The State of Natural Language Processing in Africa: An Overview

Article · April 2024

DOI: 10.15680/IJIRCCE.2024.1204002

2 authors:

Ikechukwu Onyenwe
Nnamdi Azikiwe University, Awka
49 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Onyedinma Ebele Gr
Nnamdi Azikiwe University, Awka
15 PUBLICATIONS **23** CITATIONS
SEE PROFILE

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

The State of Natural Language Processing in Africa: An Overview

[1]Onyenwe, I.E., [2]Onyedinma, E.G.

[1, 2] Department of Computer Science, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

ABSTRACT: Despite the large number of languages and native speakers in Africa, it is still an NLP-dark continent; the Language Resources and Evaluation (LRE) map[1] shows that while English has a good number of computational and corpora tools (663 reported in Corpus linguistics [2]), African languages have relatively few. This review was done before 2013.

KEYWORDS: Natural Language Processing (NLP), Part of Speech Tagging (POS), Igbo, Corpus, Corpora.

I. INTRODUCTION

There has been a growing interest in NLP in Africa [3,4,5, 6]. TALAf4 (Traitement Automatique des Langues Africaines5 (text and speech)) is a workshop (held at the JEP-TALN-RECITAL conference, 2016) with the aim of bringing together researchers in the NLP field working on African Indigenous Languages6 (AIL) through: meetings at the workshop; extracting knowledge using open source tools, standards (ISO, Unicode), and publishing the tools developed with an open license to avoid losses when a project stops and cannot be reopened for lack of resources; developing a set of best practices based on the researchers' acquaintances; setting up simple and effective methodologies based on free, or almost free, software for the development of tools; communicating methods that can eschew the use of non-existent tools; and refraining from loss of time and energy. AFLAT7 is an African Language Technology body interested in language technology research for AIL, aiming to catalogue resources (such as corpora, dictionaries, and NLP tools) for the majority of resource-scarce AIL (both current and extinct) for the benefit of researchers interested in African language technology. AILs are linguistically rich and have high divergence in typology [7], although some bear little relation to one another. The typological difference could be the effect of many ethnicities in Africa. There are four language classes in Africa: Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Khoisan [8]. The Niger-Congo family is a large language phylum that contains approximately 1500 of the languages of Africa [9, 10]. The Igbo language is among the 40-60 languages in the Kwa sub-group of the Niger-Congo family. Common shared features among the languages of Africa are in their phonology, syntax, morphology and lexicon [8], implying that NLP methods, experiments, and experiences obtained for one language could be extended to others.

There are major challenges facing under-resourced languages; most outstanding are the text to be tagged, orthography of the text, tokenization (dealing with morphology since affixations are prevalent features in most AILs), and the size of tagset. In POS tagset development and tagging for the Niger-Congo family, the first and most common issue is in their morphology and orthography-languages in this family are morphologically agglutinative in structure [11, 12]; words are formed by concatenation of morphemes that would syntactically stand as a lexical unit in other non-AIL languages, and this is mostly the case for noun and verb classes. See tables 1 and 2. POS tagset design and tagging in the above case is a non-trivial task. Firstly, which units should be classified as tokens, since words in these language types are highly inflected with morphemes, and tokenizing purely on whitespace would be linguistically misrepresenting. Additionally, it is difficult to define the boundaries of these morphemes in inflected words8 , and choosing the tagging types for each. It is also difficult to determine the best level of morphology decomposition, since these languages are so morphologically rich, and the order of occurrence of morphemes is not fixed (abiaghikwa and abiakwaghi are valid words with the same sense in Igbo**)**. Analysing and determining the best tokenization algorithm for morphologically complex languages is a lengthy and laborious process- ZulMorph, the UNISA prototype morphological analyser for Zulu, took a decade to develop [11].

Another major problem that is very common in AILs is multiword units. They may appear as separate tokens with the same tag, or combined tokens with different tags. For example, in Igbo noun classes, agentive nouns are formed through the nominalization of verbs (eg. o.gu. egwu 'singer'), and instrumental nouns are words used to refer to, or describe, instruments, which are also formed through the nominalisation of verbs (eg. ngwu ji 'digger'). In Wolof (Senegalese language), [5] found that the pronominals or focus markers and associated inflection often appear as separated words in

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

the Wolof text. A different case of the above is found in Northern Sotho, where a single word form will receive more than one tag [13]. Another prominent feature in AIL is the use of ideophones or words that evoke vivid sensations.

AIL are classified as under-resourced languages because of lacking the linguistic resources such as electronic texts, word lists, dictionaries, grammars, spell-checkers, etc. [14]. Though some languages have one or more linguistics materials available that can help kick-off natural language processing (NLP) research. But lack of the processed linguistic items is the major cause of NLP stagnant growth in African language technology and reason can be attributed to the social and political past of Africa that did not promote the use of native languages in education and commerce, until recently [15].

