Ga-RBP: A Rule-based Parser for the Syntactic Analysis of the Ga Language of Ghana

Abstract

This research presents aParts of Speech (POS) corpus and a static rule-based technique,which we refer to as Ga Rule-based Parser (Ga-RBP), for the syntactic analysis and the parsing of sentences for the Ga language of Ghana. The technique is developed to parse sentences by utilisingthe POS tagged corpus; the corpus was developed by manually tagging the Ga words with their corresponding POS tags following a standard Ga-English dictionary and custom Tagset for the language. The syntax rules were computationally defined using production rules, which establish how aword should follow the other in the right sequence to form a correct grammatical statement base on their POS. The modelgenerally analyses the sentence structure of the languageto assert its syntactic state for correctness or otherwise.

Keywords: Natural Language Processing, Parser, Parts of Speech Corpus, Syntactic Analysis, Ga Language.

I. Introduction

The major target of NLP research is to make human language computable and comprehensible by machines. This has largely driven research efforts to get various linguistic features of languages in a form that can be processed by computers via the advancement of several NLP areas. One of the core areas of NLP is syntactic analysis, where the logical structure of a language is analysed to establish its correctness according to the syntactic rules of the language. A statement can be meaningless if it is illogically constructed, or, if it does not follow its grammar rules.The meaning of a sentence for any language can therefore only be established if that sentence is correctly constructed following the syntactic rules of the language. This is why syntactic analysis is very fundamental to other NLP theories as well as NLP technologies[1,2].

Syntactic analysis forms the framework for developing parser technologies for various languages. Parsers are developed with capabilities of analysing the sentence structure to understand the order and function of words in a sentence, and how such words relate to form grammatical meaning in a particular language. One of the main resources or linguistic features necessary for developing a functional parser is the POS, since the rules of grammar vary from language to language, and words for each language have their unique POS representation in sentences, even based on context. It will usually require that these resource are first developed for a language in some way (either robustly or casually) in order to develop otheradvanced NLP tools for such languages[3].

Whiles some languages are far advanced in NLP research, there are also several languages that are underrepresented with limited resources and tools[4]; the Ga language of Ghana falls under this category of languages. The lack of, or inadequacy of computational resources for these languages limit any form of further NLP research as well as the development of applications or technologies for them. However, for languages such as English, French, German, and Chines among others, there exist large corpora and trained datasets for the various critical linguistic areas such as Morphological analysis, POS, NER, lexical analysis, semantic analysis, syntactic analysis, etc., which allows advanced research to be conducted for such languages.

The development of a parser for any language would also generally require that there are some existing basic resources like a tag lexicon, or a taggedcorpus, or a trained tagger for the language[5,6]. These existing resources are very crucial because they allow for the development of more accurate parsers for any language. Howeverfor under-resourced languages like, rule-based methods can be employed,where syntactic rules are explicitly defined together with some mechanism that use the rules to assert the state of a constructed sentence, whether it is grammatically correct or not.The basic resource which is needed in this regard is the POS corpus, which is crucial for the parser be able to identify the POS tag of word in a sentence before it can parse it.

II. Related Works

The Ga language even though under-resourced, has been explored for the computation of its grammar[7];the research lays the foundation for understanding the linguistic features of the language, including the multi-verb expressions and constructional patterns. A Ga valence dictionary is explored and compared with the Akan language of Ghana for further understanding of valence frames and verb construction in Ga. The research also relied on a lexical and other resources on which basis an extensive explanation of the language structure is presented. The research presented an import resource (even though not an NLP resource) that outlined the details of the Ga language which can be utilised for rule-based methods and for developing NLP tools, or advancing NLP research for the language.

Most African languages in general would require the use of rule-based methods, since most of them lack large datasets for probabilistic and Machine Learning approaches. The use of rule-based methods can influence the status of African languages, especially in NLP advancements, where linguistic features can be trained for further analysis, which can further open up possibilities for applications such as parsers, machine translators, information retrieval systems, self-tutored language learning systems etc[8].

Several rule-based methods have been used by different research based on the language and the target of parsing. Mohammed and Omar developed a rule-based shallow parser for the Arabic language, where they sought to target the problem of boundary identification usually faced byshallow parsing. Their parser effectively identified the entire Prepositional Phrases, Noun Phrases, and Verb Phrases boundaries of the Arabic language. The research considered the analysis of the Arabic sentence architecture for deriving more accurate rules for detecting the start and end boundaryclauses in the Arabic language[9].

