Asian Journal of Research in Computer Science

Volume 17, Issue 7, Page 1-10, 2024; Article no.AJRCOS.117592
ISSN: 2581-8260

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

Christian Otto [a] and Callistus Ireneous Nakpih [a*]

[a] *Department of Computer Science, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana.*

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajrcos/2024/v17i7473

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/117592

Original Research Article

Received: 27/03/2024
Accepted: 03/06/2024
Published: 11/06/2024

ABSTRACT

This research presents a Parts 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 utilising the 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 a word should follow the other in the right sequence to form a correct grammatical statement based on their POS. The model generally analyses the sentence structure of the language to assert its syntactic state for correctness or otherwise.

Keywords: Natural language processing; parser; parts of speech corpus; syntactic analysis; ga language.

**Corresponding author: E-mail: cnakpih@cktutas.edu.gh;*

Cite as: *Otto, Christian, and Callistus Ireneous Nakpih. 2024. "Ga-RBP: A Rule-Based Parser for the Syntactic Analysis of the Ga Language of Ghana". Asian Journal of Research in Computer Science 17 (7):1-10. https://doi.org/10.9734/ajrcos/2024/v17i7473.*

1. 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 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 other advanced 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 tagged lexicon, or a tagged corpus, 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. However for under-resourced languages like Ga, 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 to identify the POS tag of a word in a sentence before it can parse it.

2. RELATED WORKS

The Ga language even though under-resourced, has been explored for the computation of its grammar in a research work [7]; the research lays the foundation for understanding the linguistic features of the language, including its 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 and so forth [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 by shallow 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 boundary clauses 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 have been effective for myriads of reasons, scenarios and use-cases; ranging from their 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 works, 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.

3. 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 for analysing the syntactic correctness of a Ga sentence following the syntactic rules of the language. These syntactic rules are defined computationally with production rules, which establish several sequences the various POS can 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 13 word 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 dictionary which 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 corpus in 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 methods. The manual annotation remains the preferred choice for developing the POS corpus for parsing the language (for this study).

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 POS correctly follows another or otherwise. In order to produce a valid instance of a 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 following are 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 to assert 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.

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)

Algorithm 1: Ga-RBP Algorithm

```
START
Input, str S = {w_i | 0 ≤ i ≤ n}
Corpus = {(w_i,tag) | 0 ≤ i ≤ n}
token_word ← tokenize(S)
posTag(token_word ):
    set tag = None
    for w_i  in token_word:
        for w_i  in Corpus do
            Get.tag(w_i)
            tag_list = [ ]
            tag_list.append(tag(w_i))
```

```
            otherwise, display(' word not in corpus')
         end
      return tag_list
G(posTag()):
    ProductionRule ← nltk.data.load('file:productionRule.cfg)
    ga_parser ← nltk.RecursiveDescentParser(ProductionRule)
    for  tree in ga_parser.parse(posTag)
         If tree.terminal() = ProductionRule:
                 display('correct grammatical sentence')
         Otherwise, display('wrong grammatical sentence')
end for
```

4. 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 simulation explored the 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.

Test one

Rule: N → V → N → Adj
Ga sentence: mi sumɔ gbɛkɛ kpakpa

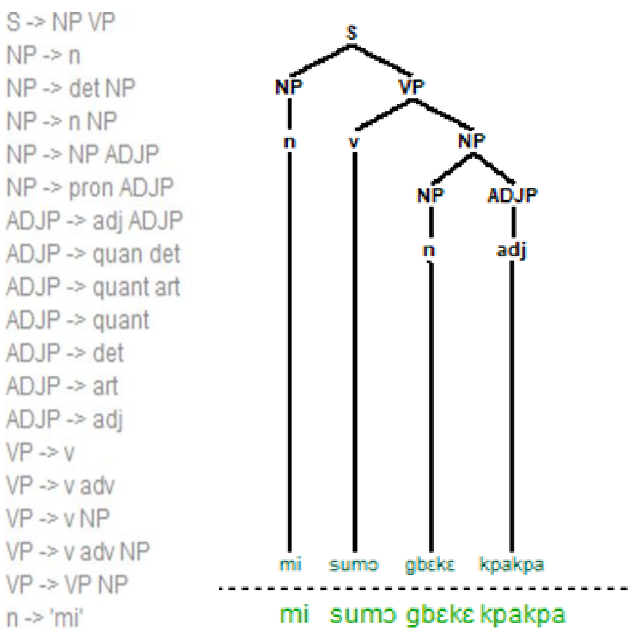

Fig. 1. Test one parse tree

The sentence "mi sumɔ gbɛkɛ kpakpa" conforms to the specified grammar rules in the following manner:

1. "mi" functions as a subject pronoun and adheres to the rule NP → Pron.
2. "sumɔ gbɛkɛ kpakpa" is a verb phrase (VP) and can be deconstructed as follows:
   - "sumɔ" is a verb (V), complying with V → sumɔ.
   - "gbɛkɛ kpakpa" is a noun phrase (NP), consistent with NP → N ADJP:

- "gbɛkɛ" functions as the noun (N), consistent with N → gbɛkɛ.
- "kpakpa" serves as the adjective (Adj), consistent with ADJP → Adj.

