Issues in Computational System for Morphological Analysis of Standard Yorùbá (SY) Verbs

Olufunke A. Oyinloye, and Odetunji A. Odejobi

***Abstract*— The focus of this study was based on the issues of a computational system for morphological analysis of Standard Yorùbá (SY) verbs. The system we developed was designed using a Finite State Automata (in order to captures the knowledge computationally) and the Unified Modelling Language. Python programming language was used for implementation and subsequently the system's output was evaluated with the use of a questionnaire as the research instrument. The reliability of system's output in relation to the respondents' response from the result generated was estimated using SPSS statistical software and Cronbachs Alpha statistical formular. The Cronbachs Alpha estimate (α) range used in the study was from 0 to 1. The reliability of the systems output in relation to the respondents response showed that the Cronbachs Alpha estimates obtained were 0.978 for the verb expression form and 0.980 for the orthography. The two Cronbachs Alpha estimates obtained (within the range $0.9 \leq \alpha \geq 1$) gave an excellent measurement that indicated a high correlation between the output generated by the developed system and the respondents' response. In conclusion, the Cronbachs Alpha estimate at the range 0.900 indicates an excellent measure of the respondents' response in relation to the system's output, in essence this showed there is a relation and consistency between the two, hence, it was concluded that the system's output is reliable.**

***Index Terms*—morphological analysis, standard yorùbá verbs, orthography, morphological analyzer, computational morphology.**

I. Introduction

Morphology studies the structure of words; it actually deals with inner structure of individual words and the laws concerning the formation of new words from morphemic pieces. [16] Morphological analysis refers to the computational processes which provide structural information about surface words in a language which results in the separation of words to produce morphological and morpho-syntactic features (properties) such as the word root (stem), tense (present, past, negation, continuous, future), person (1st, 2nd older,3rd older, 2nd younger, 3rd younger), number (singular and plural), aspect (simple, simple perfect, perfective), mood (imperative, indicative, affirmative) and case (genitive, nominative, ablative, dative). A morphological analyzer separate and identify the component morphemes of the input word, labeling them with sufficient information to be useful for further processing. Morphological generation is the inverse process of generating a surface word given a morphological analysis. Computational morphology deals with automatic word-form recognition and generation [16]. One of the core enabling technologies required in natural language processing applications is a morphological analyzer. It is an established fact in computational linguistics that a morphological analyzer is a starting point for many natural language processing applications [12], [17]. Yorùbá language is part of the Edekiri sub-branch of Niger-Congo Language family in spoken West Africa by more than 40 million speakers [10]. Yorùbá has many dialects, with a standard Yorùbá (SY) for communication and education and a tonal language with 3 tones: High (H), Medium (M), and Low (L). Yorùbá is an SVO (Subject Verb Object) language, e.g.' Olú ra àga - 'Olu bought a chair'. SY has 18 consonants (b, d, f, g, gb, h, j, k, l, m, n, p, r, s, s., t, w, y), 7 simple vowels (a, e, e, i, o, o, u), 5 nasalized vowels (an, en, in, on, un), and 2 syllabic nasals (m, n). Recent years have witnessed the gradual proliferation of computerized tools for processing natural languages with complex morphology. These tools serve language researchers by providing them with an automatic interface that enables quick and accurate analyses of corpora in different sizes. This study presents a database of a SY verbs that forms an integral part of a comprehensive data system developed with the use of Python programming language to serve as an important building-block that can be employed in comprehensive morphological analyzers for grammar and spell-checkers.

Manuscript received March 10, 2015; revised March 20, 2015. This work was supported by the TETFUND (Tertiary Education Trust Fund).

O. A. Oyinloye thanks the Federal Ministry of Education for giving financial support through TETFUND (Tertiary Education Trust Fund to ensure oral presentation and publication of this paper.

O. A. Oyinloye is with Osun State College of Education, Computer Science Department, Ilesa, (phone: 080-3056-5721; e-mail: funkeoyinloye@gmail.com).

O. A. Odejobi is with Obafemi Awolowo University, Computer Science & Engineering Department, Ile-Ife, (phone: 070-3920-6542; e-mail: oodejobi@yahoo.com,).

II. Yorùbá Morphology

Yorùbá has some productive methods of word derivation. The main morphological processes in the language include: affixation, compounding and reduplication. Affixation: use of prefix and infix to derive new words e.g

– prefix : ì+sé = ìsé (poverty)

- Infix: ilé + kí + ilé = ilékílé (any house).