Apart from the boundary identification problem in parsing sentences, which can be resolved using rule-based methods, the methods are also used by researchers for languages that have inadequate corpora. One of such languages is the Portuguese language, where a rule-based AMR parser was developed using standard rules for parsing sentences in the language. The AMR technique was used to allow for the inclusion of meaning of sentences based on the concepts and relations of words. According to the researchers, the lack of annotated corpora makes the development of parsers for the language difficult; the rule-based AMR method was then used to develop an effective parser in an attempt to bridge the gap between the language and more advanced ones (in NLP research) like the English language which has large annotated corpora and other NLP resources[10].

Other researchers have also presented rule-based dependency parsers for some rather unique text such as poetic constructed sentences in the Polish language. The technique uses a chain of word-combining-rules which operates on inputs that are *morphosyntactically*tagged rather than the use of statistical learning or formal grammar models[11]. On the other hand, other researchers have rather integrated statistical information into a rule-based lexical dependency

parser which has proven to resolve the problem of syntactic ambiguities and improved the overall parsing efficiency[12].

Rule-based techniques for developing parsers has been an effective approach for myriads of reasons, scenarios and use-cases; ranging from its effectiveness for parsing languages with Context Free Grammar (CFG) rules[13]; parsing phrases that follow some grammatical structure based on annotated data[14]; parsing text with specific rules for extraction and classification of text[15] and so forth. The utility of rule-based methods cannot be overemphasised as seen in various research work which gives credence for its further use in NLP especially for instances that require explicit definition of syntactic rules for languages with high and varied complexities, as well as for languages that are under-resourced for NLP advancement.

III. The Ga-RBP

The Ga-RBP was developed based on a recursive decent technique, which uses a top-down approach to analyse the syntax of sentences in the Ga language based on its grammar rules. The Ga-RBP depends on a POS corpus which we developed by creating a custom Tagset for the language, and then tagging the Ga words with their corresponding POS tags. The Ga-RBP references the corpus foranalysing the syntactic correctness of a Ga sentence following the syntacticrules of the language. These syntactic rules are defined computationally with production rules which establishes several sequences the various POScan be arrayed to form correct grammatical statements.

The Ga POS Tagset and Corpus

The compositing of the Tagset and the tagging of the POS were all done based on a comprehensive Ga-English dictionary which contains a list of Ga words with their corresponding POS as well as their meaning in English[16]. The Tagset composed from the dictionary has 13word classes and tags, which may not be exhaustive, given the complexity and dynamics of the Ga language. There are also several benchmark Tagset (e.g., the English Universal Tagset) that could be adopted for this purpose, however, it was more prudent to define a new Tagset following the word classes defined in the Ga dictionarywhich has slightly unique word classes and abbreviation scheme. The Table 1 presents the word classes and their corresponding POS tags.

Table 1: The Ga POS Tagset

Tag	POS Class
Det	Determiner
ADJP	Adjectival Phrase
NP	Noun Phrase
N	Noun
V	Verb
NM	Noun Modifier
NH	Noun Head
Art	Article
Adj	Adjective
Pron	Pronoun
VH	Verb Head
Adv	Adverb
Quant	Quantity

A total of 1,119 Ga words were manually tagged for developing the corpusin a tuple format for convenience of use by many learning models (Table 2 presents excerpts of the tagged corpus). Admittedly, this is a very painful exercise which could be done with other automatic methods. However, given the inadequacy of NLP resources for the Ga language, it is difficult to use any automatic or Machine learning or probabilistic tagging method, the manual annotation remains the preferred choice to develop the POS corpus for parsing the language (for this study).

Table 2: Excerpts of Ga POS Corpus

(aanyɛle, n)	(blɔfoŋme, n)	(jwere, v)
(aashikoŋ, n)	(blɔki, n)	(jwetri, n)
(aatrɛ, n)	(blɔkɔblɔkɔ, adv)	(jwetribɔɔ, n)
(aaye, excl)	(blublu, intens)	(jwɛ, v)
(aayeko, n)	(bluku, n)	(jwɛɛ, adv)
(aayelebi, n)	(bluu, n)	(jwɛi, n)
(aayenɔ, excl)	(bo, v,n,pron)	(jwɛiaŋnyo, n)
(aba, n)	(boano, n)	(jwɛjwɛɛjwɛ, adv)
(abaawa, n)	(boapia, n)	(jwɛlɛi, n)
(abada, n)	(boboo, v)	(jwɛŋ, v)
(abadai, n)	(boboobo, adv)	(jwɛŋmɔ, n)
(abakle, n)	(boda, v)	(jwiɛtɛi, n)
(abaku, n)	(bodaa, adj)	(jwii, adv)
(abalai, n)	(bodaibodai, adj)	(jwine, n)
(abantoli, n)	(bodo, v)	(jwira, v)
(abaŋ, n)	(bodobodo, n)	(ka, v,n)
(abaŋkpojurowa, n)	(bodoo, adj)	(kaa, n)
(abaŋyeli, n)	(bodua, n)	(kaabaa, n)