Class	Northern Sotho		Zulu		Igbo	
	Plural marker	Example	Plural marker	Example	Plural marker	Example
sg	mo–	*motho* 'person'	umu–	*umuntu* 'person'	nwa	*nwa mmadu* or *mmadu* 'person'
pl	ba–	*batho* 'persons'	aba–	*abantu* 'persons'	*umu*	*umu mmadu* 'persons'

Table 1: Noun classes for Northen Sotho, Zulu and Igbo. Part of table taken from [16]

	"they do not sell"	Negative morpheme	Subject	Verb root	Suffix	Prefix	Inflectional ending
Northern Sotho	ga ba rekiše	ga	ba	rek–	–iš–	–	–e
Zulu	abathengisi	a–	–ba–	–theng–	–is–	–	–i
Igbo	ha anaghi ere	–ghi	ha	–na	–ghi	a–	–

Table 2: Verbal morphology Northen Sotho, Zulu and Igbo. Part of table taken from [16]

In AIL, a verb may comprise subject, concord, a verb stem (bears the basic meaning) and inflectional ending [16]. Morphemes prefixed to the verb root may include lexical class such as object concords, potential and progressive, negative, and participle morphemes. There might be derivational or extensional suffixes appearing between a verb stem and inflectional ending as in table 2.

According to [16], all verbal derivatives can be blindly tagged as verbs or can be morphologically analysed. If the latter, then tagging will be based on the verbal suffix's lexical functions. Morphological ambiguity is resolved using contextual information. For example, verbal derivation like that one in table 3 can be blindly tagged as verbs, or alternatively, first be morphologically analysed and then tagged the verbal suffixes based on their lexical functions. Compare table 3 with table 4. The tables 3 and 4 are illustrative excerpts from [16] of Northern Sotho morphological analysis in their tagset designs.

Module Composition
root + reciprocal + standard modifications

Module Composition	Abbreviations morpheme	Stems and Derivations
root + reciprocal + standard modifications	VRRec	rekana
	VRRecPer	rekane
	VRRecPas	rekanwa
	VRRecPerPas	rekanwe

Table 3: Derivations of the verb reka

rekana 'V'	*rek*– 'Vroot' –*an*– 'Rec' –*a*
rekane 'V'	*rek*– 'Vroot' –*an*– 'Rec' –*e* 'Per'
rekanwa 'V'	*rek*– 'Vroot' –*an*– 'Rec' –*w*– 'Pas' –*a*
rekanwe 'V'	*rek*– 'Vroot' –*an*– 'Rec' –*w*– 'Pas' –*e* 'Per'

Table 4: Alternative in POS tagging of the verb reka

II. LOW-RESOURCED LANGUAGES CORPORA AND TAGSETS

This section reviews low-resourced languages tagsets and corpora that have been developed for African Indigenous languages9 (AIL) and non-African. Finally, we discuss English since it is one of the most spoken languages of the world and some African countries use it as their official language (e.g. Nigeria).

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

	Swahili	Northern Sotho	Zulu	Cilubà
Number of sentences	152,877	9,214	3,026	422
Number of tokens	3,293,955	72,206	21,416	5,805
POS Tagset size	71	64	16	40
% of ambiguous words	22.41	45.27	1.50	6.70
Average% of unknown words	3.20	7.50	28.63	26.93

Table 5: Corpus and tagset information for Swahili, Northern Sotho, Zulu and Cilubà. The percentage ratios are computed on 10-fold cross validation. Unknown words are previously unseen words in the training data. Source: Table taken from [17]

Table 5 presents four AIL corpora and tagsets statistics by [17]. The Swahili corpus is part of Helsinki Corpus of Swahili (HCS) tagged in Standard Swahili text using SALAMA10 (Swahili Language Manager) by [18]. In addition to POS tags, HCS contains other information, such as the word base form (lemma), morphology, noun class affiliation and verbal morphology. HCS consists different text styles, such as texts from Deutsche Welle newswire to represent Swahili news and excerpts from a number of textbooks (eg. prose, fiction, education, and sciences). The size of HCS is 12.5 million words11 and tagset used contains about 302 tags [19].