Therefore, the arrangement of the sentence "mi sumɔ gbɛkɛ kpakpa" aligns with the following structure:

Sentence (S) comprises a Noun Phrase (NP) followed by a Verb Phrase (VP):

S → NP VP
- NP (mi)
- VP (sumɔ gbɛkɛ kpakpa)
  - V (sumɔ)
  - NP (gbɛkɛ kpakpa)
    - N (gbɛkɛ)
    - ADJP (kpakpa)

Test two

Rule: N → V → N → N → Art
Ga sentence: oko yi gbɛkɛ nuu ko

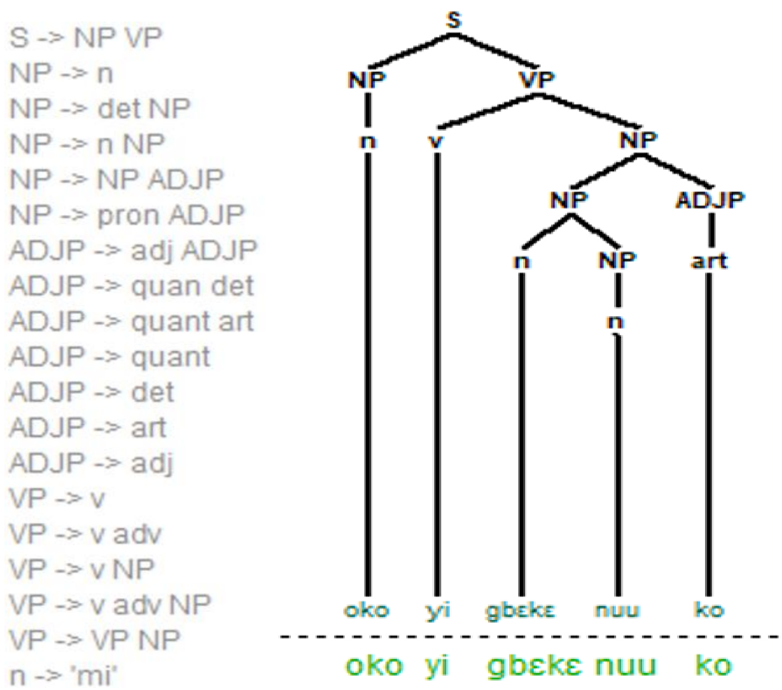

Fig. 2. Test Two Parse Tree

The sentence "oko yi gbɛkɛ nuu ko" adheres to the provided grammar rules as outlined below:

1. "oko" functions as a subject noun and conforms to the rule NP → N.
2. "yi gbɛkɛ nuu ko" constitutes a verb phrase (VP) and can be deconstructed as follows:
   - "yi" is a verb (V), consistent with V → yi.
   - "gbɛkɛ nuu ko" is a noun phrase (NP), consistent with NP → N NP:
     - "gbɛkɛ" functions as the noun (N), consistent with N → gbɛkɛ.
     - "nuu ko" represents another noun phrase (NP), consistent with NP → N Art:
       - "nuu" is a noun (N), consistent with N → nuu.
       - "ko" is an article (Art), consistent with Art → ko.

Therefore, the sentence "oko yi gbɛkɛ nuu ko" corresponds to the following structure:

S → NP VP
- NP (oko)
- VP (yi gbɛkɛ nuu ko)
  - V (yi)
  - NP (gbɛkɛ nuu ko)
    - N (gbɛkɛ)
    - NP (nuu ko)
      - N (nuu)
      - Art (ko)