Compounding: combination of two independent words e.g - ilé + ìwé = iléìwé (school)

Reduplication: derivation of nouns by a total reduplication of an existing noun e.g. ọmọ 'child' = ọmọọmọ 'grand-children'

- derivation of nominal items/adjectives from verbs through a partial reduplication of verbs

e.g. jẹ 'to eat' = jíjẹ 'edible' [11]

Standard Yorùbá Language is a non- inflectional language, that is, the root word or verb stem does not change in its form, and the mapping of words to their analysis constitutes a regular relations, for example the verb stems in past tense form shows that they do not change in form but remain regular, that is, Lọ becomes – ti Lọ, Se becomes - ti Se, likewise Sọ becomes - ti Sọ. On the other hand, English language is inflectional meaning that the root word or verb stem does change in its form, for example the verb stems in past tense form shows that they do change in form, that is, go changes to – went, do changes to – did and say changes to – said. Also, SY does not make use of suffix while English language uses suffix, for example, the verb stem "Lọ" in SY remains regular while verb stem "break" in English language made use of suffix(s). Because of the regularity of SY morphology, it is easy to model it computationally e.g. Concatenation: ti + verb, that is, ti + Lọ, ti + Se, ti + Lọ, ti + Se.

III. The Yorùbá verb morphology

The syllable in Yorùbá is the smallest tone bearing unit. The five basic syllable types in Yorùbá are Vowel (V) only, Consonant and a Vowel (CV), Syllabic Nasal (N), Nasalised Vowel (Vn) and Consonant Nasalised Vowel (CVn). All multi-syllabic words in the language are combinations of these five syllable types:

V: o - you (2nd/sing/subj) She/He/It - 3rd/sing/subj
a - we (3rd/plural/subj) you - (2nd/sing/obj)

CV: jo - to dance.

Vn: un an in on en.

N: a n jó - we are dancing.

CVn: gún - to pound

There are different types of Yorùbá Verbs, but for the purpose of this study, all the Standard Yorùbá verbs are grouped into two, namely: (i) monosyllabic verbs, that is, verbs with only one syllable, for example, sùn - to sleep, jí - to wake up and rìn - to walk, and (ii) polysyllabic verbs, that is, verbs with two or more syllable, for example, fẹ́ràn – to like, rántí - to remember, gbàgbé - to forget, tẹ̀lé - to follow, láálí - to insult.

IV. Morphological Analysis of SY Verbs

SY items joined with the verb stem by the simple operation of concatenation to generate different expression forms of SY verb stem, for example, is the verb stem "wa – to come" in Figure 4 shows analysis of the different expression forms the SY verb stem can take in a sentence. This work is limited to the morphological analysis of SY monosyllabic and polysyllabic verbs, which are not splitted, in terms of tense (present, past, negation, continuous, future), person (1st first person – 1S, second person older - 2O, third person older - 3O, second person younger - 2Y, third person younger - 3Y. Also, it can be expressed in form of tense, that is, present - pr, past - pt, continuous - ct, future - fr, and negation - ne. Furthermore, it can be expressed in form of numbers, that is, singular - sg and plural - pl, also a, 2nd older, 3rd older, 2nd younger, 3rd younger), number (singular and plural) only. The tree diagram in Figure 1 represents the joint effect of the rules that constraint the realization of morpheme and the shapes of stems in SY regular verbs.