Northern Sotho corpus annotation by [14] contains 10000 tokens and 56 POS tags. Microsoft Excel environment was used for the annotation based on the following reasons: computer-literate users in Northern Sotho are familiar with Microsoft Office suite and POS tagging in Excel could speed up annotation. [13] designed Nothern Sotho tagset based on the lexical and morphological criteria. The structure of the tagset are into two annotation levels of EAGLES, namely; obligatory and recommended. The authors used the obligatory level to distinguish the Nothern Sotho tagset into nine different classes: concords, pronouns, nouns, adjectives, verbals, morphemes, particles, questions, and others. From the obligatory classes, a fine-grained tagset which has 141 tags was developed. Following this was additional morphosyntactic distinctions, which led to 262 different types of morphemes. There are five features considered by [13] in Nothern Sotho tagset, namely; (1) the class membership feature, which is a classification of POS tags based on different classes; (2) the personal attribute features- a classification based on first and second persons (e.g., PERS); (3) the feature set of morphemes- morphemes are classified based on their lexical functions; (4) the feature set of particles- all the possible values of particles are considered (hortative, copulative, locative, etc.). For example, in [16] work, all verb forms are tagged “V” except copulative verb “VCOP” and participle-like words are tagged each with its grammatical function; (5) a further step to indicate whether a copulative is negated, and some features (eg. locative) of the top-level tagset.

Table 6 shows various tagsets by different authors. The last row of the table, [20] disregards the morphosyntactic distinctions in the tagset of [13]

Authors	Tagset size	±Noun class	Tool?
(Van Rooy and Pretorius, 2003)	106	– noun class	No
(De Schryver and De Pauw, 2009)	56	– noun class	Yes
(Kotze, 2008)	Partial	N.R.	Yes
(Taljard et al., 2008)	141/262	+ noun class	No
(Gertrud Faaß et al., 2009)	25/141	+ noun class	yes

Table 6: Various tagsets sizes for Nothern Sotho from [20]

tokens	Tags associated to each ambiguous token	Frequency
a	CDEM6:CO6:CS1:CS6:CPOSS1:CPOSS6:PAHORT:PAQUE:PRES	2304
go	CO2psg:CO15:COLOC:CP15:CS15:CSLOC:Csindef:PALOC	2201
ka	CS1psg:PAINS:PATEMP:PALOC:POSSPRO1psg:POT	1979
le	CDEM5:CO2ppl:CO5:CS2ppl:CS5:PACON:VCOP	1690
ba	AUX:CDEM2:CO2:CS2:CPOSS2:VCOP	1509

Table 7: Most frequent and ambiguous words in the Northern Sotho corpus, taken from [16].

to reduce 141 tags to 25 top-level tags. Their aim was geared towards building a standard and structured tagset for Nothern Sotho. The high lexical ambiguity of Nothern Sotho as shown in table 5 and 7 is an evidence that languages with disjunctive writing system12 apparently possess a high level of words with more than one tags. Possible solutions used in Nothern Sotho’s multiword problems as proffered by [13] are: (1) to run tokens together with their tags without

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

intervening spaces. (2) to use portmanteau tags13 , that is, keeping the combined tokens together, accompanied by relevant tags, which could be segregated by means of some symbol or punctuation marks. (3) to separate the fused words during lexicon-based pre-tagging using a unique lexicon as a stoplist.

The POS tagged Zulu's corpus is called Ukwabelana, which came from the Zulu's fiction and Bible translation texts [21, 7]. According to [16], Zulu and Nothern-Sotho corpora were prepared in the department of African Language of the University of Pretoria, South Africa. The sources of the corpora are the newspaper reports, academic texts, and internet. And those sources that are not electronically available were OCR-scanned and hand-cleaned. The followings are the description of Ukwabelana corpus: there are about 100,000 common Zulu word types and 30,000 Zulu sentences, of which 10,000 words are morphologically tagged and 3000 sentences are POS tagged [21]. [16] also report that untagged corpora of Northern Sotho and Zulu comprise of 6.5 million tokens and 5.2 million tokens compiled by African Language Department in the University of Pretoria. Table 5 shows that about 98% of words in Zulu corpus do not need disambiguation because it is rich in morphology and has conjunctive14 writing systems.

	Verbs	Pronouns	Particles
Noun			
Adjectival	Proper	Absolute	associative
Deverbative	Auxiliary	demonstrative	instrumental
Locative	copulative	Quantitative	locative
		Possessive	Possessive
			Qulificative

Table 8: Sub-categorization of the main word classes of Tswana

[17] present the Cilubà small POS labelled corpus of 6,000 tokens (see table 5). However, a full description of the tagset used is lacking.

In Tswana, word classes are divided on the basis of similarities between certain words. The major types found in Tswana are nouns, verbs, pronouns, particles, adverbs, idiophones and interjection. Nouns and verbs are open classes on the basis of their morphological productivity while pronouns, particles, adverbs, interjections, and idiophones are in the close classes group since they are morphologically unproductive. Fine-grained categories of the above core word classes are based on the grounds of similarities between words within a specific word category [22]. In table 8, noun is sub-categorized into adjectival, deverbative and locative.