Ga-RBP Algorithm

The Ga-RBP algorithm was developed to use production rules based on the syntactic rules of the Ga language. The production rules evaluates a sentence following the logics of the syntactic rules to assert if a parts of speech correctly follows another or otherwise. In order to produce a valid instance of a query-language construct recursively, the production rule will have to be applied until only terminal symbols remain.The definition of the production rule consists of a combined sequence of nonterminal and terminal symbols that use several operators. The followingare the production rules for the model.

S → NP VP

NP → N| Det NP | N NP | NP ADJP | Pron ADJP

ADJP → Adj ADJP | Quan Det | Quant Art | Quant | Det | Art | Adj

VP → V | V Adv | V NP | V Adv NP| VP NP

The algorithm generally takes a sentence as input, tokenizes the sentence and then checks the corpus for the POS tag of the word. If the word is not found in the corpus, the algorithm will terminate, otherwise, the algorithm will proceed to check the next word for its POS tag and if that word correctly follows the preceding word based on the production rules. The production rule utilised by the algorithm ensures that all terminal symbols follow any of the possible syntactically correct rules before it can render a sentence as grammatically correct.

Algorithm 1: Ga-RBP Algorithm

1. START
2. Input, str S = $\{w_i | 0 \leq i \leq n\}$
3. Corpus = $\{(w_i, tag) | 0 \leq i \leq n\}$
4. token_word ← tokenize(S)
5. posTag(token_word):
6. set tag = None
7. **for** w_i intoken_word:
8. **for** w_i in Corpus **do**
9. Get.tag(w_i)
10. tag_list = [ ]
11. tag_list.append(tag(w_i))
12. **otherwise**, display(‘ word not in corpus’)
13. **end**
14. **return**tag_list
15. G(posTag()):
16. ProductionRule ← nltk.data.load('file:productionRule.cfg)
17. ga_parser ← nltk.RecursiveDescentParser(ProductionRule)
18. **for** tree in ga_parser.parse(posTag)
19. **If**tree.terminal() = ProductionRule:
20. display(‘correct grammatical sentence’)
21. **Otherwise,**display(‘wrong grammatical sentence’)
22. **end for**

IV. Experimental Results

The implementation and simulation of the Ga-RBP was done to illustrate its parsing capabilities for the Ga language. A few Ga sentences were used to demonstrate whether the parser accurately follows the various logics and syntactic rules defined for it or otherwise. The simulationexploredthe possibilities and differences in the Recursive Descent Parsing method. For each test, a parse tree is generated to assert the syntactic flow of the words in a sentence, and how each word-class based on its POS follows other words, as well as how the pattern or sequence of words are recognised to be congruent with the defined rules or otherwise.