Test three

Rule: N → Art → V → N
Ga Sentence: Papa lɛ na mi

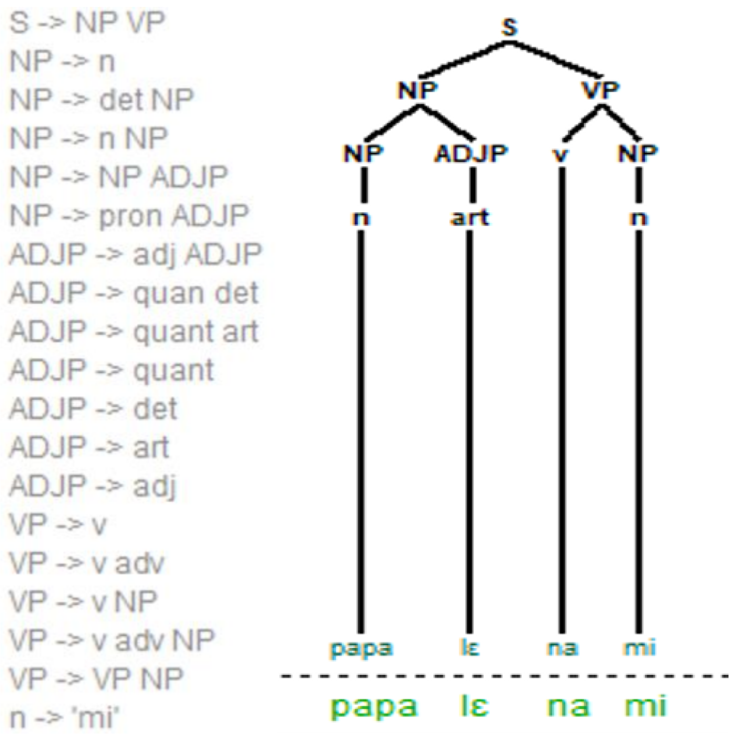

Fig. 3. Test Three Parse Tree

The sentence "Papa lɛ na mi" follows the provided grammar rules as follows:

1. “Papa lɛ” is a noun phrase (NP) which can be deconstructed as follows;
   - “Papa” as a noun (N) and,
   - “lɛ” as the article (Art).

2. “na mi” is a verb phrase (VP) which can be deconstructed as follows;
   - “na” as a verb (V) and,
   - “mi” as a pronoun, forming a noun phrase (NP) consistent with NP → Pron.

Therefore, the sentence "Papa lɛ na mi" corresponds to the following structure:

S → NP VP
- NP (Papa lɛ)
    - N (Papa)
    - Art (lɛ)
- VP (na mi)
    - V (na)
    - NP (mi)

Test Four

Rule: N → N → Adj → Art → V → N → Quant
Ga sentence: gbɛkɛ nuu diŋ lɛ na lɔlɛ ɛyo

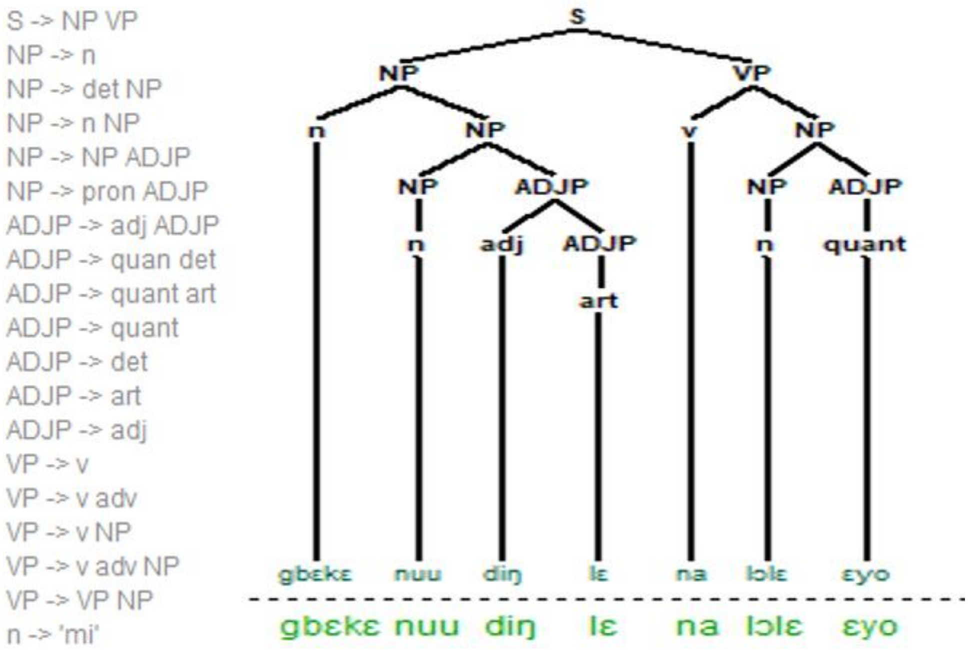

Fig. 4. Test Four Parse Tree

The sentence " gbɛkɛ nuu diŋ lɛ na lɔlɛ ɛyo " conforms to the specified grammar rules in the following manner:

1. "gbɛkɛ nuu" is a noun phrase (NP) which can be deconstructed as follows;
    - "gbɛkɛ" as a noun (N) and,
    - "nuu" as another noun, creating a nested noun phrase (this is also known as Noun Head (NH) in the Ga syntactic definitions) following the rule **NP → N NP**.
2. "diŋ lɛ na lɔlɛ ɛyo" is a verb phrase (VP), which can be deconstructed as follows:
    - "diŋ lɛ na" as a verb phrase (VP), consisting of;
        - "diŋ" as a verb (V)
        - "lɛ" as an article (Art)
        - "na" as another verb (V)
    - "lɔlɛ ɛyo" as a noun phrase (NP), deconstructed as;
        - "lɔlɛ" as an adjective (Adj), and
        - "ɛyo" as a quantifier (Quant) modifying the noun phrase (NP)

The simulation of the Ga-RBP demonstrates the 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 of sentences using the defined syntactic rules of the language.

5. 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 are important resources, 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. Technologies such as machine translation of Ga to other languages will need the parser to analyse the syntactic structure of the language for performing the translations. 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 defined 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.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Cereda PR, Miura NK, Neto JJ. Syntactic analysis of natural language sentences based on rewriting systems and adaptivity. Procedia computer science. 2018 Jan 1;130:1102-7.
   DOI: 10.1016/j.procs.2018.04.164.
2. McRoy S. Grammars and Syntactic Processing; 2021.
   Accessed: May 06, 2024. [Online].
   Available:https://uwm.pressbooks.pub/naturallanguage/chapter/syntacticprocessing/
3. Pakzad A, Minaei Bidgoli B. An improved joint model: POS tagging and dependency parsing. Journal of AI and Data Mining. 2016 Jan 1;4(1):1-8.
   DOI: 10.5829/idosi.JAIDM.2016.04.01.01.
4. Pratik, Joshi, Sebastin, Santy, Amar, Budhiraja, Kalika, Bali, and Monojit, Choudhury, 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World', presented at the Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. 2020;6282–6293.
   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. 2016 Jan 1;230:173-91.
   DOI: 10.1016/j.artint.2015.09.002.
6. Sibarani EM, Nadial M, Panggabean E, Meryana S. A study of parsing process on natural language processing in bahasa Indonesia. In2013 IEEE 16th International Conference on Computational Science and Engineering. 2013;309-316. IEEE.
   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), European Language Resources Association (ELRA). 2020;36–44.
8. Hurskainen A. Sustainable language technology for African languages. in *The Routledge Handbook of African Linguistics*, 1st ed., A. Agwuele and A. Bodomo, Eds., Routledge. 2018;359–375.
   DOI: 10.4324/9781315392981-19.
9. Mohammed MA, Omar N. Rule Based Shallow Parser for Arabic Language. JCS. 2011;7(10):1505–1514.
   DOI: 10.3844/jcssp.2011.1505.1514.
10. Rafael T. Anchiêta and Thiago, Alexandre Salgueiro Pardo. A Rule-Based AMR

Parser for Portuguese', presented at the Ibero-American Conference on Artificial Intelligence, Springer, Cham. 2018;341–353.
DOI: 10.1007/978-3-030-03928-8_28.

11. Marek, Korzeniowski and Jacek, Mazurkiewicz, 'Rule Based Dependency Parser for Polish Language', presented at the International Conference on Artificial Intelligence and Soft Computing, Springer, Cham. 2017;498–508.
DOI: 10.1007/978-3-319-59060-8_45.
12. Yoon-Hyung Roh, Ki-Young Lee, and Young-Gil Kim. Incorporating Statistical Information of Lexical Dependency into a Rule-Based Parser', presented at the Pacific Asia Conference on Language, Information, and Computation, City University of Hong Kong. 2009; 493–500.
Accessed: May 07, 2024. [Online].
Available:https://typeset.io/papers/incorporating-statistical-information-of-lexical-dependency-1nwu3k7aeq
13. Miloš Jakubíček. Effective parsing using competing CFG rules. presented at the Text, Speech and Dialogue, Springer-Verlag. 2011;115–122.
DOI: 10.1007/978-3-642-23538-2_15.
14. Syntactic Parser Assisted Semantic Rule Inference, (Jun. 19, 2014). Accessed: May 07, 2024. [Online Video].
Available:https://typeset.io/papers/syntactic-parser-assisted-semantic-rule-inference-4sokrcs5js
15. Rule-Based Shallow Parsing to Identify Comparative Sentences from Text Documents; 2016.
DOI: 10.1007/978-981-10-0287-8_33.
16. Dakubu MEK. Ga-English Dictionary. Institute of African Studies, University of Ghana; 1973.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/117592