According to [5], Wolof is a well documented language better than other West Atlantic languages (Sub-family of Niger-Congo). There are two main aspects of the language's grammar; first, Wolof is rich in morphology derivation for nouns and verbs, and secondly, inflectional elements, pronouns or clitics are treated as separate tokens or as verbal suffixes. Though in the tagset design, they remained neutral regarding how to tokenize these elements, since their main goal is to design a reliable and informative tagset with respect to the syntactic function of the linguistic elements. Therefore, the internal criteria design is less important. [5] started Wolof tagset design from scratch since no previous tagset had been designed for the language. The sources used by the authors for Wolof corpus and tagset developments are the Wolof Bible, dictionaries, and grammars books. Table 9 lists Wolof different tagset sizes by [5]. Coarse-grained tagset in Wolof contains adverbs, prepositions, articles, comparatives, conjunctions, determiners, inflectional markers, nouns, pronouns, particles, verbs, reflexives, foreign language material, and punctuation. One of the difficulties encountered during the POS tagset design for verbs was its finiteness, and the possible step adopted by the authors to find a solution was to follow a particular work of a linguist who proposed three categories for verb finiteness. These categories are POS tagged in their tagset as VVFIN, VVNFN, VVINF corresponding to finite, deficiently finite and infinite verbs respectively. Also, there was an issue of multiword units. In this case, they used the standard tokenization format where tags are assigned to each token separated by lexical space at the first level. For example, 'inflectional sentence focus

Tagset Name	Detailed	Medium	General	Standard
Tagset size	200	44	14	80
Tags name	ATDs.b.P	ATDs	AT	ARTD
	ATDp.y.R	ATDp	AT	ARTD
	ATDs.b.SF	ATDSF	AT	ARTF
	ATDs.w.SF	ATDSF	AT	ARTF
	ATDs.ñ.SF	ATDSF	AT	ARTF
	I.1p.CF.PF	ICF	I	ICF
	I.1p.DiFut.IMPF	IFUT	I	IFUT
	I.3p.NF.PF	INF	I	INF

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

I.1p.VF.PF	IVF	I	IVF
I.1s.SuF.IMPF	ISUF	I	ISuF
I.3p.SF	ISF	I	ISF
...	...	...	...

Table 9: Different granularities found in Wolof tagset Bamba Dione et al. (2010)

marker' followed by 'sentence focus particle'. Thus, the multiword 'maa ngi' is POS tagged as 'maa/ISF ngi/UPSF', where ISF means sentence focus inflection marker and UPSF is a sentence focus particle. Their tagset granularity is into four types: a fine-grained of 200 different classes, which they used to annotate the entire gold standard corpus; a medium coarse tagset of 44 tags; more coarse tagset using the 14 common grammatical classes; .and a standard tagset of 80 tags which is define as useful for morphosyntactic studies of Wolof [5].

Yoruba is one of the major languages used in South western and North central of Nigeria. Its annotated corpus was developed from the Yoruba-English and English-Yoruba dictionaries, YLP lexical database containing 450,000 words and Yoruba lexical analyser. An output of 312,562 annotated corpus with POS tags was achieved [23]. The lexical database is the work of Awoyele released to Linguistic Data Consortium (LDC) in 2008.

Amharic is a language in the Semitic family [24, 3]. It is spoken in Ethiopia by about 30 million speakers as first or second language [3]. During the Amharic tagset development, [24] identify some orthographic system issues, such as: allowing words to be delimited by space, words are formed by joining two or more words together to form a lexical unit, non existence of capital letters in the writing system, and the use of only consonants and long vowels. The short vowels are left for the readers to fill the gaps. The steps [25] adopted in developing Amharic corpora and annotation are corpora collection and manual tagging, automatic POS tagging, morphological analysis, and further refinement and application of the resources. Sources of their untagged corpora are Ethiopian News Headlines (having approximately 3.5 million words in Amharic text), Walta Information Center (consisting of 8715 Amharic news articles)– partly annotated with appropriate POS tags by human annotators [25], and two bilingual corpora of Amharic-English consisting of government policy files which are collected from the Ethiopian Ministry of Information web. There have been three different tagsets developed for Amharic corpus POS tagging. The first two came from linguists in the Ethiopian languages Research Center (ELRC) at Addis Ababa University (AAU) [25, 3]. The basic

htp	**Name**	um–	aba–	um–	imi–	ili–	ama–	isi–	izi–	in–	izin–	ulu–	ubu–	uku–
	Class	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10	n11	n14	n15