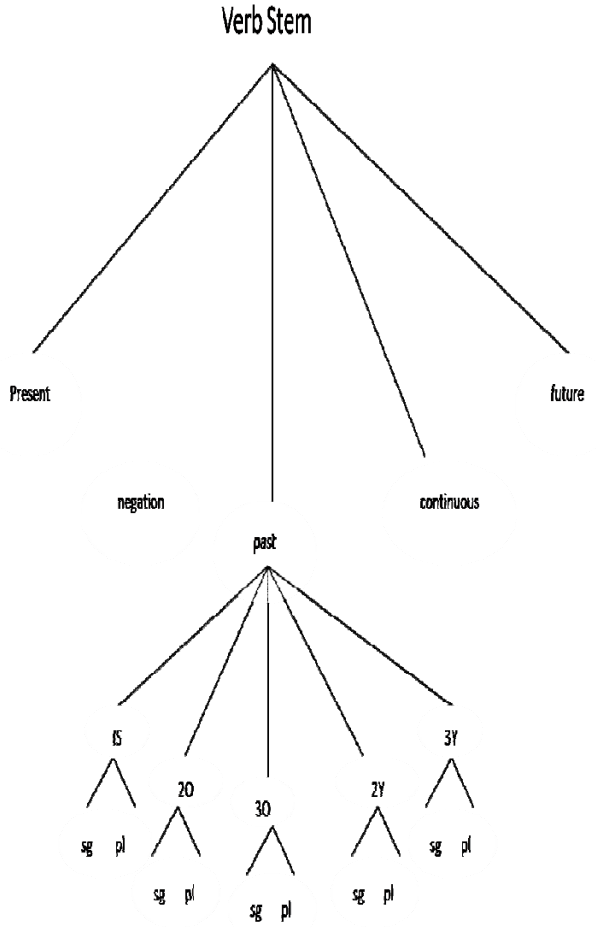

Fig. 1. Parse tree for SY verb morphology

V. Existing related works

A number of works have been reported on the computational modeling of African-language morphology. For example, Statistical data - based approach was used by [13] and [5], Corpus – based approach was applied by [5] and [6]. Finite State Approach was used by [2] and [14]. KART-Theory approach was used by [4] and [9] for Russian morphology. Also KART-Theory approach was used by [8] for Hebrew morphology, Again, KART-Theory approach was used by [7] for Yorùbá morphology.

Morphology of SY Yorùbá verbs was presented by [7] with the use of realizational approach. The aim is to identify the ways in which default – inheritance relations describe language morphology. Realizational approach was applied to the study of (SY) verbs and KART-Theory was used for its morphological analysis. Although KART-Theory was successfully implemented for verb morphology of Yorùbá

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

language, it is rather too complex since the inflectional part of the system had to be suppressed to achieve the aim. Result shows that KART- Theory has been successfully implemented for verb morphology of different languages like, Hebrew, Latin and French with limitations [8].

Morphological analysis of all part of speech was presented by [1] in Amharic Language with the use of Finite State Technique, based on the assumption that the mapping of words to their analysis constitutes a regular relation, i.e. the underlying forms constitute a regular set, the surface forms constitute a regular set, and there is a (possibly many–to–many) regular relation between these sets. Result shows that handling non–concatenative (or partially concatenative) languages using finite–state (FS) techniques is more challenging in languages whose morphotactics is morph concatenation only, finite–state (FS) techniques are straightforward to apply. Hence finite–state (FS) techniques are straightforward to apply for this research purpose

VI. Implementation

The data used for the development of this system were obtained mainly from a dictionary of the Yorùbá Language. The data collected comprises mainly of 117 monosyllabic and 134 polysyllabic (non-splitting) SY verbs with their corresponding meaning. Analysis of database consists of the following phonological structure of SY syllable: CV, CVCV, CVCVCV, CVCVCVCVCV, CVCVn, CVn, N, CVnCV. The frequencies of the data which were analyzed comprised of total number of 16 monosyllabic and polysyllabic verbs adding up to 100 respondents who are all literate and native speakers of the language in all. Total number of morphological categories of verbs analyzed were 20000.

The frequency of age ranges of all the respondents e.g the number of respondents whose age falls within the range 30-39 years is 33. The frequency of marital status of single respondents is 35 and married is 65. In addition, the gender frequency of male and female respondents were 44 and 56 respectively. A computational model for the knowledge was designed using Finite State Automata. Finite-state technology is considered the preferred model for representing the phonology and morphology of natural languages [15]. The model has been used to computationally analyze natural languages such as English, German, French, Finnish, Swahili, to mention a few cases [3], and its main advantage is that it is bidirectional - it works for both analysis and generation.

It is on this basis that the technology was selected to be applied on the morphological analysis of Standard Yorùbá verbs. The advantages of using finite-state networks include time complexity that is linear in the length of a string that is processed, and the ability to reverse the morphological analyzer and obtain a morphological generator with no extra effort [3].

The system was implemented using Unified Modelling Language and a python Programming language. The Figure 2 below shows the Activity diagram of the model.

