.1 Voting with Latin, Gothic, Ancient Greek, Old Church Slavonic

The projected UD Part of Speech Tags were used to tag the Yorùbá words.

For each word in each language, the part of speech with the highest frequency for that word was selected. Projected POS tags from Latin, Gothic, Ancient Greek, Old Church Slavonic might disagree for reasons such as erroneous source

annotations, incorrect word alignments, or legitimate differences in POS between translation equivalents. This is resolved by taking a majority vote. By letting several languages vote on the correct tag of each word, our projections become more robust, less sensitive to the noise in our source-side predictions and word alignments. In the case were there were equal number of votes for different parts-of-speech, we randomly selected one.

The union of word types for the four languages consists of 5,967 words. Table 3.22 shows the number of word types that were tagged with the projected UD part of speech by voting via the four languages. 14 % of the word types had no projected UD part of speech from all the four languages.

The UD part of speech and corresponding frequency is shown in 3.23.

Total word types	5967
Word types with POS	5144
Word types without POS	823

Table 3.22: Result of POS Voting using Latin, Gothic, Ancient Greek and Old Church Slavonic

POS	**Number of word types**
VERB	1894
NOUN	1723
ADJ	621
PROPN	428
ADV	183
NUM	73
PRON	66
ADP	60
DET	32
CCONJ	24
SCONJ	15
INTJ	11
AUX	7
X	6

Table 3.23: POS Statistics from POS Voting

3.3.2 Voting with Latin, Gothic, Ancient Greek, Old Church Slavonic, English, French and Vietnamese

The projected UD Part of Speech Tags were used to tag the Yorùbá words.

For each word in each language, the part of speech with the highest frequency for that word was selected. After this, a voting technique was used, the final part of speech tag for a word was the most frequent tag based on the votes from the languages. Adding more source languages significantly improves the performance of our POS tagger. The union of word types from the seven parallel consists of

35,430 words. Table 3.24 shows the number of word types that were tagged with the projected UD part of speech by voting via the seven languages.

The UD part of speech and corresponding frequency is shown in 3.25. By using parallel data from English, French and Vietnamese we are only able to cover more word types but also more parts-of-speech. Using parallel data from English, French and Vietnamese introduces a new part-of-speech *PART* which was not captured using Ancient Greek, Gothic, Latin and Old Church Slavonic.

Total word types	35430
Word types with POS	27162
Word types without POS	8268

Table 3.24: Result of POS Voting using Latin, Gothic, Ancient Greek and Old Church Slavonic

POS	**Number of word types**
NOUN	12594
VERB	4774
PROPN	2961
ADJ	2431
NUM	1946
ADV	792
ADP	386
PRON	365
INTJ	230
DET	157
PUNCT	151
AUX	121
X	87
CCONJ	68
SCONJ	52
SYM	26
PART	20

Table 3.25: POS Statistics from POS Voting

3.3.3 Analysis

One of the problems of cross-lingual POS tagger evaluation is absence of manually annotated test data. Since there is no manually annotated corpus for Yorùbá, we decided to select the top five words from each part-of-speech category and translate them to English.

Table 3.27, 3.26, 3.28, 3.29, 3.30, 3.31, 3.32, 3.33 3.34 3.35 3.36 shows the top Yorùbá words belonging to each UD POS category from the POS voting results and their approximate English translation.
These top Yorùbá words were translated by a native speaker of both Yorùbá and English.

While it is not always guaranteed that translational equivalents preserve the POS category, it is often the case. Hence we take these results as an indirect supporting evidence that our algorithm produces the desired output.

Word	**English Translation**
fi	put
sọ	speak
lọ	go
wá	come
mú	take
rí	see

Table 3.26: Top Yorùbá Verbs from POS Voting

Word	**English Translation**
èèyàn	person
ọmọ	child
kristẹni	christian
ìgbà	time
ọdún	year

Table 3.27: Top Yorùbá Nouns from POS Voting

Word	**English Translation**
tí	whom, who
wọ	them
rẹ	his, her, its
wa	us
mi	I, me

Table 3.28: Top Yorùbá Pronouns from POS Voting

Word	**English Translation**
gbogbo	all
àwọn	they
yìí	this
yẹn	those
tàwọn	theirs

Table 3.29: Top Yorùbá Determiners from POS Voting

Word	English Translation
kan	one
méjì	two
méje	seven
mẹ́ta	three
ọ̀kan	one

Table 3.30: Top Yorùbá Numerals from POS Voting

Word	English Translation
ní	in
sí	to
fún	for
láti	with
nínú	inside

Table 3.31: Top Yorùbá Adpositions from POS Voting

Word	English Translation
tó	if
pé	that
kí	that
bá	-
bí	as, if, although

Table 3.32: Top Yorùbá Subordinating Conjunctions from POS Voting

Word	English Translation
ni	to be
ti	have
jẹ́	to be
wà	to be
máà	will

Table 3.33: Top Yorùbá Auxiliary from POS Voting

Word	English Translation
sì	and, also
àti	and
tàbí	or
ṣùgbọ́n	but
àbí	or

Table 3.34: Top Yorùbá Coordinating Conjunction from POS Voting

Word	English Translation
náà	-
nígbà	when
bẹ́ẹ̀	so, thus

Table 3.35: Top Yorùbá Adverbs from POS Voting

Word	English Translation
ọ̀pọ̀	plenty
rere	good
dára	good
pátákí	important
pọ̀	plenty

Table 3.36: Top Yorùbá Adjectives from POS Voting

By introducing 3 more languages we were able to capture interesting language phenomena. The UD part of speech **PART** was discovered via alignment between English and Yorùbá.

Table 3.37 shows an English and equivalent Yorùbá sentence, its alignment is shown in Figure 3.6 to illustrate this.

Yorùbá	ọkùnrin náà àti aya rẹ̀ sì wà ní ìhòòhò, ojú kò sì tì wọ̀n .
English	and they were both naked, the man and his wife, and were not ashamed .

Table 3.37: Yorùbá and English sentence to illustrate the UD POS PART Projection

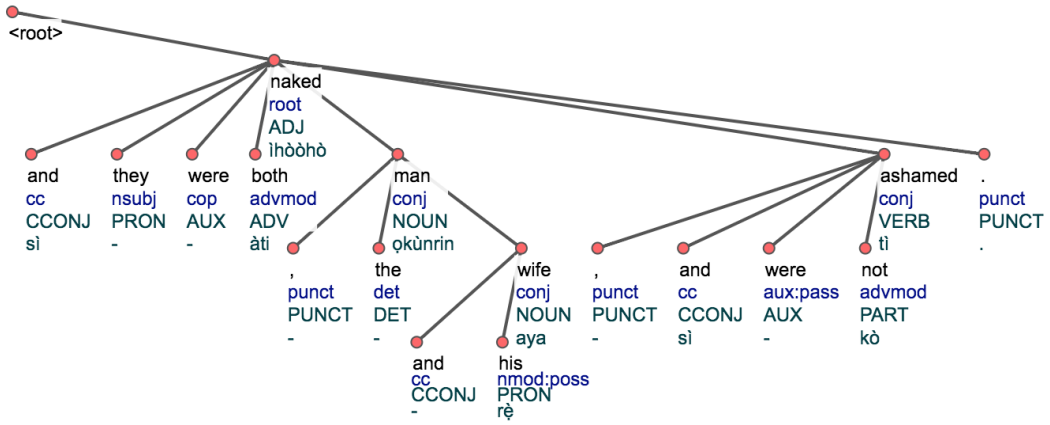

Figure 3.6: Example alignment between Yorùbá and English sentence. Analyses visualized using CoNLL-U file viewer

The words which had no UD part of speech tag from the projection were mostly words occurring with a frequency of 1 and words with wrong diacritics.