Table 10: Xhosa noun class prefixes developed from [26]

tagset has 10 common grammatical classes, and one other tag (UNC) for problematic words. The 10 basic types were further subdivided into 30 types (describe in the work of [27]) to accommodate extended lexical functions attached to conjunction, pronoun, preposition, numerals and verbs. The third tagset was made by Sisay in 2005 [25]. This tagset (Sisay) was used in POS tagging experiments based on Conditional Random Fields. The manually POS tagged corpus15 of Amharic originally contains 210000 words from 1065 Amharic news articles tagged using 30 grammatical classes [25]. In the POS tagging experiment of [3], the three tagsets were adopted. Firstly, the largest 30 tagset developed by ELRC. Secondly, the 11 basic tagset that contains 10 grammatical classes. And thirdly, tagset by [28] was used for comparison reasons. To retain the core tags, the full tagset was mapped to only 10 tags such that UNC is mapped to residual, CONJ and PREP are mapped to adposition, and N and PRON mapped to noun [3].

[26] proposes corpus-based approach for developing tagset and training data for Xhosa language of South Africa. They chose this method because of the challenges of linguistic phenomena most AIL are facing, such as agglutinative or morphemic merging languages. The corpus-based approach enables information retrieval from enriched corpus, which is achieved through annotating linguistic facts. The annotations are used to derive specific linguistic, grammatical and lexical patterns from the corpus. Instead of manual tagging of Xhosa, the authors proposes a computer-based-drag-and-drop tagger and the training corpus data developed will be used to train a POS tagger for the language. Xhosa tagset design goes a bit further than the two normal tagset create levels: core POS tags and syntagmatic morphological categories. There is also paradigmatic distinctions, which tries to identify the paradigmatic inflections within a particular syntagmatic morphological class. For example, the word abantwana "children" in the first level will be tagged "N". In the second level, the degree of granularity is increased through POS tagging each of the prefixal, stem and suffixal morphemes based on their lexical functions. Here, abantwana will be tagged as

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

"a/PREF+ba/PREF+ntw/NSTEM+ana/SUF"16 . While in the third level, instead of the prefix PREF in the second level, they find a distinct POS tag for each of the noun class prefixes from the predefined list in the table 10.

2.1. NON-AFRICAN INDIGENOUS LANGUAGES (AIL)

Sherpa is spoken in Nepal (South Asia) and Sikkim communities. There are about 200,000 speakers who live in Nepal, 20,000 in Sikkim and 800 in Tibet17 . In the past, Sherpa language is spoken without letters. But in the recent years, Sherpa language scripts are based on the Sambota scripts, which is Tibetan orthography [29, 30]. According to [30], there are limited written text available for Sherpa language, therefore, the tagset developed is for the written texts available in the language. The tagset was prepared for tagging Sherpa texts in Sambota scripts following Tibetan orthography, which led to the use of tokenization that is based on Tibetan orthography. Sherpa language does not have any inflection in regard to gender, person, and a number due to its agglutinative form. It is rich in derivational morphology and word order is subject-object-verb. In Sherpa noun phrases, modifiers follow the head noun and there is no morphological marker to show tense. Tenses are expressed by the interaction of adverbs, aspect and evidential marking. Sherpa tagset was developed in [30]. It contains 86 tags, which includes minor and major categories in the Sherpa written texts. Priority was given to the morphological and syntactic aspect during the design phase rather than the semantic aspect. The tagset is hierarchical morphosyntactic based-features. For compatibility and interoperability, general labels (NN, NP, JJ, CC, etc.) were used for grammatical types that are common across languages. Though, prominent lexical features attached to these types were further divided into subcategories in decomposable form. The written texts used lack some features like suffixes, the number and case markers in the nominal categories. Uniform lexical markers with no morphophonemic changes are separated from the nominals and given a separate tag as suffix while all others with morphophonemic changes are given separate tag apart from suffix. The verbal forms, aspect, mood and evidential markers are treated as suffix and given separate tags. The auxiliary and copular verbs, nominalizers, are treated as separate tags. The Sherpa verbal categories take negative markers as prefix. Though, at times it comes in between the verb root and causative marker to cause the negative form of the verb. In the tagset design, the negative affix is separated and given a tag. Negative marker can equally occur in adjective as prefix, it is separated from adjective and given a tag as done in the verb. The onomatopoeic and echo-words (that is, words that imitate the sound they denote, as ideophones in Igbo) were given separate tags. There is no well defined Sambota scripts as regard to syntactic punctuation marks for off words, clauses, and enumeration. [30] sub-categorized Sherpa's punctuation mark into three; syllable, word, and sentence boundary markers in the text and proposed a separate for them. Symbols such as brackets, mathematical operators are given separate tags.