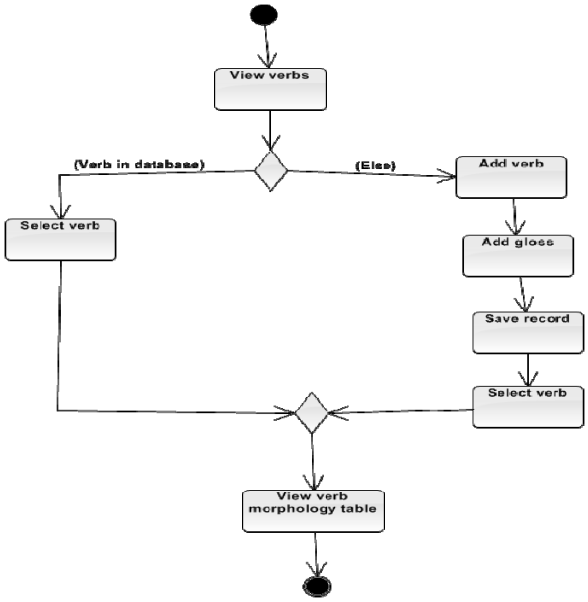

Fig. 2. Activity diagram of the model.

VII. Analysis of the Proposed System

The architecture of the proposed model in this research is depicted in Figure 3. Here, a rough explanation of the processing flow for a user's input verb stem is given. A user enters a natural language stem to demand assistance in doing his task with application software. The input verb stem is entered through the Yoruba keyboard on the module named verb stem. The verb stem links to the SY verb formation rule referring to the verb database module, which is a necessity for storing or adding SY verb stems.

There is a link between the SY verb formation rule and the verb database, to determine if input stem can be found in the database, if not, verb stem can be added at the click of add button, into the database. At the same time, the SY verb formation rule serves as the interface where the verbs stem to be analyzed is entered on the module named morphological analysis. This takes input verb stem to produce morphological and morpho-syntactic features (properties) such as, tense, person, and number. Finally, the resulting verb expression forms are generated as the output.

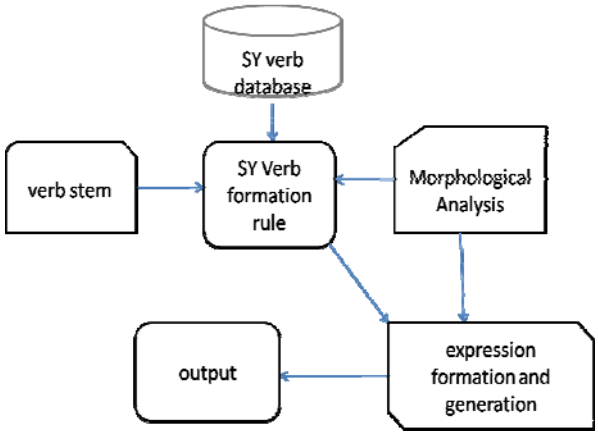

Fig. 3. Morphological Analyzer Model

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VIII. Evaluation Procedures

(a) A dichotomous analysis was carried out on the table generated from the designed system and measure instrument i.e. the questionnaire. (b) A statistical software was used to generate the mean and variance scores of the analysis specifically SPSS. (c) Cronbachs alpha statistical formular was applied on the results in (b) above to estimate reliability of the system's output in relation to the respondents' response in terms of SY verb expression and orthography. (d) A conclusion was drawn from the result of the evaluation.

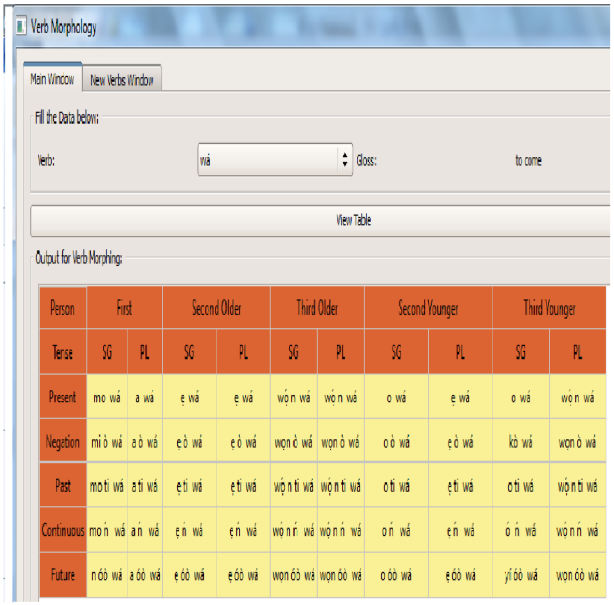

Fig. 4. System's Output VerbMorpher – Showing verb expressions for 'wa - meaning to come'