3.3.4 Manual Annotation of POS

In order to perform a more robust evaluation of how well our part-of-speech tagger works, we manually tagged 200 most frequent word types which had no projected UD part of speech. Also, 100 intersecting sentences from Latin, Gothic, Ancient Greek, Old Church Slavonic were manually tagged and compared with the result gotten via projection. An accuracy of 86% was obtained. Most of the errors were from function words which had multiple parts-of-speech because we only assign the most probable tag to each word type, without taking the context into account, ambiguous words inevitably introduce errors.

3.3.5 POS Tagger Training Using UDPipe

We tagged 113717 sentences with part-of-speech based on the ouput of the POS Voting after which these sentences were trained with UDPipe version 1.2.0.
The parameters used for training were the default UDPipe tagger parameters, the only change made was to the number of iterations. We used 10 iterations instead of 20. Table 3.38 and 3.39 show the statistics of the train and test data. 87.86% accuracy was obtained on the test data.

In order to test robustness of the tagger across text types, the Watchtower Corpus was used as the training set and the Bible was used as the test set. Even though both corpora consists of biblical texts, they are significantly very different.

The difference in accuracy gotten from UDPipe tagger (87.86%) and simple projection 86% might be due to the following reasons:

- For the simple projection, we only tested on 100 sentences, which is a really small fraction of our data. Increasing the test sentences might give a better accuracy.
- For the UDPipe tagger, training and testing was done on more data. This might be the reason we got a higher percentage.

Sentences	113717
Word Types	20170
Tokens	2519198

Table 3.38: Statistics of Training Data for UD POS

Sentences	30866
Word Types	18656
Tokens	945233

Table 3.39: Statistics of Testing Data for UD POS

3.3.6 Challenges/Observations

Wrong diacritics

Some words were wrongly typed (wrong diacritics). Most of these words were tagged null by the POS voting.

Function words with multiple POS

Some words in Yorùbá have multiple part of speech. Function words such as *kà*, *kí*, *la*, *lé*, e.tc. For example, the word *kà* can function as an adverb and as a verb. These ambiguous function words were assigned the most probable tag(tag with the highest votes). Another method would have been to assign these tags based on context but the behaviour of these function words are not clear in a lot of instances.

4. Dependency Parsing

4.1 Manual annotation of Dependency relations

As part of this work, 100 sentences from the bible were annotated using the Universal Dependencies version 2 guidelines Nivre et al. [2016] [1].
Although this is not a full-fledged annotation work because a full-fledged annotation work would require hiring of additional annotators and computing inter-annotator agreement, still, it is an important contribution and currently the only available 'treebank' of Yorùbá.

Table 4.1 shows the number of sentences, tokens, word types and average sentence length. Table 4.2 shows the statistics of the annotated Part of Speech. Table 4.3 shows the statistics of the annotated Dependency relations.
The dependencies were annotated using Arborator, a manual dependency tool which supports editing of POS tags and dependency relations in an easy to use drag and drop interface Gerdes [2013]. This treebank has been released in Universal Dependencies data release version 2.2.

One major challenge encountered in the annotation is the function words, function words in Yorùbá could have multiple parts-of-speech and at times it is difficult to assign them to a particular part-of-speech.

Sentences	100
Average Sentence Length	27
Word Types	453
Tokens	2688

Table 4.1: Annotated data statistics

[1] `http://universaldependencies.org/`

4.1.1 Part of Speech

POS	Number of tokens
PRON	473
PUNCT	451
VERB	408
NOUN	344
AUX	190
ADP	181
SCONJ	150
CCONJ	150
ADV	100
PROPN	72
DET	69
ADJ	50
PART	34
NUM	11
INTJ	4
X	1

Table 4.2: Part of Speech statistics

4.1.2 Dependency Relations

Table 4.3 lists the 37 universal syntactic relations used in UD v2. It is a revised version of the relations originally described in Universal Stanford Dependencies: A cross-linguistic typology de Marneffe et al. [2014]. Universal Stanford Dependencies came first, its revised version was UD v1 Nivre et al. [2016] and the second revised version is currently UD v2 [2].

Table 4.3 also shows the UD Dependency relations statistics based on the manually annotated corpus.

Nouns and Pronouns

Figure 4.1 illustrates an example of a 'nsubj' dependency relation in Yorùbá where the child is a noun, Yorùbá sentence structure follows majorly the 'SVO' format.

[2]http://universaldependencies.org/

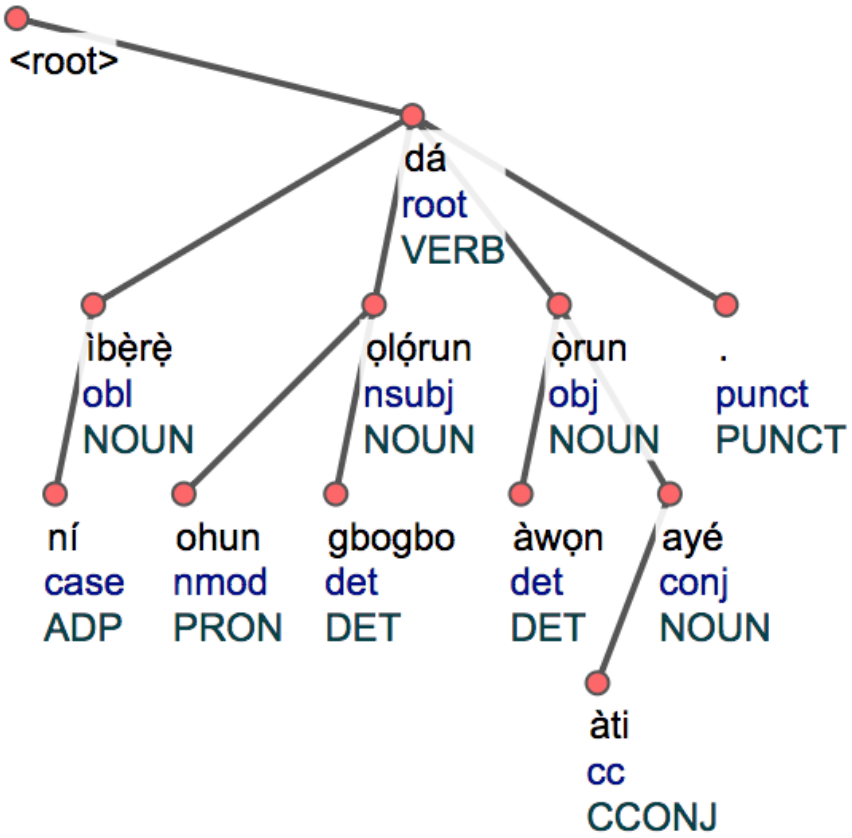

Figure 4.1: Example of a nsubj dependency relation where the child is a noun **English Translation**: *In the beginning god created the heaven and the earth .*

Auxiliaries and Verbs

In Yorùbá, the same word can function as a verb or an auxiliary depending on context. For instance, the word 'wà' can act as a verb or an auxiliary in a sentence. When it functions as a verb, it can mean the following: 'to exist', 'to dwell/abide', 'to dig', 'to pull a boat'. These other meanings won't have relationship 'aux' with their head. Figure 4.2 shows when 'wà' is a verb and functions as the root of a sentence.