Kurdish is a Northwestern Iranian language spoken in Eastern Turkey. [27] were able to build a medium-scale morphological lexicon for Kurmanji Kurdish using freely available lexical resources. The lexical categories list was developed from Kurdish reference grammar. This contains grammar lists as nouns, verbs, pronouns, numerals, adjectives, pre-, post- and circumposition, complementizers and several particles. Kurdish morphological lexicon called Kurlex was developed through morphological description within Alexina framework. This is achieved through converting their lexical resources into Alexina18 format and using them to extract as much information as possible. A tagset consisting of 36 tags was designed and developed.

[31] presented an initiative project by Open Linguistic Resources Channelled towards InterDiscipline research (ORCHID) geared towards developing linguistic resources for Thai and Japanese languages to support NLP research. The ORCHID corpus for Thai contains about 400k words of the National Electronics and Computer Technology Centre (NECTEC) proceedings in Thailand. NECTEC focuses its research on NLP for Thai language. ORCHID Thai corpus was developed from limited resources with most of the text entered into the system through keyboard. Apart from automatic POS tagging, all other processes were manually executed with limited software support. The Thai original POS class has 13 grammatical classes with 45 subcategories. For their research aim, the POS classes were redefined, some POS tags added to clarify ambiguity, and this led to a new 14 word classes with 47 subcategories. The redefinition of the original POS tags affected the classifier (CLAS) and prefix (FIXP) classes. As a measure to alleviate POS tagging difficulties in manual process, problematic cases were illustrated in their tagging scheme to act as a guidelines in determining the correct POS tagging type in the cases of potential ambiguity. An example of such guideline between verb and preposition is given based on these two classes having the same lexical forms, and making distinctions between them is difficult in POS tagging. In order to clarify how they will be tagged if encountered, the authors made the following intuitive guidelines (1) preposition cannot be negated, while verb can. (2) preposition status can be tested by moving the preposition phrase around within the same sentential context. Preposition always accompanies the proceeding noun under movement, but verb does not. ORCHID is the first project to build Thai tagged corpus.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

[32] describes the development of the Punjabi tagset for the purpose of machine learning POS tagging. Before their work, only one tagset sized 630 fine-grained tags was in existence. This tagset consists of all the tags for the various word categories, word specific tags and tags for punctuations. [33] used 630 fine-grained tagset to implement HHM tagger for Punjabi, in which 503 tags out of proposed 630 tags were found in 8-million words of Punjabi corpus. Corpus source was online collection. [33] started a different tagset for the purpose of their work. The tagset was developed by using coarse-grained granularity for representing morphosyntactic features of Punjabi, which led to a tagset size of 40 tags. The tagset developed was compared with the existing tagsets for Indian languages.

[34] describes the development of automatic POS tagging of Urdu texts from scratch. He started with tagset design and guidelines for manual POS and post-editing tagging. The tagset design complied with the EAGLES standard on morphosyntactic annotation where necessary. The Urdu grammar used as a model for the tagset design is based on [35]. The tagset size developed is 400 POS tags and manual POS tagging was undertaken to obtain POS tagged corpus for Urdu which serves as a training data for the implementation of POS tagger for Urdu language.

Nelralec is Nepali Language Resources and Localisation for Education and Communication, designed to develop corpus and computational linguistics in Nepal language. This is via the implementation of new corpus-based lexicography methods in a new and empirical Nepali dictionary. Justification of POS tagging for Nepali are based on annotating the new Nepali National Corpus (NNC) with POS tags to ensure its status as a state-of-the-art language resource, help in corpus-based lexicography, provide an upgraded resource for language engineering implementations, and to widen the range of survey available to future researchers exploiting the corpus in the analysis of the grammatical and textual structures of Nepali [36]. Tagset for Nepali was developed by a team of linguists from Tribhuvan University [37]. The initial set of categories was based on the Nepali grammar of [38]. Iteratively, the tagset was implemented using a small data samples, discussed, re-evaluated, and then re-tested for several weeks. The tagset is hierarchical in nature, for example in VVYN1F, V − indicates verbs, V V − indicate finite verbs, V V Y − indicate third person finite verbs, etc. There are two structural features in the tagset, (1) the Nepali postpositions, which are specially written as affixes on the nouns or other words that they control, are treated as discrete tokens in this scheme of study. This gives the tagset the tolerance needed to handle the very large range of potentially possible arrangements of case. (2) And the tense, aspect and modality are not marked up on finite verbs, which are categorized solely depending on their agreement marking – a needful simplification for handling the very complex verbal inflections of Nepali, which, along with the use of compound verbs, could not be marked by the tagset without the use of several additional categories" [36]. Nepali tagged corpus for training and testing automated system was created by a team of analysts undertook the tags insertion by hand into one of the texts. The process involves tokenization, assigning a tag, assembling lists of morphological rules and exceptions, and so on. All were executed by hand. However, as the size of linguistic knowledge in the manually annotated dataset grew, it became possible to include that knowledge into a preparatory version of the automatic tagger, which was then run on the texts prior to manual investigation. Manual annotation of a 350000 word subsection of the 1 million word Nepali National Corpus Core Sample took several months [36].

