Setswana Parts of Speech Tagging: Indirect Relative

Gabofetswe Malema *malemag@ub.ac.bw*
Department of Computer Science
University of Botswana
Gaborone, Botswana

Ontiretse Ishmael *ontyishmael@gmail.com*
Department of Computer Science
University of Botswana
Gaborone, Botswana

Boago Okgetheng *okgethengb@gmail.com*
Department of Computer Science
University of Botswana
Gaborone, Botswana

Goaletsa Rammidi *rammidig@ub.ac.bw*
Department of Computer Science
University of Botswana
Gaborone, Botswana

Abstract

Setswana relatives have been shown to have a wide range of structures compared to other qualificatives. They can take negation, tense and can be recursively extended using other qualificatives, adverbs, noun phrases, and verb phrases. Studies have also shown that the structure of indirect relatives is more challenging as it is less regular compared to that of direct relatives. As a result, proposed Setswana Relatives taggers performed badly on indirect relatives. In this study, we propose the use of noun phrase and verb phrase constructs to represent the structure of indirect relatives at a high level. This approach shows that indirect relatives are also regular making them also amenable to the use of regular expressions for their identification in a sentence. This study investigates the extent to which noun phrases and verb phrases could be used to construct a regular structure for indirect relatives. We developed patterns for indirect relatives which we then implemented in Python NLTK regular expressions. The proposed tagger has a recall of 69% and precision of 62%. The tagger fails in some instances due to challenges in identifying its sub-components of noun phrases, verb phrases, and qualificatives.

Keywords: Parts of Speech Tagging, Setswana Relative, Rule based POS Tagging.

1. INTRODUCTION

Part of speech tagging (POS) is a process of identifying functions of a word in a given sentence. These functions for a given tagset typically include nouns, verbs, adjectives, adverbs, demonstratives, and pronouns for most languages. Identification of a word's grammatical function in a given sentence enables one to analyze the sentence. POS tagging is therefore an important phase in developing many NLP applications such as Machine translation, grammar checkers, Text to Speech, Information retrieval, and entity recognition systems (Kumar,2011; Suzuki,2020). This makes POS tagging a necessary exercise for advanced NLP applications such as those mentioned above.

Natural Language Processing (NLP) is characterized by ambiguity and POS tagging is no exception. Some words fall under multiple POS and their tags depend on the context of the sentence they are found in. POS tagging techniques are generally categorized as supervised and unsupervised (Awwaln,2020; Demille,2020). Supervised POS tagging models learn from a pre-tagged corpus. The performance of these models generally increases with the increase in size of the training data. Unsupervised POS tagging models do not require pre-tagged corpora. Instead, they use advanced computational methods to automatically discover information, patterns, and structures from the given data. Both supervised and unsupervised POS tagging models can use rule-based and stochastic techniques. Rule-based techniques rely on handwritten rules to determine the tag or function of a word. These rules could also include contextual information about the word, its morphological form, and in some cases punctuation. Crafting rules needs a deeper knowledge of the target language, and they can be time-consuming. Studies have shown these techniques to be more accurate but less robust. Stochastic techniques include frequencies, probability or statistical information to determine POS tags. Studies have shown them to be robust as they can handle unknown words.

Indirect relatives clauses are used to express ideas such as whose, to whom, by whom, and for whom. They are used when the subject of a second clause is not the same as the subject of the first clause but related indirectly to that first subject. Examples,

Ngwana yo ba mo ratang (the child they love)

In this example, the concord *yo* refers to *ngwana (the child)* and *ba* refers to some people (they). That is, the child is indirectly referred to through them. Previous studies have shown that Setswana relatives are generally regular and could therefore be identified through pattern matching (Malema,2020; Malema,2022). However, it has been pointed out that indirect relatives have a more open structure making them difficult to represent as patterns. In this study, we demonstrate that indirect relatives are also generally regular when broken down using verb and noun phrases. Based on this approach, we show that indirect relatives too can be handled using pattern matching. Relatives could be classified into different categories including object, subjectival possessive, objectival possessive, associative adverbial, locative adverbial, and instrumental.