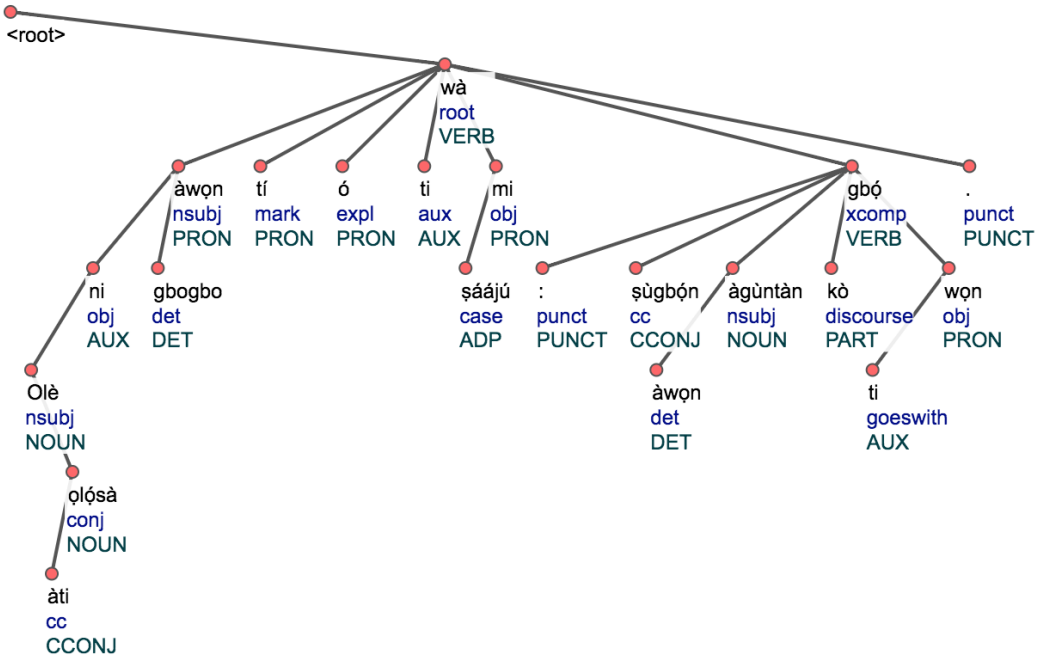

Figure 4.2: Word 'wà' functioning as a verb and root of the sentence
English Translation: *All that ever came before me are thieves and robbers: but the sheep did not hear them.*

Serial Verbs

'Compound:svc' (compound:serial verb construction) is an interesting dependency relation found in Yorùbá. Figure 4.3 illustrates this. The sentence contains two serial verbs namely **wọ̀** and **pàpọ̀**.
Wọ̀ - to put on, to enter, to set (as the sun)
Pàpọ̀ - to join, to mingle together, to unite
Wọ̀ pàpọ̀ - to gather together
As discussed in Chapter 1, we have different categories of serial verbs in Yorùbá. **Wọ̀ pàpọ̀** falls under the category where both words act as a verb but the meaning of the verb is based on just one of the two verbs, in this case, the meaning is based on **pàpọ̀**.

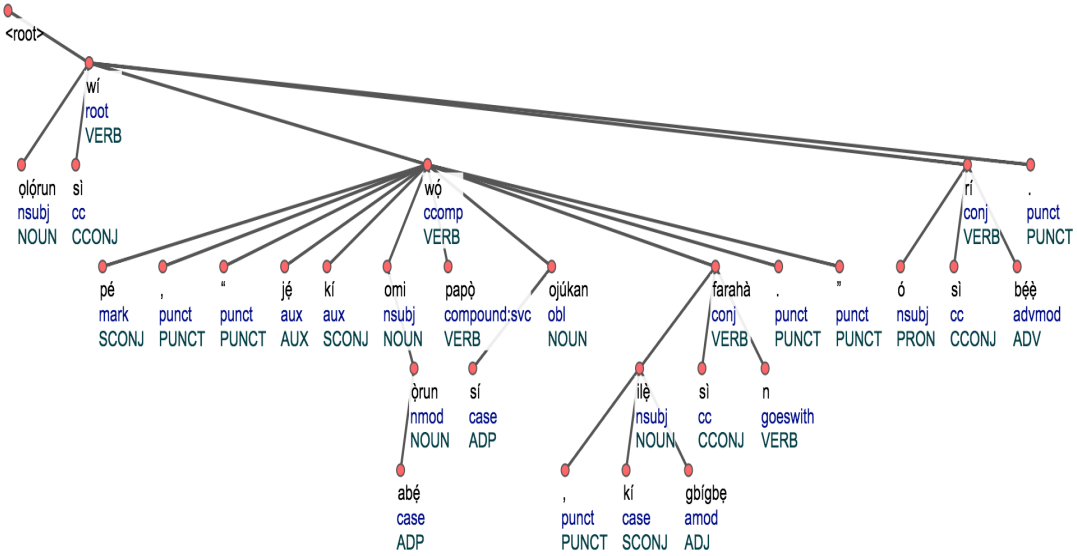

Figure 4.3: Example of ‘compound:svc’
English Translation: *And god said , let the waters under the heaven be gathered together unto one place , and let the dry land appear : and it was so .*

Determiners

Determiners in Yorùbá behave in a similar way as English. Also, some pronouns in Yorùbá function as determiner, example of such is the pronoun ‘àwọn’. Figure 4.4 shows it behaviour as a pronoun and figure 4.5 shows its behaviour as a determiner.