The Kazakh (spoken by Republic of Kazakhstan) tagset was designed in the internal criterion principle where a POS tag is followed by a paradigm string, whose locations mean certain grammatical aspects, say a verb mood, and take certain values. For POS that take inflectional suffixes, there are respective paradigms along with generative scopes, that is, the upper bound limit on a number of possible tags that can be generated from a given POS and the different compositions of the corresponding paradigms. The maximum size of the tagset (36 tags) is equivalent to the total generative capacity (3844 tags). Depending on the extent of granularity needed for an application, some or even all grammatical aspects may be deleted or included back, providing additional adjustability. For example, Mektepke bardym. "I went to school". KLC tagset will represent this sentence in POS tags and its phrasal structure as follows: Mektepke/ZEP A0N0S0P3C3 (ZEP – impersonal noun; A0 - inanimate; N0 - singular; S0 - no possessor; P3 - 3rd person; C3 - dative case) bardym/ET G0T3M1V0P1 (ET - regular verb; G0 - not negated; T3 - past tense; M1 - indicative mood; V0 - active voice; P1 - 1st person) ./. [39].

III. CONCLUSION AND FUTURE WORK

We looked at various sizes of tagset and corpus data for languages, challenges associated with the design and development (especially in African Indigenous languages (AIL)) and how best to resolve it, and the guideline necessary for start-up design of a tagset. For this course, we studied existing tagsets and corpora design and developments for various languages. The strength and limitations of each tagset and/or corpus development were taken into account as guides to ensure standardization in creating our tagset and corpus. The transferring of tagset tags onto a corpus through

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

the tagging process and the ambiguous assumptions underlying the various operations are made clear, as in the case of how best to undergo the morphological analysis of verbs or what should be the best size of a tagset.

REFERENCES

1. LRE MAP. https://en.wikipedia.org/wiki/LRE_Map[Accessed: 31/07/2016]
2. African Language Corpora. https://corplinguistics.wordpress.com/tag/swahili/[Accessed: 31/07/2016]
3. Björn, G., Fredrik, O., Atelach, A. A., and Lars, A. (2009). Methods for amharic part-of- speech tagging. In Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages AfLaT 2009, pages 104–111. TEHNOGRAFIA DIGITAL PRESS 7 Ektoros Street, Athens, Greece.
4. Tachbelie, M. Y., Abate, S. T., and Besacier, L. (2011). Part-of-speech tagging for under-resourced and morphologically rich languages — the case of amharic. HLTD, pages 50–55.
5. Bamba Dione, C. M., Kuhn, J., and Zarrieß, S. (2010). Design and development of part-of-speech-tagging resources for wolof (niger-congo, spoken in senegal). In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta. European Language Resources Association (ELRA).
6. Trushkina, J. (2006). The north-west university bible corpus: a multilingual parallel corpus for south african languages. Language Matters: Studies in the Languages of Southern Africa. UNISA Press, 37(2):227–245.
7. Mariya, K. (2012). Towards Adaptation of NLP Tools for Closely-Related Bantu Languages: Building a Part-of-Speech Tagger for Zulu. PhD thesis, Saarland University.
8. Alejandro, G. and Beatriz, A. (2013). Languages of africa. http://www.languagesgulper.com/eng/Africa.html. Accessed: 2016-01-10.
9. Gordon, R. (2005). Languages of the World, Fifteenth Edition. Ethnologue Dallas: SIL International.
10. Demuth, K., Faraclas, N., and Marchese, L. (1986). Niger-congo noun class and agreement systems in language acquisition and historical change. In Proceeding of a Symposium, Eugene, Ore., 1983, volume 7, page 453. John Benjamins Publishing Co. Amsterdam/Philadelphia.
11. Bosch, S. E., Pretorius, L., , and Fleisch, A. (2008). Experimental bootstrapping of morphological analysers for nguni languages. Nordic Journal of African Studies., 17.
12. Poulos, G. and Louwrens, L. (1994). A Linguistic Analysis of Northern Sotho. Pretoria: Via Afrika Limited.
13. Taljard, E., Faaß, G., Heid, U., and Prinsloo, D. J. (2008). On the development of a tagset for northern sotho with special reference to the issue of standardisation. Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies. AOSIS, 29(1):111–137.
14. De Pauw, G. and De Schryver, G.-M. (2009). African language technology: the data-driven perspective. In Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics, Bozen-Bolzano, 13th-14th November 2008, pages 79 – 96. European Academy.
15. Mariya, K. (2012). Towards Adaptation of NLP Tools for Closely-Related Bantu Languages: Building a Part-of-Speech Tagger for Zulu. PhD thesis, Saarland University.
16. Heid, U., Taljard, E., , and Prinsloo, D. J. (2006). Grammar-based tools for the creation of tagging resources for an unresourced language: the case of northern sotho. In 5th Edition of International Conference on Language Resources and Evaluations.
17. De Pauwy, G., de Schryverz, G.-M., and de Looy, J. v. (2012). Resource-light bantu part-of-speech tagging. In Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages, SALTMIL 8-AFLAT 2012, pages 85–92. European Language Resources Association (ELRA).
18. Arvi, H. (2004). Tagset of swatwol a two-level morphological dictionary of kiswahili. http://www.aakkl.helsinki.fi/cameel/corpus/swatags.pdf. Accessed: 2016-03-22.
19. Hurskainen, A. (2004). Swahili language manager: A storehouse for developing multiple computational applications. Nordic Journal of African Studies, 13(3):363–397.
20. Gertrud, F., Ulrich, H., Elsabé, T., and Danie, P. (2009). Part-of-speech tagging of northern sotho: disambiguating polysemous function words. In AfLaT '09 Proceedings of the First Workshop on Language Technologies for African Languages, pages 38–45. Association for Computational Linguistics Stroudsburg, PA, USA.
21. Spiegler, S., van der, S. A., and Flach, P. A. (2010). Additional material for the ukwabelana zulu corpus. Technical report, Intelligent Systems Group University of Bristol.
22. Berg, A., Pretorius, R., and Pretorius, L. (2012). Exploring the treatment of selected typological characteristics of tswana in lfg. In Proceedings of the 17th International Lexical Functional Grammar Conference (LFG 2012), pages 85–98. CSLI Publications.
23. Adedjouma, S. A., John, O. R. A., and Mamoud, I. A. (2013). Part-of-speech tagging of yoruba standard, language of niger-congo family. Research Journal of Computer and Information Technology Sciences, 1:2–5.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204002 |