2. LITERATURE REVIEW

Although POS tagging has been done for developed languages such as English, Spanish, French, and others, little work has been done for low resources languages such as Setswana. Few studies done on Setswana parts of speech tagging include (Faaß,2009; Dibitso, 2019) in which statistical methods are used. In most previous works, parts of speech tagging is studied on single word parts of speech. In Setswana, parts of speech such as adverbs, adjectives, possessives and relatives are made up of multiple words. This paper investigates tagging of an indirect relative using regular expressions. It builds on previous studies in research works in (Malema,2020; Malema,2022) which proposed rule-based POS taggers for Setswana relatives and Setswana qualificatives in general. In (Malema,2022) rules for identifying basic direct and indirect relatives are put as patterns in a trie data structure. For a given input in the form of a text file, the text is first tokenized and individual words are tagged. The tagged file is then searched for relative patterns using a trie. Although the proposed method generally performs well with a recall rate of 78%, several challenges reduce performance including difficult representation of some indirect relatives. It was observed that in general indirect relatives have a complex and open structure compared to direct relatives in Setswana. In (Malema,2020) rules for identifying qualificatives are represented as regular expressions. An input sentence is first tokenized then regular expressions are used to identify basic qualificatives. The study in (Malema,2020) aimed to identify extended or complex qualificatives. That means, further rules were developed and implemented to see if an identified basic qualificative could be extended with what follows it or not. For example, in the sentence in (1).

Motabogi yo o oleng (the runner who fell) (1)
Motabogi yo o oleng maabane (the runner who fell yesterday) (2)

the verb *oleng (fell)* could be extended by an adverb as in (2). The adverb of time *maabane (yesterday)* is also part of the relative. In (Malema,2020) the main objective was to develop rules that could determine if the next word is also part of the relative or not. This study recorded a recall rate of 74% reporting similar challenges to those found in (Malema,2022). In this study, we propose general rules for identifying indirect Setswana relatives. Specifically, we proposed rules that target constructs that were found to be open and therefore not easy to put in a general pattern form. A rule-based approach is preferred in this study as from our observation from previous studies Setswana POS such as relatives and other qualificatives have a regular structure which can be taken advantage of using patterns or rules. Stochastic approaches require data that is not available for a scarcely resourced language such as Setswana. The study also provides a baseline for future studies.

3. SETSWANA

2.1 Setswana Indirect Relative

Setswana relatives generally have a regular structure as shown in (3).

relative concords + root of relative. (3)

The relative concord must match the noun by class and the root can be a verb, adjective, or adverb. Setswana direct relative concords come in pairs and match the noun described by class. In the case of an indirect relative, the first concord refers to the indirectly referred noun as was illustrated in (1). In this case *yo* refers to the indirect person related to the direct persons referred to using concord *ba*. Indirect relatives could be categorized syntactically according to how the antecedent relates to the predicate in the sentence. These categories include plain objectival, subjectival possessive, objectival possessive and adverbial relationship (Cole,1955). Examples are

Indirect Relatives of Plain Objectival Relationship:

Bana ba re barutang (children that we teach)

Dilotsekesa di busang (things which I did not return)

Monna yobathobamoratang (the man that people like)

Indirect Relatives of Subjectival Possessive Relationship:

Ngwana yo batsadi ba gagwe ba tlhokafetseng (the child whose parents died)

Batsadi ba bana ba bone ba paletsweng (parents whose children failed)

Indirect Relatives of objectival Possessive Relationship:

Motsadi yo koloi e thudileng ngwana wa gagwe (the parent whose child was hit by a car)

Batsadi ba ke rutang bana ba bone (the parents whose children I teach)

Indirect Relatives of Adverbial Relationship:

Batho ba ke berekang le bone (the people I work with /the with whom I am working)

Motsadi yo ke berekang le bana ba gagwe (the parent whose children I am working with)

Koloi e ba mo thudileng ka yone (the car with which they hit him/her with)