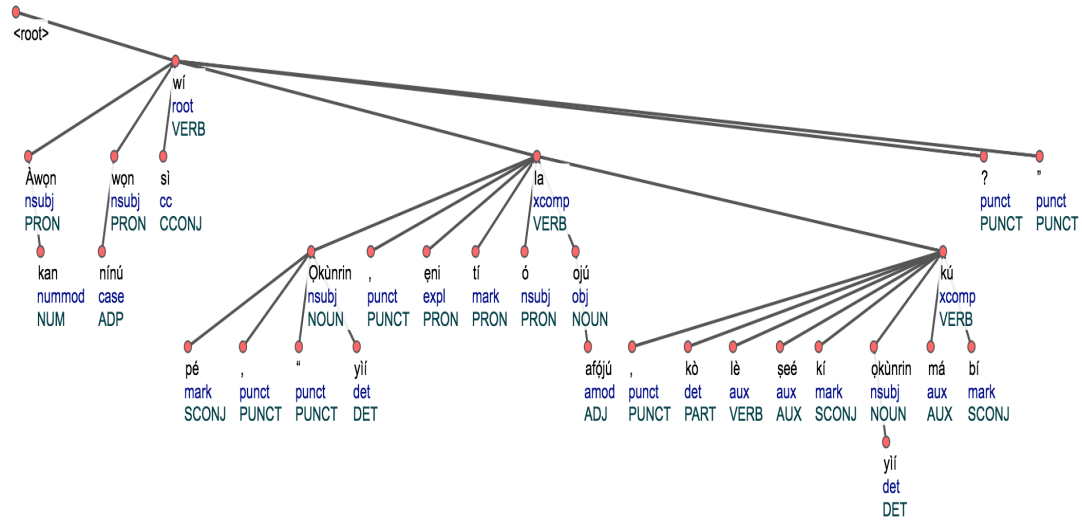

Figure 4.4: Example of ‘àwọn’ as a pronoun
English Translation: *And some of them said, Could not this man, which opened the eyes of the blind, have caused that even this man should not have died?*

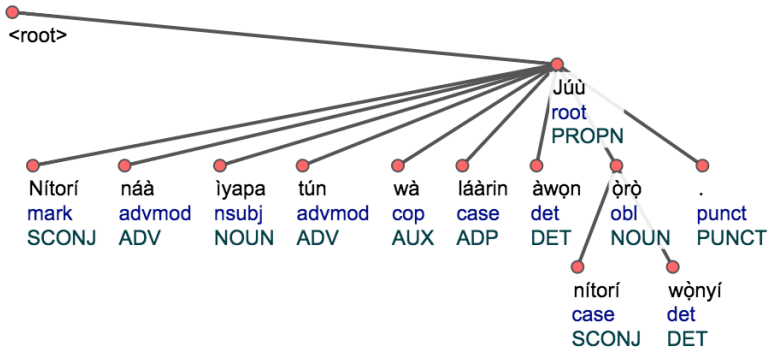

Figure 4.5: Example of a 'àwọn' as a determiner
English Translation: *There was a division therefore again among the Jews for these sayings.*

Dependency relation	Dependency relation	Number of tokens
acl	adjectival clause	11
advcl	adverbial clause modifier	8
advmod	adverbial modifier	87
amod	adjectival modifier	41
appos	appositional modifier	0
aux	auxiliary	160
case	case marking	168
cc	coordinating conjunction	156
ccomp	clausal complement	38
clf	classifier	0
compound	compound	19
conj	conjunct	71
cop	copula	43
csubj	clausal subject	31
dep	unspecified dependency	0
det	determiner	77
discourse	discourse element	28
dislocated	dislocated elements	0
expl	expletive	31
fixed	fixed multiword expression	0
flat	flat multiword expression	0
goeswith	goes with	36
iobj	indirect object	10
list	list	1
mark	marker	212
nmod	nominal modifier	101
nsubj	nominal subject	362
nummod	numeric modifier	10
obj	object	213
obl	oblique nominal	42
orphan	orphan	0
parataxis	parataxis	1
punct	punctuation	436
reparandum	overridden disfluency	0
root	root	100
vocative	vocative	0
xcomp	open clausal complement	162
compound:svc	compound for serial verbs	30
compound:prt	compound for particle verbs	2

Table 4.3: UD Dependency relations and Statistics

4.2 Projecting Dependencies

We are interested in projecting dependency relations from a resource-rich language such as English to a low-resource language Yorùbá. The idea here is based on Hwa et al. [2005]. The main idea is to annotate the English parallel corpus using UDPipe, project the analysis to Yorùbá, and then train a model using UDPipe on the resulting annotations after which we test on manually annotated data. Yorùbá sentences and the equivalent sentences in English were word-aligned using Fast Align Dyer et al. [2013].

We chose English because it is a low-morphology language and there exists plenty of parallel data between English and Yorùbá.

Given sentence pair *(E, F)* and a set of syntactic relations for E, where $E = e_1,....e_n$ is an English sentence and $F= f_1,...,f_m$ Yorùbá equivalent, dependency relations (denoted as *R(x, y)*) are projected from English for the following situation (Hwa et al. [2005])

- **one-to-one** if e_i is aligned with a unique f_x and e_j is aligned with a unique f_y, if $R(e_i, e_j)$, conclude $R(f_x, f_y)$.
- **unaligned (English)** if e_j is not aligned with any word in F, then create a new empty word f_y such that for any e_i aligned with a unique f_x, $R(e_i, e_j) \Rightarrow R(f_x, f_y)$
- **one-to-many** if e_i is aligned with $f_x,...,f_y$, then create a new empty word f_z such that f_z is the parent of $f_x,...,f_y$ and set e_i to align to f_z instead. This is called a *Multiply-Aligned Component, or (MAC).*
- **many to one** if $e_i,...,e_j$ are all uniquely aligned to f_x, then delete all alignments between $e_k (i \leq k \leq j)$ and f_x except for the head of $e_i,...,e_j$.
- **many-to-many** decomposed into a two-step process: first perform one-to-many, then perform many-to-one
- **unaligned foreign** leave them out of the projected tree

The projection architecture is shown in Figure 4.6 Hwa et al. [2005].
Table 4.4 shows an equivalent Yorùbá and English sentence. Figure 4.7 shows an alignment between UDPipe parsed English sentence and its Yorùbá equivalent. Figure 4.8 shows the resulting projected dependency tree when the Direct Projection Algorithm is applied to the English-Yorùbá sentence in table 4.4. From figure 4.7, we can see that not all the Yorùbá words are aligned, some words are left out.

From figure 4.8 we can see that the Direct Projection Algorithm captures the main dependency relations from English to Yorùbá despite the fact that they are two different languages. Despite the fact that main dependency relations were captured, some errors were encountered in the projection. Firstly, the new dependency tree assumes the English order. This is a problem because the structure of sentences in English and Yorùbá are not exactly the same. In Yorùbá, possessive pronouns come after a noun whereas it is the opposite in English as seen in figure 4.10 .

Also, due to the really low morphology nature of Yorùbá, some function words are left unaligned.
Table 4.5 shows another equivalent Yorùbá and English sentence. Figure 4.7 shows the alignment between UDPipe parsed English sentence and Yorùbá from table 4.5. Figure 4.10 shows the resulting projected dependency tree when the of Direct Projection Algorithm is applied to table 4.5.