24. Gebre, B. G. (2010). Part of speech tagging for Amharic. PhD thesis, University of Wolverhampton Wolverhampton.
25. Björn, G. and Lars, A. (2009). Experiences with developing language processing tools and corpora for amharic. In IST-Africa 2010 Conference Proceedings, pages 1–8. Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2010.
26. Allwood, J., Grönqvist, L., and Hendrikse, A. P. (2003). Developing a tagset and tagger for the african languages of south africa with special reference to xhosa. Southern African Linguistics and Applied Language Studies, 21:223–237.
27. Girma, A. D. and Mesfin, G. (2006). Manual annotation of amharic news items with part-of- speech tags and its challenges. Ethiopian Languages Research Center Working Papers, 2:1–16.
28. Sisay, F. A. (2005). Part of speech tagging for amharic using conditional random fields. In Workshop on Computational Approaches to Semitic Languages., pages 47–54. ACL (2005).
29. Sang, Y. L. (2005). Sherpa orthography. Technical report, Korea Research Institute for Languages and Culture.
30. Gelu, S. (2010). Pos tagset design for sherpa text. Technical report, Central Department of Linguistics Tribhuvan University, Kathmandu.
31. Sornlertlamvanich, V., Takahashi, N., and Isahara, H. (1999). Building a thai part-of-speech tagged corpus (orchid). In J Acoust Soc Japan.
32. Kumar, D. and Josan, G. S. (2012). Developing a tagset for machine learning based pos tagging in punjabi. International Journal of Applied Research on Information Technology and Computing, 3:132–143.
33. Sapna, K., Ravishankar, M., and Sanjeev, K. S. (2011). Pos tagging of punjabi language using hidden markov model. An International Journal of Engineering Sciences, 2.
34. Hardie, A. (2003). The Computational Analysis of Morphosyntactic Categories in Urdu. PhD thesis, University of Lancaster.
35. Schmidt, R. (1999). Urdu: an essential grammar. London: Routledge.
36. Nelralec (2006). A part-of-speech tagger for nepali. http://www.lancaster.ac.uk/staff/hardiea/nepali/postag.php#tagset. Accessed: 2016-01-10.
37. Hardie, A., Lohani, R., Regmi, B., and Yadava, Y. (2005). Categorisation for automated morphosyntactic analysis of nepali: introducing the nelralec tagset (nt-01). Technical report, Nelralec/Bhasha Sanchar Working Paper 2.
38. Acharya, J. (1991). A descriptive grammar of Nepali. Washington, D.C.: Georgetown University Press.
39. Aibek, M., Zhandos, Y., Islam, S., and Anuar, S. (2014). On certain aspects of kazakh part-of-speech tagging. In Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference, pages 1–4. IEEE.