Ntsalaka yo ke dirisang koloi ya gagwe (my cousin whose car I am using)

Bana ba sekoloto sa bone se duetsweng ke batsadi ba bone (children whose debt was paid by their parents)

Morutabana yo bana ba buang jaaka ene (the teacher who the children speak like)

Lefelo le dikgomo di fulang mo go lone (the place where cattle graze)

Although, linguistically indirect relatives are categorized as in the examples above, computationally we look at them in a similar way in terms of their structure. That is, some categories structurally are the same. We therefore, do not look at them by linguistic categories. In (Malema,2020) complex qualificatives were identified by developing rules on how to extend basic qualificatives. The developed rules work for most qualificatives. However, they do not work well with indirect relatives. The structure of indirect relatives exposes opportunities for expansion in the middle of most of the relatives rather than at the end of a relative. We describe the structure of indirect relatives using noun phrases (NPs) and verb phrases (VPs) in the following subsection.

2.2 Setswana Relatives Structure using NPs and VPs

A basic Setswana sentence structure consists of a noun phrase (NP) and a verb phrase (VP) (University of Botswana,2000). For example,

Bana ba/NP ya sekolong/VP (children are going to school). (4)

The NP can be a word or a group of words. It could be just a noun or a noun plus a quantifier, demonstrative, pronoun, or qualificative. The VP is made up of a verb part and complements. Complements are the goal of the action of the verb. Complements include NP, adverb, quantifiers, and pronouns. With NP and VP structure in mind we can have a look at basic indirect relatives and how they could be extended making them more complex in terms of length. Our observation is that verbs and nouns in the middle of an indirect relative can be extended making the indirect relative lengthy and look more complex and without a regular pattern. Our strategy is to first identify NPs and VPs in a given sentence. In the next pass, we then determine if identified NPs and VPs are part of an indirect relative or not. That means, our pattern for indirect relatives reduces all possible NPs and VPs to just NP or VP. In this way, the structure is made simpler even though the NPs and VPs could be long and complex.
Examples (5) and (6) below illustrate the point.

Ngwana yo batsadi ba gagwe ba tlhokafetseng (the child whose parents passed away) (5)

Ngwana yo batsadi ba gagwe ba ba neng ba bereka kwa Palapye ba tlhokafetseng
ka kotsi ya koloi maabane bosigo (the child whose parents were working at Palapye and died by car accident yesterday night) (6)

The indirect relative in (5) could be expanded by describing the noun *batsadi (parents)* further using adjectives, qualificatives, and other descriptives. It must be noted, as pointed out in (Malema,2020) that nouns can be extended using descriptives and verbs could be extended using NPs and adverbs in a recursive manner. Based on this we can generate more complex indirect relatives from basic indirect relatives. In example (6), we have expanded the noun *batsadi* by including more information through a relative (*ba ba neng ba beraka kwa Palapye*) and also expanded the verb *tlhokafetseng* (passed away) by explaining how the parents (*batsadi)* died (*tlhokafetseng*). *ka kotsi ya koloi (by car accident)* is the adverb added to the verb. The noun and the verb could be expanded in many ways to suit the situation being talked about. Our observation is that, although the noun and verb can take an infinite number of additions, we can represent these two as NP and VP respectively. That means all nouns extensions are basically NPs and all verb extensions are VPs. Therefore, if NPs and VPs could be identified in a relative then we can identify the complete relative. Indirect relatives of the form of examples (5) and (6) could be generalized into a form as in (7)

CC1 NP SC VP (7)

Where CC1 is the relative concord according to the noun, NP represents the following noun phrase followed by its subject concord, and lastly a verb phrase. Even if NP and VP are lengthy and complex, if they are identified we can identify the indirect relative when combined with CC1 and SC in the order shown above. Below we show that other indirect relatives can be simplified similarly. The list is not exhaustive, we show some of the patterns we found in Setswana documents.