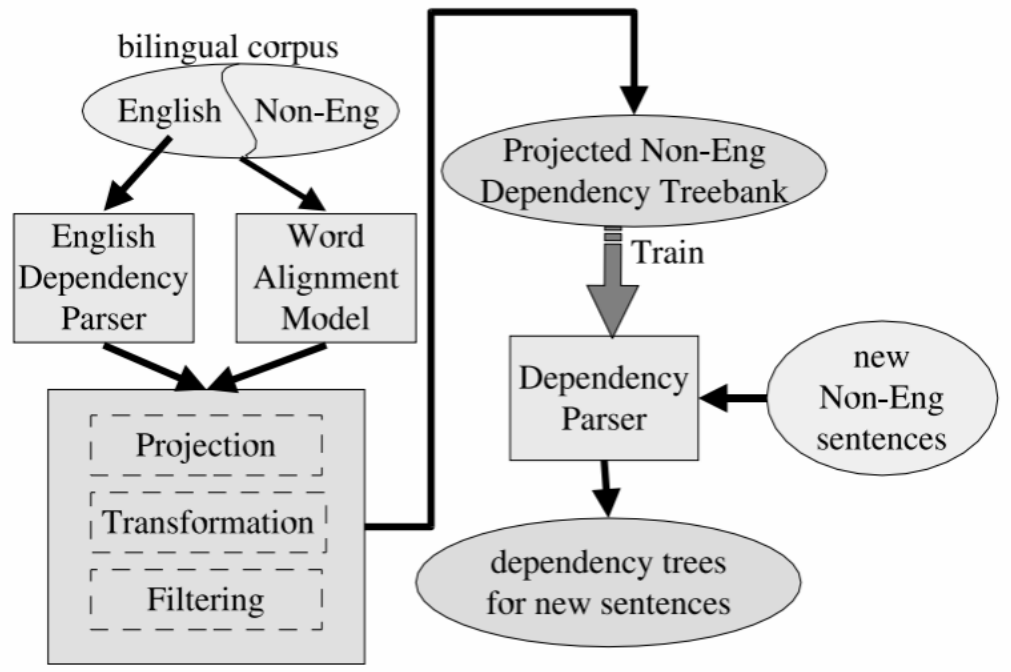

Figure 4.6: Projection architecture

Yorùbá	ìrètí ìgbàlà lè jẹ́ kéèyàn mọ̀kàn le kódà nígbà tíná bá jó dóríi kókó pàápàá.
English	the hope of being saved can help a person to hold on even in the direst of circumstances.

Table 4.4: Yorùbá and English sentence

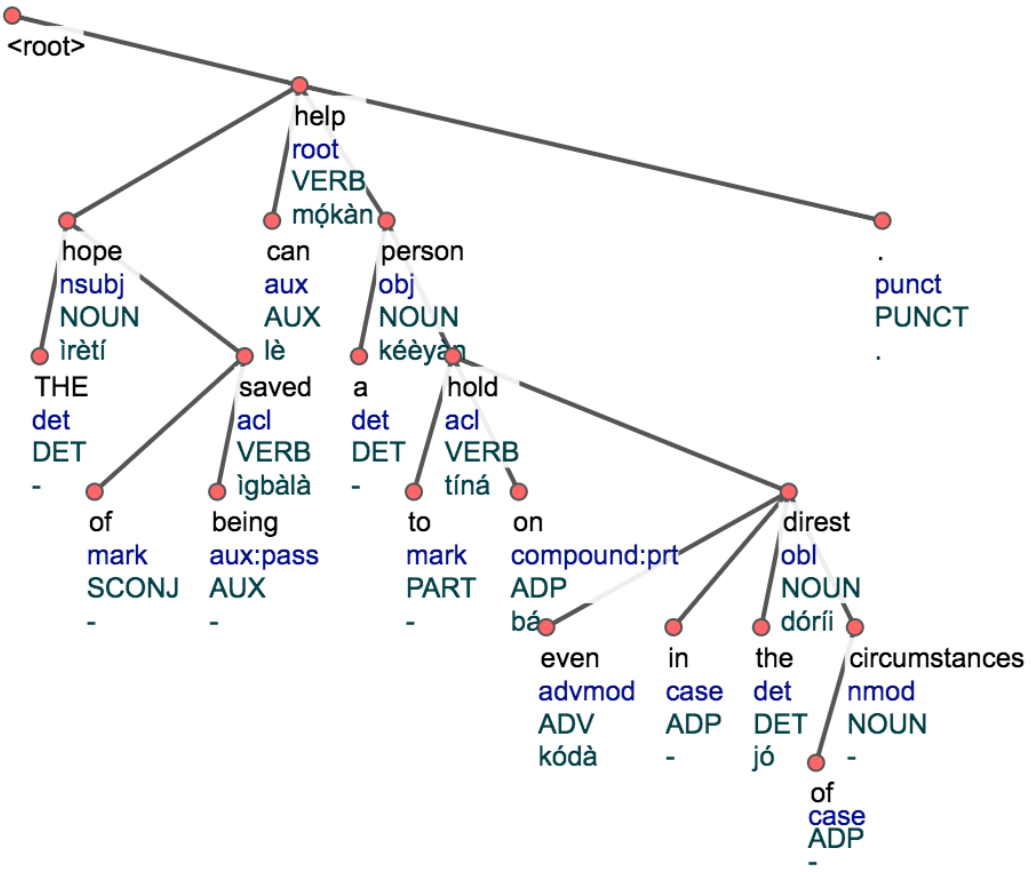

Figure 4.7: Example alignment between Yorùbá and English sentence.
The English tree is not a manual annotation but output of the UDPipe parser.
Analyses visualized using CoNLL-U file viewer