IX. Results of Evaluation

A portion of the dichotomous analysis data, carried out to obtain the correct respondents output when considered in terms of expression forms and orthography of the standard Yorùbá Verbs, and compared same with the system's output is shown in Table 1 and Table 2 respectively. Table 1 for example, shows that, for all `100' morphological categories, that is, first person singular in present tense form (1Ssgpr), generated from the system (machine output) a total number of 95 was the correct value generated from the respondents (human output), with the corresponding mean and variance of `0.9500' and `0.048' respectively. Likewise, Table 2 shows that, for all 100 morphological categories, that is, second person older plural in negation tense form (2Oplne), generated from the system (machine output) a total number of 60 was the correct value generated from the respondents (human output), with the corresponding mean and variance of 0.6000 and 0.242 respectively. All the results were generated with the use of SPSS statistical software. The Cronbachs Alpha estimate of the respondents' output considered in terms of verbs expression forms in relation to the systems' output is `0.978', and this falls within the range 0.9, also the resulting Cronbachs Alpha estimate of the respondents output considered in terms of orthography in relation to the system's output is `0.980' also falls within the range 0.9. Since the Cronbachs Alpha estimate both fall within the range 0.9, then it shows that the respondents, output is excellent and is therefore consistent in relation to the systems output, hence the output of the system is reliable.

In addition, the above result also shows that morphological analysis of Standard Yorùbá verbs with the use of Finite State Automata (FST) based technique as a tool, can be easily applied to derivational languages whose morphotactics is morph concatenation only and not non-concatenative [1]. While on the contrary the KART - theory was difficult though successfully implemented for morphological analysis of Standard Yorùbá verbs, because writing KART specifications for the language requires considerable effort, and there is need to often discard attempts and rethink how to represent the target morphology [7].

TABLE I

TABLE SHOWING ANALYSIS IN TERMS OF VERBS EXPRESSION

Morphological Categories	No of verbs from machine output	No of verbs from human output	Mean	Variance
1Ssgpr	100	95.00	0.9500	0.048
1Splpr	100	89.00	0.8900	0.099
2Osgpr	100	73.00	0.7300	0.199
2Oplpr	100	78.00	0.7800	0.173
3Osgpr	100	76.00	0.7600	0.184
3Oplpr	100	94.00	0.9400	0.057
2Ysgpr	100	92.00	0.9200	0.074
2Yplpr	100	72.00	0.7200	0.204
3Ysgpr	100	90.00	0.9000	0.091
3Yplpr	100	95.00	0.9500	0.048
1Ssgne	100	83.00	0.8300	0.143
1Splne	100	80.00	0.8000	0.162
2Osgne	100	65.00	0.6500	0.230
2Oplne	100	69.00	0.6900	0.216
3Osgne	100	64.00	0.6400	0.233
3Oplne	100	78.00	0.7800	0.173
2Yslne	100	72.00	0.7200	0.204
2Yplne	100	67.00	0.6700	0.223
3Ysgne	100	75.00	0.7500	0.189
3Yplne	100	81.00	0.8100	0.155

TABLE 2

TABLE SHOWING ANALYSIS IN TERMS OF ORTHOGRAPHY

Morphological Categories	No of verbs from machine output	No of verbs from human output	Mean	Variance
1Splpr	100	96.00	0.9600	0.039
1Ssgpr	100	93.00	0.9300	0.066
2Osgpr	100	73.00	0.7300	0.199
2Oplpr	100	75.00	0.7500	0.189
3Ysgpr	100	72.00	0.7200	0.204
3Yplpr	100	90.00	0.9000	0.091
2Ysgpr	100	94.00	0.9400	0.057
2Yplpr	100	74.00	0.7400	0.194
3Ysgpr	100	82.00	0.8200	0.149
3Yplpr	100	82.00	0.8200	0.149
1Ssgpr	100	93.00	0.9300	0.066
1Ssgne	100	79.00	0.7900	0.168
1Splne	100	72.00	0.7200	0.204
2Osgne	100	63.00	0.6300	0.235
2Oplne	100	60.00	0.6000	0.242
3Osgne	100	60.00	0.6000	0.242
3Oplne	100	68.00	0.6800	0.220
2Yslne	100	68.00	0.6800	0.220
2Yplne	100	68.00	0.6800	0.220
3Ysgne	100	73.00	0.7300	0.199
3Yplne	100	79.00	0.7900	0.168