Motsadi yo koloi e thudileng ngwana wa gagwe (the parent whose child was hit by a car)
CC1 NP SC VP
yo/CC1 koloi/NP e/SC VP

batsadi ba ke rutang bana ba bone (the parents whose children I teach)
CC1 CC3 V NP ADV
Ba/CC1 ke/CC3 rutang/V banaba bone/NP

Where ADV represents an optional adverb that goes with V after the object NP.

Batho ba ke berekang le bone (the people I work with /the with whom I am working)
CC1 CC3 VP
Ba/CC1 ke/CC3 berekang le bone/VP

Motsadi yo ke berekang le bana ba gagwe (the parent whose children I am working with)
CC1 CC3 VP
yo/CC1 ke/CC3 berekang le bana ba gagwe/VP

koloi e ba mo thudileng ka yone (the car with which they hit him/her with)
CC1 CC3 CC4 VP
e/CC1 ba/CC3 mo/CC4 thudileng ka yone/VP

ntsalaka yo ke dirisang koloi ya gagwe (my cousin whose car I am using)
CC1 CC3 V NP

yo/CC1 ke/CC3 dirisang/V koloi ya gagwe/NP

bana ba sekoloto sa bone se duetsweng ke batsadi ba bone (children whose debt was paid by their parents)
CC1 NP SC VP
ba/CC1 sekoloto sa bone/NP se/SC duetsweng ke batsadi ba bone/VP

morutabana yo bana ba buang jaaka ene (the teacher who the children speak like)
CC1 NP SC VP
yo/CC1 bana/NP ba/SC buang jaaka ene/VP

lefelo le dikgomo di fulang mo go lone (the place in where cattle graze)
CC1 NP SC VP
Le/CC1 dikgomo/NP di/SC fulangmo go lone/NP

se mmu wa bone o se tlhokang (what their soil needs)
CC1 NP SC CC2 VP
Se/CC1 mmuwa bone/NP o/SC se/CC2 tlhokang/VP

se o ka reng se thubegile (the one that looks like/seems its broken)
CC1 o ka reng CC2 VP
Se/CC1 o ka reng se/CC2 thubegile/VP

e ba ka e bonang (the knowledge they could get)
CC1 CC3 ka CC2 VP

e/CC1 ba/CC3 ka e/CC2 bonang/VP

dilo tse batho ba di buang (things that people say)
CC1 NP SC CC2 VP
tse/CC1 batho/NP ba/SC di/CC2 buang/VP

In Setswana the position of an adverb is usually after the corresponding verb. However, there are instances where it is placed well before the verb including in relatives. Here are some examples.

yo kwa tshimologong ya pego a bidiwang Boago (one who at the beginning of the report is named Boago)
CC1 ADV SC VP
yo/CC1 kwa tshimologong ya pego/ADV a/SC bidiwang Boago/VP.

In this example, the adverb is put before its corresponding verb. The position of the verb could be changed without changing the meaning of the sentence. In all these examples above the relative can be modified by introducing different tenses and negation.