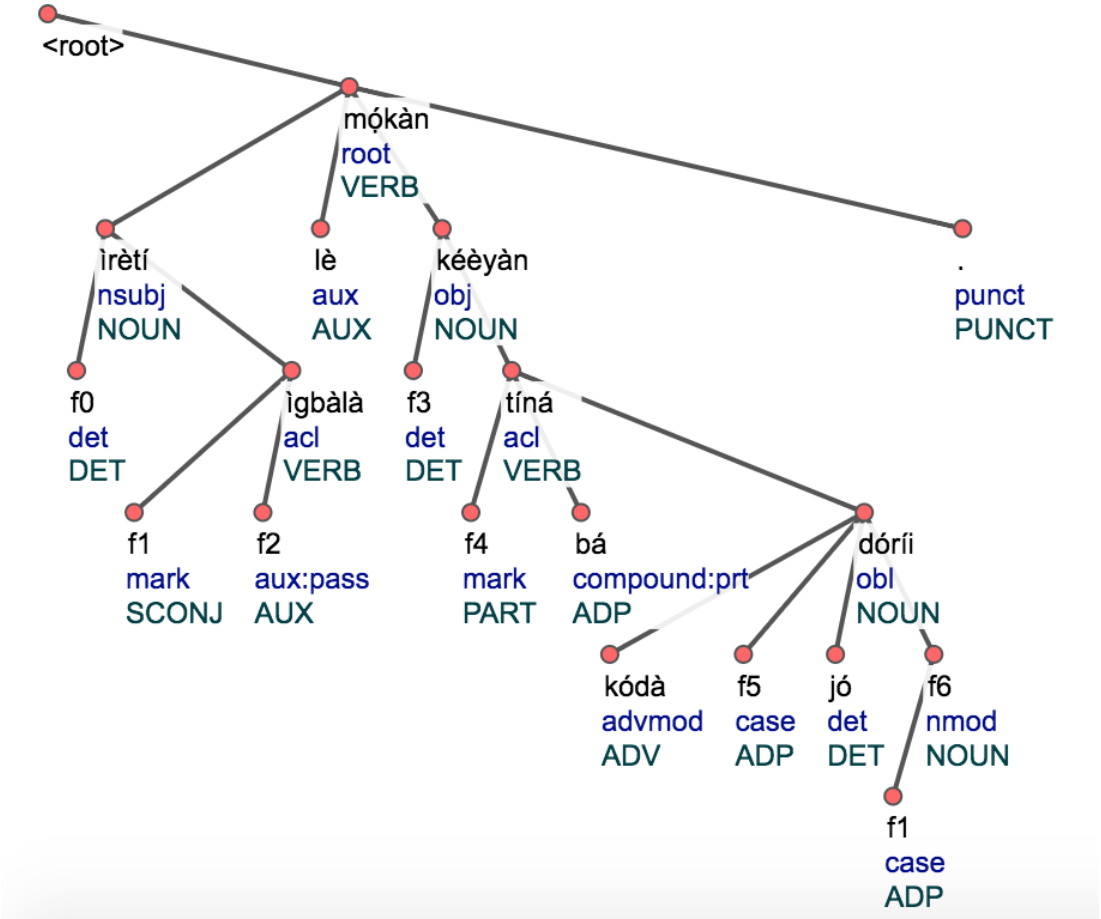

Figure 4.8: Result of the dependency projection algorithm
Analyses visualized using CoNLL-U file viewer

Yorùbá	jésù ni ẹni tí ẹbọ rẹ̀ mú àwọn àpẹẹrẹ alásọtẹ́lẹ̀ wọ́nyẹn ṣẹ.
English	Jesus was the one whose sacrifice fulfilled those prophetic pictures.

Table 4.5: Yorùbá and English sentence

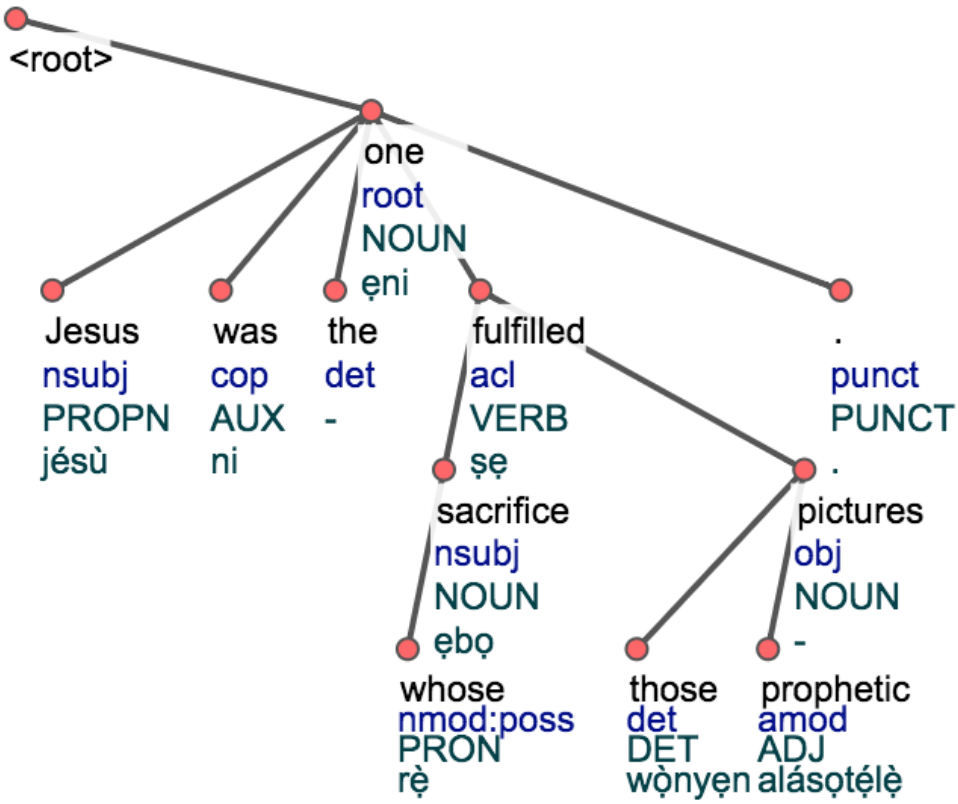

Figure 4.9: Example alignment between Yorùbá and English sentence.
The English tree is not a manual annotation but output of the UDPipe parser.
Analyses visualized using CoNLL-U file viewer

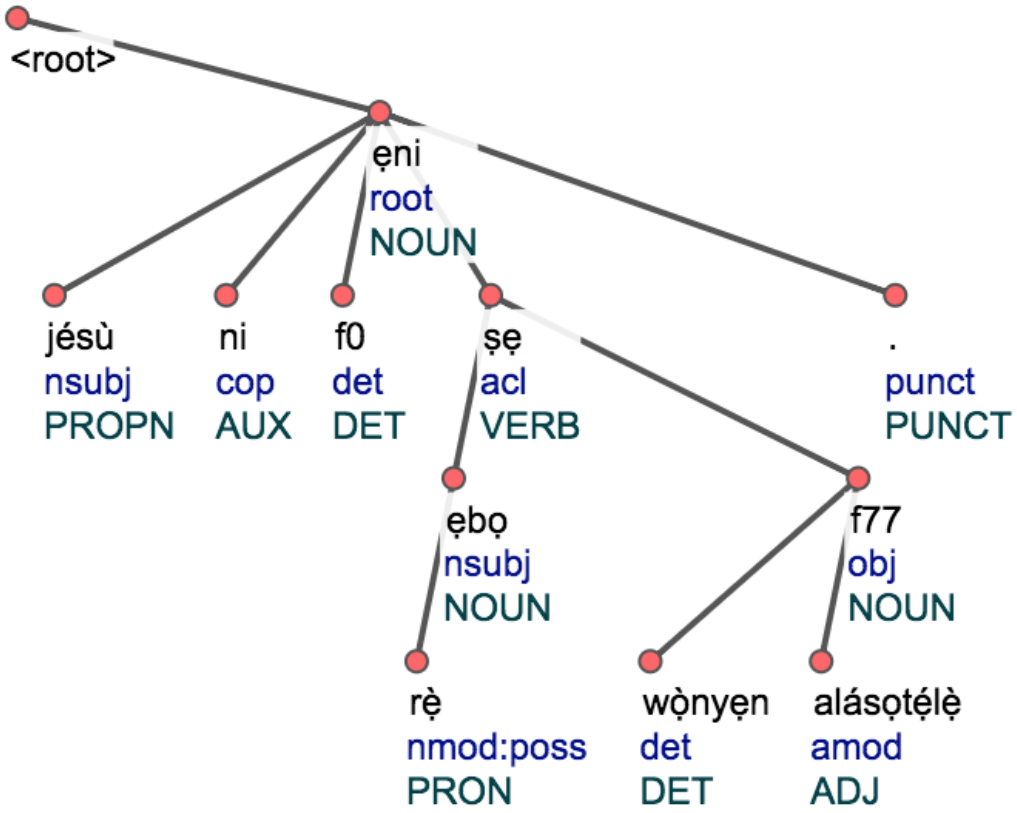

Figure 4.10: Result of the dependency projection algorithm
Analyses visualized using CoNLL-U file viewer