Test one

Rule: N → V → N → Adj

Ga sentence: mi sumɔ gbɛkɛ kpakpa

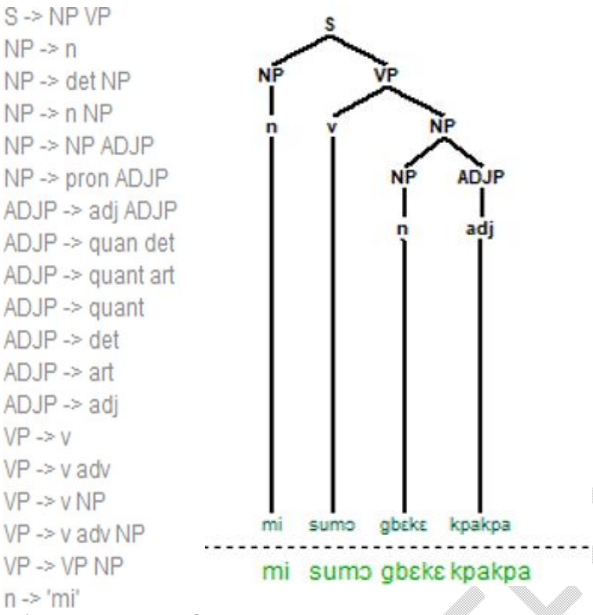

Figure 1: Test One Parse Tree

Test two

Rule: N → V → N → N → Art

Ga sentence: okoyigbɛkɛnuu ko

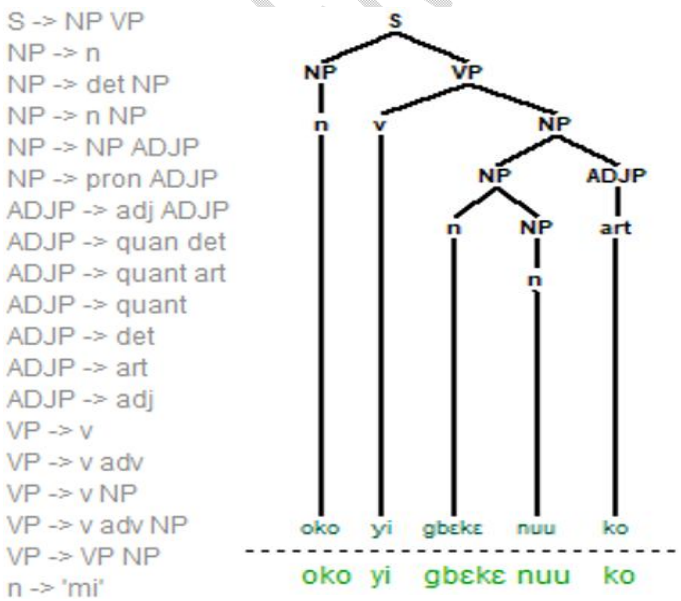

Figure 2: Test Two Parse Tree

Test three

Rule: N → Art → V → N

Ga Sentence: Papa lɛna mi

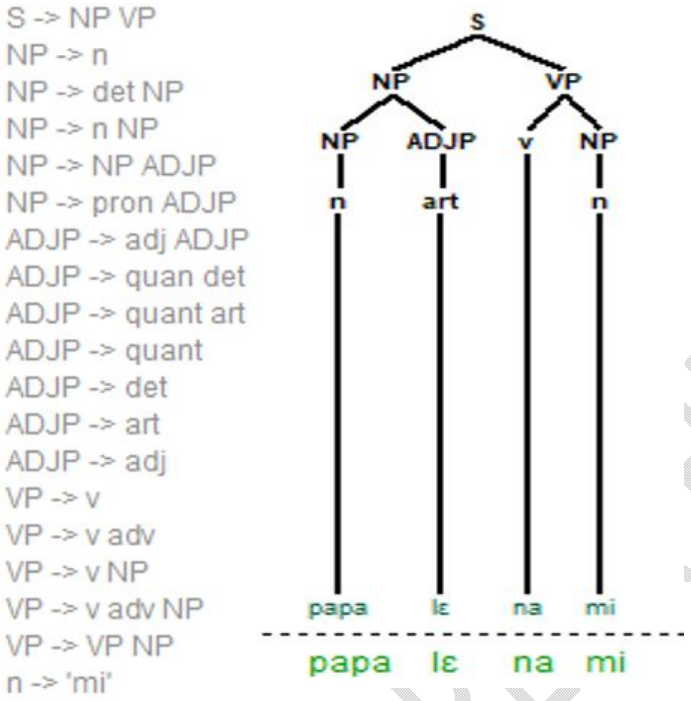

Figure 3: Test Three Parse Tree

Test Four

Rule: N → N → Adj → Art → V → N → Quant

Ga sentence: gbɛkɛnuudinlɛnalɔlɛɛyo

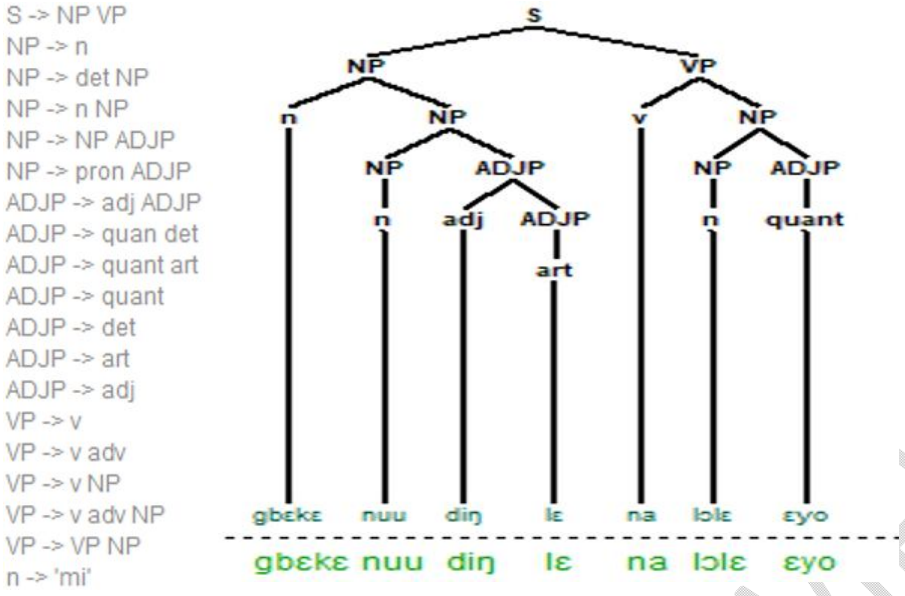

Figure 4: Test Four Parse Tree

The simulation of the model demonstratesthe ability of the model to recognise the syntactic structure of the Ga language. The model generates a parse tree following a Recursive Descent approach which illustrates the assertion that is done using the defined syntactic rules of the language.