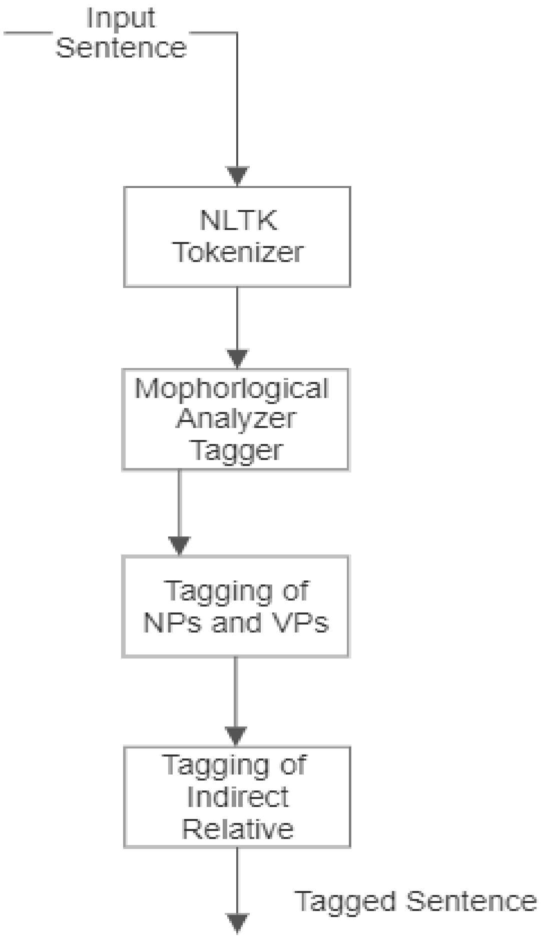

FIGURE 1: Block Diagram of Setswana Indirect Relative POS tagger.

4. PROPOSED TAGGER AND IMPLEMENTATION

The main contribution of this study to tagging Setswana indirect relatives is that the parts of the relative that were declared irregular in previous studies are in fact VPs and VPs. Based on the structure of indirect relatives described in the previous section we propose a tagger using regular expressions. Setswana indirect relatives were analyzed and represented using VPs, NPs, concords and other parts of speech as illustrated in Section 3. These structures are stored as patterns or regular expressions. The main task then is to identify the components or parts of speech that form an indirect relative in a given phrase of sentence.

The process of tagging is summarized in the following steps:

1. Tokenize a given sentence/phrase in by words.
2. Tag tokenized sentence.
3. Scan tagged sentence from 2 from right to left building VPs and VPs and relatives using VP and NP rules and stored relative patterns.

Figure 1 shows a block diagram of the proposed Tagger composing of a tokenizer and three specific taggers. An input sentence is first tokenized into individual words which are then tagged using morphological analyzers developed in (Malema,2016; Malema,2018).The generalized patterns are stored in a file representing regular expressions in Python NLTK. The patterns include indirect relatives, other qualificatives and adverbs which are used to build relatives and NPs and VPs. VPs and NPs are built through rules which could be stored in a file or be part of the code. They could also be put in the form regular expressions. In our analysis of Setswana sentence, we have observed that processing the from right to left gives the expected tree-like structure of the sentence. That is, showing how substructures are combined to form other structures such as VPs and NPs. This process is similar to that of (Malema,2020; Malema,2022) except that in (Malema,2022) a trie data structure was used to hold patterns instead of regular expressions. In the final phase, another pass is made in the sentence searching for indirect relative patterns also using the NLTK Regular Expressions.

5. RESULTS AND ANALYSIS

The proposed indirect relative POS tagger was given a text document with 52 identified indirect relatives of different forms/structures. The text document was made up of passages from local newspapers (Botswana Daily News,2022) and Jehovah's Witness website which has a Setswana version (Jehova's Witness,2022). From our experiments, the tagger positively identified 36 relatives, falsely identified 22, and had 16 false negatives resulting in a precision of 62%, recall of 69% and F1-Score of 65%. The taggers' performance depends on the identification of qualificatives, adverbs, NPs, VPs and individual word tagging. From our experiments, we found out that there are challenges in identifying these structures which may lead to false identifications or misses. The POS tagger correctly tags short or simple relatives with high accuracy. Short or simple relatives have simple NPs and VPs which have reduced ambiguity. The longer a relative is the more it is likely to have ambiguity in some of its components. Below we describe some of the challenges that we found through experimentation.

Ambiguity – As also reported in (Malema,2020) and (Malema,2022), not all structures of qualificatives, NPs and VPs are unique. Therefore, in some cases, one structure may be mistaken for another structure. In our study, we have not developed all disambiguation rules to determine the correct structure for such cases. For example, the pattern *VRB ka NP* could be interpreted in at least two ways. *Ka* could be an adverbial concord for adverbs of manner. With this interpretation, we have a verb followed by an adverb forming and a verb phrase. However, *ka* could also function as a conjunctive, meaning because, in this case, the two components do not form a verb phrase. As stated above longer relatives tend to have more ambiguity resulting in false positives.