Dependencies were projected between 113717 sentences. Table 4.6 shows the statistics of the train data. Table 4.8 shows the training parameters. These parameters are shown here to facilitate reproducibility of the results.
The dependency parsing was evaluated using Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS), the two most common methods for evaluating the accuracy of dependency parsers. Labeled attachment refers to the proper assignment of a word to its head along with the correct dependency relation. Unlabeled attachment simply looks at the correctness of the assigned head, ignoring the dependency relation.
Table 4.9 shows the result on manually annotated data. UDPipe version 1.2.0 was used for training and testing the model.

Sentences	113717
Word Types	34992
Tokens	1944773

Table 4.6: Train data

Sentences	100
Word Types	453
Tokens	2688

Table 4.7: Test data (Gold corpus)

system	projective
oracle	dynamic
structured interval	8
single root	1
embedding_upostag	20
embedding_xpostag	0
embedding_lemma	0
embedding_deprel	20
iterations	10
hidden layer	40
batch size	10
learning rate	0.0200
learning rate final	0.0010
l2	0.5000
early stopping	0

Table 4.8: Parameters for Parser training using UDPipe version 1.2.0

LAS %	**UAS %**
35.53	50.60

Table 4.9: Results of parser on Manually annotated data (100 sentences)

5. Evaluation

In Chapter 3, we presented how we automatically used word-aligned bilingual corpora to project part-of-speech annotations from English, French, Vietnamese, Ancient Greek, Old Church Slavic and Gothic to Yorùbá, we manually annotated 100 sentences, 2688 tokens with universal dependency relations in Chapter 4. In this Chapter, we are going to train a parser on half of the manually annotated sentences and see how it performs. We are interested in seeing if it performs better than the treebank projection in Chapter 4 even though the training data is extremely small.

Also, the parser will be trained and tested using cross validation in order to get a more accurate estimate of our model's prediction performance.

5.1 Training Dependency Parser using UDPipe

The manually annotated dataset in Chapter 3 was trained using UDPipe version 1.2.0. The parsing is performed using Parsito, which is a transition-based parser using a neural-network classifier Straka et al. [2015]. Three transition systems were explored for parsing, the projective stack-based arc standard system with shift, left_arc and right_arc transitions, the swap which is a fully non-projective system which extends projective system by adding the swap transition and link2 which is a partially non-projective system which extends projective system by adding left_arc2 and right_arc2 transitions.

5.1.1 50:50 Train and Test data

The 100 sentences were split into test and train data using a 50:50 ratio and trained using UDPipe Straka and Straková [2017]. Table 5.1 and 5.2 shows the statistics of the test and train data.
Table 5.4 shows the parameters for training the Parser. Table 5.5 shows the result of the parser (UAS and LAS accuracies) and tagger (Universal Part-of-Speech Tag (UPOSTAG) accuracy) with the modified parameters. The best result was gotten using the projective transition system.

Sentences	50
Word Types	325
Tokens	1361

Table 5.1: 50% Test data

Sentences	50
Word Types	317
Tokens	1327

Table 5.2: 50% Train data

iterations	25
early stopping	0
suffix rules	8
prefix max	4

Table 5.3: Parameters for Tagger training

system	projective
oracle	dynamic
structured interval	8
single root	1
embedding_upostag	20
embedding_xpostag	0
embedding_lemma	0
embedding_deprel	20
iterations	10
hidden layer	200
batch size	10
learning rate	0.0200
learning rate final	0.0010
l2	0.5000
early stopping	0

Table 5.4: Parameters for Parser training

Transition system	**LAS %**	**UAS %**	**UPOSTAG%**
projective	**50**	**58.34**	85.23
swap	47.32	55.25	85.23
link2	46.51	53.93	85.23

Table 5.5: Results of tagger and parser on test data using 50% train and 50% test data

UPOSTAG%	**UAS %**	**LAS %**
74.80	49.82	34.75

Table 5.6: Results of Projected Tagger and Parser on 50% test data

5.1.2 Training and Testing using Cross Validation

Using cross validation gives a more accurate estimate of our model's prediction performance compared to dividing the test and train data into a 50:50 ratio.

The 100 sentences were split into ten folds of test and train data using a 90:10 ratio and trained using UDPipe Straka and Straková [2017]. After this, the results were averaged. It was trained with the parameters in Table 5.4.

As expected, the result gotten from cross validation is better than that gotten using 50:50 train and test data.

Table 5.7 shows the result of the parser and tagger on manually annotated data using cross validation. Table 5.8 shows the comparison between results gotten from training a model on projected dependencies, projected part-of-speech tags and training on manually annotated data. The best result was gotten by the model trained on manually annotated data.

Transition system	**AverageLAS %**	**Average UAS %**	**Average UPOSTAG%**
projective	**58.1**	**68**	89.56
swap	54.60	64.18	89.56
link2	55.26	65	89.56

Table 5.7: Results of parser and tagger on test data using cross-validation

	LAS %	**UAS %**	**UPOSTAG%**
Model trained on manual annotation and tested using Cross Validation	**58.1**	**68**	**89.56**
Model trained on manual annotation using 50% train and 50% test	50	58.34	85.23
Parser Model trained on projected dependencies and tested on manually annotated data (100 sentences)	35.53	50.60	-
Tagger Model trained on projected part-of-speech tags and tested on manually annotated data (100 sentences)	-	-	76.71
Model trained on manually annotated sentences (100 sentences) and tested on UDPipe test data	-	-	72.76

Table 5.8: Model trained on manual annotation vs projected dependencies

5.1.3 Analysis

From table 5.8, the parser model trained on projected dependencies and tested on 100 manually annotated sentences obtained a low UAS and LAS because of linguistic differences between Yoruba and English, application of post transformation rules will give better results.

The model trained on projected part-of-speech annotations achieves a upostag accuracy of 76.71 when tested on the 100 manually annotated sentences. This is low compared to accuracy of 89.56 gotten via cross-validation. We analysed the output of the model trained on projected part-of-speech annotations and discovered that most of the errors were from the function words. Table 5.9 shows some of this errors. From this errors, we can see the ambiguity of these function words. For example the word ṣe can mean 'to do, to act, to cause', it can act as an

auxiliary or a verb. The word 'ní' can mean 'to be, to have, to say, to posses, to obtain, from', it can act as an auxiliary, verb or adposition. To get a better accuracy, language specific rules will have to be added to the output of the projected annotations.