V. Conclusions

This research presents an important advancement for the Ga language for NLP research and technologies.The POS Tagset and corpus developed for the language is an important resource, not only for the Ga-RBP but for other NLP research as well. The development and implementation of the Ga-RBP which is based on a rule-based method, presents a tool for analysing the syntactic structure of Ga sentences and parsing them to establish their grammatical state. This contribution is especially important because other NLP resource can be advanced from, or based on the Ga-RBP for their development. Resources such the machine translation of the language to other languages will need the parser to analyse the syntactic structure of Ga in order to perform the translation. The development of other NLP tool such as, Question Answering tools, Ontology Construction, Sentiment Analysis, Natural Language Generation etc., can utilise the Ga-RBP as a basic resource. That notwithstanding, we recognise the inherent limitations of rule-based parsers, which may not be able to accurately analyse sentence that have words that are not in the corpus, or within the define rules. We therefore, for future research, intend to integrate other automated techniques into the Ga-RBP to broaden the analytical scope of the Ga-RBP.

VI. References

1. Cereda, P., Miura, N. & Neto, J. Syntactic analysis of natural language sentences based on rewriting systems and adaptivity. *Procedia Computer Science***130**, 1102–1107 (2018).

2. McRoy, S. Grammars and Syntactic Processing. (2021).

3. A, Pakzad & Minaei, Bidgoli. An improved joint model: POS tagging and dependency parsing. *JAIDM***4**, (2016).
4. Pratik, Joshi, Sebastin, Santy, Amar, Budhiraja, Kalika, Bali, & Monojit, Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. in 6282–6293 (Association for Computational Linguistics, 2020). doi:10.18653/V1/2020.ACL-MAIN.560.
5. Chen, W., Zhang, M., Zhang, Y. & Duan, X. Exploiting meta features for dependency parsing and part-of-speech tagging. *Artificial Intelligence***230**, 173–191 (2016).
6. Sibarani, E. M., Nadial, Mhd., Panggabean, E. & Meryana, S. A Study of Parsing Process on Natural Language Processing in Bahasa Indonesia. in *2013 IEEE 16th International Conference on Computational Science and Engineering* 309–316 (2013). doi:10.1109/CSE.2013.56.
7. Hellan, L. A Computational Grammar of Ga. in *Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL)* 36–44 (European Language Resources Association (ELRA), 2020).
8. Hurskainen, A. Sustainable language technology for African languages. in *The Routledge Handbook of African Linguistics* (eds. Agwuele, A. & Bodomo, A.) 359–375 (Routledge, 2018). doi:10.4324/9781315392981-19.
9. Mohammed, M. A. & Omar, N. Rule Based Shallow Parser for Arabic Language. *JCS***7**, 1505–1514 (2011).
10. Rafael, T. Anchiêta & Thiago, Alexandre Salgueiro Pardo. A Rule-Based AMR Parser for Portuguese. in 341–353 (Springer, Cham, 2018). doi:10.1007/978-3-030-03928-8_28.
11. Marek, Korzeniowski & Jacek, Mazurkiewicz. Rule Based Dependency Parser for Polish Language. in 498–508 (Springer, Cham, 2017). doi:10.1007/978-3-319-59060-8_45.

12. Yoon-Hyung Roh, Ki-Young Lee, & Young-Gil Kim. Incorporating Statistical Information of Lexical Dependency into a Rule-Based Parser. in 493–500 (City University of Hong Kong, 2009).

13. Miloš Jakubíček. Effective parsing using competing CFG rules. in 115–122 (Springer-Verlag, 2011). doi:10.1007/978-3-642-23538-2_15.

14. *Syntactic Parser Assisted Semantic Rule Inference*. (2014).

15. *Rule-Based Shallow Parsing to Identify Comparative Sentences from Text Documents*. 355–365 (Springer, Singapore, 2016). doi:10.1007/978-981-10-0287-8_33.

16. Dakubu, M. E. K. *Ga-English Dictionary*. (Institute of African Studies, University of Ghana, 1973).