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

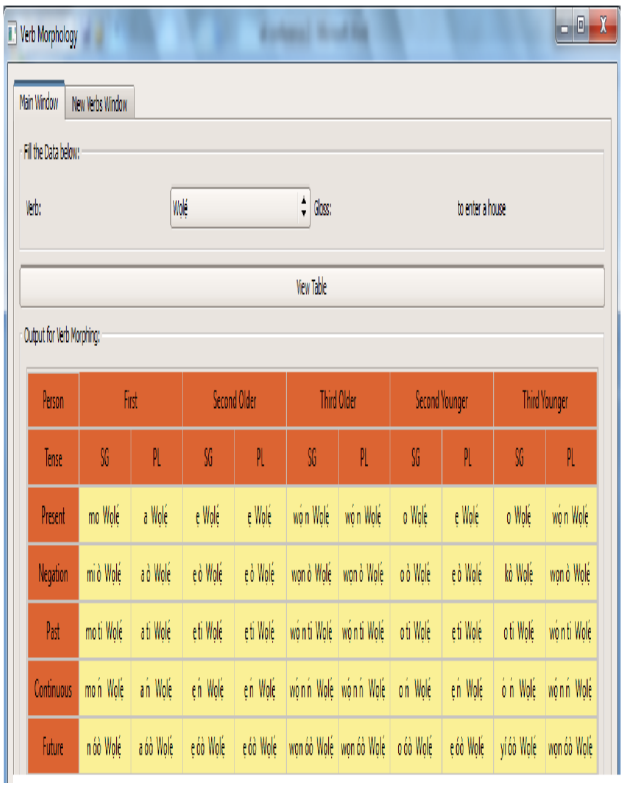

Fig. 5. System's Output VerbMorpher - Showing verb expressions for 'wọlé - meaning to enter'

Standard Yorùbá items joined with the monosyllabic verb "wá – meaning to come" by the simple operation of concatenation to generate different expressions, for example, first person singular in present tense form (ISsgpr) – mo wá, first person plural in present tense form (ISplpr) – a wá, second person older singular in present tense form (2Osgpr) – e wá, second person older plural in present tense form (2Oplpr) – e wá, third person older singular in present tense form (3Osgpr) – wón wá, third person older plural in present tense form (3Oplpr) – wón wá, second person younger singular in present tense form (2Ysgpr) – o wá, second person younger plural in present tense form (2Yplpr) – e wá, third person younger singular in present tense form (3Ysgpr) – o wá, third person younger plural in present tense form (3Yplpr) – wón wá, second person older plural in negation tense form (2Oplne) – e ò wá, third person younger plural in future tense form (3Yplfr) – wón óò wá, these and all other verb expressions are as shown in Figure 4.

Likewise, Figure 5 shows verb expressions generated for the polysyllabic verb "wọlé – meaning to enter", for example, first person singular in present tense form (ISsgpr) – mo wọlé, first person plural in present tense form (ISplpr) – a wọlé, second person older singular in present tense form (2Osgpr) – e wọlé, second person older plural in present tense form (2Oplpr) – e wọlé, third person older singular in present tense form (3Osgpr) – wón wọlé, third person older plural in present tense form (3Oplpr) – wón wọlé, second person younger singular in present tense form (2Ysgpr) – o wọlé, second person younger plural in present tense form (2Yplpr) – e wọlé, third person younger singular in present tense form (3Ysgpr) – o wọlé, third person younger plural in present tense form (3Yplpr) – wón wọlé, third person older plural in past tense form (3Oplpt) – wón ti wọlé, second person younger plural in continuous tense form (2Yplct) – e óò wọlé, first person plural in future tense form (1Splfr) – a óò wọlé.

X. Recommendation

The system's output VerbMorpher generated from the study above is therefore recommended to be used as an input for other applications such as:

- a spell checker for Yorùbá language
- a dictionary since Yorùbá outputs lemmas
- a syntax analyzer for Yorùbá language
- a language learning system for vocabulary and grammar depending on how it can be developed.