Word	**Gold tag**	**Predicted tag**
sì (and, to)	CCONJ	ADP
lè (can)	AUX	VERB
kuro (from, leave)	ADP	VERB
bá (with, should)	SCONJ	AUX
máa (will)	AUX	VERB
ń (-)	AUX	VERB
ní (have, to)	AUX	ADP
ṣe (to do, to act, to make)	AUX	VERB

Table 5.9: Analysis of some errors from 100 manually annotated sentences tested on Projected Tagger trained with UDPipe 1.2.0

Conclusion

This thesis has shown that a part of speech tagger and parser can be built for Yorùbá, a low resource language using Parallel data available in English, French, Vietnamese, Ancient Greek, Gothic, Latin and Old Church Slavonic via exploring parts-of-speech annotation projection by Yarowsky and Ngai [2001] and syntactic relations projection by Hwa et al. [2005].

By using languages such as Ancient Greek, Gothic, Latin and Old Church Slavonic we were not able to build a robust part-of-speech tagger because these languages are significantly different from Yorùbá and a lot of Yorùbá words were left unaligned.
This is due to Yorùbá being an analytical language, with almost no morphology and a lot of function words while the other four languages are all morphologically rich. Thus the function words in Yorùbá correspond to morphological features in the other languages, and there are no independent words that could be aligned with the Yorùbá function words.

By adding resource rich languages such as English, French and Vietnamese we are able to build cover more word types, tokens, and parts-of-speech. We built a robust part-of-speech tagger and trained a tagger using UDPipe 1.2.0 and obtained an accuracy of 87.86% on our test data.

As part of this thesis, 100 Yorùbá sentences were manually annotated using the Universal Dependencies (UD) annotation format. An accuracy of 89.56% was gotten when we trained our part-of-speech tagger on this manually annotated corpus using cross validation.

We also explored the method of syntactic relations projection developed by Hwa et al. [2005]. The parser trained on dependencies projection from English using the direct projection algorithm yields a UAS 50.60% of and LAS of 35.53% when tested on manually annotated data.
However, the parser trained on manually annotation yields UAS of 68% and LAS of 58.1%. In order to obtain better results using the dependency projection algorithm, post transformation rules must added to the output of the depenency projection algorithm since English and Yorùbá are significantly different languages.

We have presented Universal Dependencies (UD) for Yorùbá following the Universal Dependencies(UD) annotation fomat. The Universal Dependencies (UD) for Yorùbá contains 100 sentences and 2688 tokens. The treebank is made freely available in the Universal Dependencies version 2.2 repository.

This treebank will aid the development of part-of-speech taggers or treebanks for other low-resource African languages in the Niger-Congo family that have parallel data with Yorùbá.

Bibliography

Stephen Monday Adewole. *Yoruba Word Formation Processes*. PhD thesis, University of California at Los Angeles, 1995.

Željko Agić. Cross-lingual parser selection for low-resource languages. In *UDW@NoDaLiDa*, 2017.

Željko Agić, Dirk Hovy, and Anders Søgaard. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272. Association for Computational Linguistics, 2015. URL `http://www.aclweb.org/anthology/P15-2044`.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. Multilingual projection for parsing truly low-resource languages. *TACL*, 4:301–312, 2016. URL `https://transacl.org/ojs/index.php/tacl/article/view/869`.

Željko Agić, Barbara Plank, and Anders Søgaard. Cross-lingual tagger evaluation without test data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 248–253. Association for Computational Linguistics, 2017. URL `http://aclweb.org/anthology/E17-2040`.

Oladele Awobuluyi. *Essentials of Yoruba Grammar*. Oxford University Press, Ibadan, Nigeria, 1978. ISBN 0-19-575300-3.

Yiwola Awoyale. Nominal compound formation in yoruba ideophones. *Journal of African Languages and Linguistics*, 3:139–157, 1981.

Ayo Bamgbose. *A Grammar of Yoruba*. Cambridge University Press, UK, 1966.

Ayo Bamgbose. Deprived, endangered, and dying languages. *Diogenes*, 41(161): 19–25, 1993.

Ayo Bamgbose. *Fonoloji ati Girama Yoruba*. University Press PLC, Ibadan, 2010.

Ezra Black, Fred Jelinek, John Lafferty, David M. Magerman, Robert Mercer, and Salim Roukos. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 31–37, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. doi: 10.3115/981574.981579. URL `https://doi.org/10.3115/981574.981579`.

Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395, June 2015. ISSN 1574-020X. doi: 10.1007/s10579-014-9287-y. URL `http://dx.doi.org/10.1007/s10579-014-9287-y`.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 600–609, 2011.

Hal Daumé and Daniel Marcu. Learning as search optimization: Approximate large margin methods for structured prediction. *CoRR*, abs/0907.0809, 2009. URL `http://arxiv.org/abs/0907.0809`.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, 2014.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. Simpler unsupervised pos tagging with bilingual projections, 08 2013.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*, 2013.

A.Y. Folarin. *Lexical Phonology of Yoruba Nouns and Verbs*. University of Kansas, 1989. URL `https://books.google.co.uk/books?id=y2HNngEACAAJ`.

Victoria Fossum and Steven Abney. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP'05, pages 862–873, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-29172-5, 978-3-540-29172-5. doi: 10.1007/11562214_75. URL `http://dx.doi.org/10.1007/11562214_75`.

Dan Garrette, Jason Mielens, and Jason Baldridge. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *ACL*, 2013.

Kim Gerdes. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97, Prague, 2013. URL `https://arborator.ilpga.fr`.

Rebecca Hwa, Phillip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel texts. In *Natural language engineering*, page 11(3):311–325, 2005.

Shen Li, João V. Graça, and Ben Taskar. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1389–1398, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390948.2391106`.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair),

Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. A universal part-of-speech tagset. *CoRR*, abs/1104.2086, 2011. URL `http://arxiv.org/abs/1104.2086`.

Milan Straka and Jana Straková. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/K/K17/K17-3009.pdf`.

Milan Straka, Jan Hajič, Jana Straková, and Hajič Jan jr. Parsing. Universal dependency treebanks using neural networks and search-based oracle. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories TLT 14*, 2015.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P14/P14-5003.pdf`.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. Cross-lingual part-of-speech tagging through ambiguous learning. In *EMNLP*, 2014.

David Yarowsky and Grace Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*, pages 377–404, 2001.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, pages 109–116, 2001.

Daniel Zeman. Reusable tagset conversion using tagset drivers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In *IJCNLP*, 2008.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. Unsupervised and Lightly Supervised Part-of-Speech Tagging Using Recurrent Neural Networks. In *29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Shangai, China, October 2015. URL `https://hal.archives-ouvertes.fr/hal-01350113`.

List of Figures

List of Tables

List of Abbreviations

HMM Hidden Markov Model.

LAS Labelled Attachment Score.

POS Part-of-Speech.

UAS Unlabelled Attachment Score.

UD Universal Dependencies.

UPOSTAG Universal Part-of-Speech Tag.