Missing rules – One major challenge with this relative tagging approach is that there are no existing tagged data and no systematic way of finding all possible forms of the relative. This is also a common limitation in most rule-based techniques. Missing rules results in false negatives.

Errors from Morphological Analyzers – Individual words were tagged using morphological analyzers we developed in our earlier works. These analyzers still have challenges in reducing some words to their root form which affects their tagging performance. Errors in morphological analyzers result in false negatives and false positives.

6. CONCLUSIONS

POS tagging is one of the fundamental phases in most NLP applications. Indirect relatives have been identified as one of the most difficult qualificatives to identify due to their irregular structure by previous studies. In this study, we have shown that indirect relatives are also regular as other qualificatives by using sentence components formations of NPs and VPs. This approach reduces the indirect relative to a regular pattern though at a high level. This makes it easier to represent them using regular expressions. The proposed tagger obtained a high accuracy level which shows that identification of indirect relatives using regular expressions is a viable option in the absence of tagged corpora as is the case with Setswana language. Successful tagging of relatives and other parts of speech make it possible to do more sophisticated applications such as sentence analysis and construction, learning tools and translation. Results of this study contribute to the development of Setswana language processing tools. Further work on subcomponents of the proposed tagger, morphological analyzers, and identification of qualificatives, adverbs, NPs, and VPs, will improve its overall performance.

7. REFERENCES

Awwaln, J, Abdullahi, S. E, and Evwiekpaefe A, E, (2020), “PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES”, FUDMA Journal of Sciences (FJS), Vol. 4, No. 2, pp. 712 – 721, June 2020.

Botswana Daily News online (2022), www.dailynews.gov.bw.

Cole D. T, An Introduction to Tswana Grammar, Longmans and Green, Cape Town,1955.

DemillieW.B. (2020), “Analysis of Implemented part of Speech tagger approaches: The case of Ethiopian Languages”, Indian Journal of Science and Technology 2020;13(48):4661—4671.

Dibitso M.A, Owolawi, P.A, and Ojo S.O, (2019), Parts of Speech tagging for Setswana African Language, Internatinal Multidisciplinary Information Technology and Engineering Conference, No 2019, pp. 1 – 6.

Faaß G., Heid U., Taljard E., and Prinsloo D., Parts-of-speech tagging of Northern Sotho Disambiguating polysemous function words, Proceedings of the EACL 2009 Workshop on Language Techniques for African languages, AfLat 2009, March 2009, pp. 38—45.

Jehovah’s Witness online (2022), www.jw.org/tn.

Kumar V., Mukherjee S. M., and Mehta G. K., (2011). Sentiment and mood analysis of weblogs using POS tagging based approach. In International Conference on Contemporary Computing. Springer.

Malema G,MotlogelwaN,Okgetheng B, Mogotlhwane O (2016) Setswana Verb Analyzer and Generator, International Journal of Computational Linguistics (IJCL) Vol 7, Issue 1, 2016.

Malema G,MotlhankaM,Okgetheng B, Motlogelwa N (2018), Setswana Noun Analyzer and Generator, International Journal of Computational Linguistics (IJCL) Vol 9, Issue 2, pp 32 – 40, 2018.

Malema G, Tebalo B., Okgetheng B., Motlhanka M., and RammidiG., (2020) “Complex Setswana Parts of Speech Tagging”, Proceedings of the First Workshop on Resources for African Indigenous

Languages (RAIL), pages 21–24 Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020 c European Language Resources Association (ELRA), licensed under CC-BY-NC 21.

Malema, G and Ishmael O. (2022), Parts of Speech Tagging: A Setswana Relative, Journal of Physics Conferences Series, February 2022, 2188(1):012002.

Suzuki, M., Komiya K., Sasaki M.,ShinnouH. (2018), “Fine-tuning for Named Entity Recognition Using Part-of-Speech Tagging”, 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong, 1—3 December 2018, pp. 632 – 640.

University of Botswana, Department of African Languages and Literature, (2000), The Structure of Setswana Sentences: An Introduction, Lentswe La Lesedi, 2000.