XI. Suggestion for Future Work

This study concentrated on the issues of a morphological Lexical Analyzer for only Standard Yorùbá monosyllabic and polysyllabic verbs using computational model in the form of finite automata, which is one of the essential parts of natural language processing systems. The entire plan for this research to cater for all verbs as well as other parts of speech in Standard Yorùbá language word categories to be analyzed by finite state automata, may be considered in the future. This will result into a comprehensive morphological analyzer for the Standard Yorùbá language. The morphological analyzer will be an input for many other planned applications such as learning systems and machine translation.

Acknowledgment

The authors are grateful to all who have contributed to the success of this project. Special thanks to Dr. Akanbi, Dr. Eludiora, Dr. Iyanda, Mrs Naina, Mr Akinade and all Computing and Intelligent System Research Group in the department of Computer Science and Engineering at Obafemi Awolowo Ile-Ife (ifecisrg.org), for their assistance, and support during the course of this work.

References

[1] S. Amsalu, and D. Gibbon. *Finite State Morphology of Amharic*, number 25, Germany. University Bielefeld Universitstrasse, 2003.

[2] L. R. Bahl, J. K. Baker, P. S. Cohen, A. G. Cole, F. Jelinek, B. L. Lewis and R. L. Mercer. *Automatic recognition of continous spoken sentences from a finite state grammar*. In *Proc. IEEE-ICASSP'79*, 1979.

[3] K. R. Beesley, and L. Karttunen, *Finite State Morphology. Stanford*, CA: CSLI Publications, Stanford University, 2003.

[4] G. G. Corbett, and N. M. Fraser, *Network Morphology: A DART account of Russian nominal inflection*. Journal of Linguistics, (29):113-142, 1993.

[5] A. Clark, *Partially supervised learning of morphology with stochastic transducers*. In Natural Language Processing Parcific Rim Symposium NLPRS, pp. 341-348, Tokyo, Japan, 2001.

[6] L. Deng, *Speech recognition using autosegmental representation of phonological units with interface to the trended HMM*. Speech Communication, (46): 211-222, 2005.

[7] R. Finkel, and O. Odejobi, *A Computational Approach to Yoruba Morphology*. In Language Technologies for African Languages AfLaT, volume 2, pp. 25-31, Athens, Greece. Association for Computational Linguistics, 2009.

[8] R. Finkel, and G. Stump, *A default inheritance hierarchy for computing Hebrew verb morphology*. In Literary and Linguistic Computing, volume 22 ofdx.doi.org/10.1093/llc/fqm004, pp. 117-136, 2007.

[9] A. Hippisley, *A network morphology account*. The Slavonic and East European Review, 2(74):201-222, 1996.
[10] Y. O. Laniran, and G. N. Clements, *Downstep and high rising: interacting factors in Yoruba tone production*. Journal of Phonetics, 2(31):203-250, 2003.
[11] K. Owolabi, *Noun-Noun construction in Yoruba: A Syntactic and Semantic Analysis.* PhD thesis, University of Ibadan, Nigeria, (1976).
[12] L. Pretorius, and E. S. Bosch, *Finite-State Computational Morphology:* An analyzer prototype for Zulu. Machine Translation, 3(18):195-216, 2003.
[13] S. Seneff, and C. Wang, *Statistical modeling of phonological rules through linguistic hierarchies*. Speech Communication, (46):204 - 216, 2005.
[14] M. Szarras, and S. Furui, In Proc. of the Intl. Conference on state transducer based modelling of morphosyntax with application. In Acoustics, *speech and Signal Processing*, pp. 368-371, Hong Kong, China, 2003.
[15] S. Wintner, Strengths and Weaknesses of Finite-state Technology: A Case Study in Morphological Grammar Development. *Natural Language Engineering*, 4(14):457- 469, 2008.
[16] S. Yona, *A Finite-State Based Morphological Analyzer For Hebrew.* Master's thesis, Department of Computer Science, Faculty of Social Science, University of Haifa, Technion, Haifa, Israel, 2004.
[17] S. Yona, and S. Wintner, *A Finite State Morphological Grammar for Hebrew*. In Computational Approaches to Semitic Languages, pp. 9-16, Technion, Haifa, Israel, 2005.

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)