Some Salient Issues in the Unsupervised Learning of Igbo Morphology

O. U. Iheanetu, O. Oha

Abstract - The issue of automatic learning of the morphology of natural language is an important topic in computational linguistics. This owes to the fact that morphology is foundational to the study of linguistics. In addition, the emerging information society demands the application of Information and Communication Technologies (ICT) to languages in ways that demand human-like analysis of language and this depends to a large extent on the ability to undertake computational analysis of morphology. Even though rule-based and supervised learning approaches to the modeling of morphology have been found to be productive, they have also been discovered to be costly, cumbersome and sucseptible to human errors. Contrarily, unsupervised learning methods do not require the expensive human intervention but as in everything statistical, they demand large volumes of linguistic data. This poses a challenge to resource scarce languages such as Igbo. Furthermore, being a highly agglutinative language, Igbo features certain morphological processes that may not be easily accommodated by most of the frequency-driven unsupervised learning models available. this paper takes a critical look at some of the identified challenges of inducing Igbo morphology as a first step in devising methods by which they can be addressed.

Keywords: Igbo; Igbo morphology; Rule-based learning; Unsupervised learning; Computational morphology

I. INTRODUCTION

Computational approaches to the modelling of morphology can take either of two approaches; rule-based or statistical approaches. The Rule-based approach is well suited for resource-scarce languages [9] such as Igbo. Apart from the high cost associated with rule-based methods in terms of data annotation and manual coding, it requires a fore knowledge of the morphological rules of the target language as produced by linguists and is therefore susceptible to human errors. In addition, rule-based approach is not scalable.

Recent trends in computational learning of morphology favour statistical methods of language modelling because of the low cost of building such language models. A learning process is successful in the event that '*learners are exposed to a rich input of the target language'* [13]. This concept forms the basis of statistical learning methods like the Unsupervised Learning Method (ULM).

Manuscript received July 15, 2017; revised August 10, 2017.
Olamma Iheanetu is with the Department of Computer and Information Sciences, Covenant University, Ota., Nigeria. olamma.iheanetu@covenantuniversity.edu.ng.
Obododimma Oha is a Professor of Stylistics with the department of English, University of Ibadan, Ibadan, Nigeria. obodooha@gmail.com

ULMs require large volumes of linguistic data; a scarce commodity for resource-scarce languages such as Igbo.
This method does not require a fore knowledge of the morphological rules of the target language as the system can learn from the many examples embedded in the large data available in its *environment.*

ULM mimics the passive way a child learns a first language. They are less cumbersome to implement because the high cost of annotation and manual coding are excluded.

These circumstances do not augur well for the computational modelling of Igbo, the major challenge being scarcity of Igbo linguistic data such as annotated and tone-marked Igbo corpus in computer readable form. Although this situation doesn't present a major challenge to rule-based models (RBM), it poses serious challenges in ULMs.

Most unsupervised learning algorithms for morphology induction are primarily hinged on natural language behaviour. Some of the basic behaviour of language include arbitrary segment occurrence and frequent segment occurrence. The fact that unsupervised learning algorithms based on natural language behaviour makes the ULM approach a viable option for morphological induction because such models are scalable to other languages. However, we observe that existing ULMs are not suitable for Igbo language. Although some of these models work well for Igbo inflected words, the models are not able to cater for some other words which are mostly derived words from reduplication, circumfixation, compounding and interfixation processes in Igbo. In addition, because Igbo is highly agglutinative, capable of having as high as four cascades of affixes, this nature poses a challenge for the identification of all affixes. Given these situations therefore, there is a need to device novel unsupervised methods that take into cognizance, the resource scarcity of Igbo as well as the other peculiar characteristics of the Igbo morphology.

II. PREVIOUS LITERATURE

Unsupervised learning models of morphology are usually based on common behaviours of natural language. The expectation is that such models would suit most of the world's languages especially languages with concatenative morphology and a low average number of morpheme per word. However, the underlying heuristics used for many of the popular models may not be able to properly cater for some morphological process in certain languages. These models were developed based on languages such as English in which orthographic inconsistencies necessitate adjustments in the heuristics used. Such adjustments may not be necessary for Igbo because of the high level of regularity in its orthography.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In the early 80s, [19] described the Finnish morphology using the two level rule. The success of this approach was very remarkable and the two level (**twol**) formalism since then has been used in describing RBMs for various languages. Likewise [6], [24], [22], [12] and [20] have described morphological models for Catalan, German, Persia, Sanskrit and Syriac languages respectively. In 2002, [7] proposed the use of Finite State Automata (FSA) in the modelling of morphology. This is based on the fact that natural language can be represented as Regular Expressions (RE) and REs can be conveniently represented as FSA. They introduced a Xerox Finite State based Tool (XFST), developed at Xerox lab for the computational modelling of languages employing the rule-based approach. Ever since, many studies have applied finite state techniques to the modelling of morphology. [23] and [5] described aspects of Setswana verbs and Igbo verbs respectively using XFST. Although these studies recorded success, the common denominator is that the manual annotation and the coding process is laborious and time consuming. According to [8], it took an average of 3000hrs to manually label and code the Zulu morphological analyzer.

Unsupervised learning approach is a more recent approach to morphological modelling. [25] employed a direct approach to the unsupervised learning of morphology by identifying highly productive paradigms by examining morphologically rich subsets of the input lexicon.

[16] based the extraction of morpheme boundaries on the Arbitrary Character Assumption (ACA) and Frequent Flyer Assumption (FFA). The idea is that words are normally arbitrarily occurring segments. However, some segments occur rather more frequently than others. These frequently occurring segments, which have same length with the arbitrarily occurring segments are therefore hypothesised to be affixes. This is typical of the behaviour of affixes, although there could be outliers, and these were duly catered for in the study.

[15] employed the Minimum Description Length (MDL), as basis for identifying correct morpheme boundaries. Having bootstrapped the morpheme boundary identification problem with the successor frequency heuristic proposed by [17] the morphology which gives the least minimum description length is then regarded as the best morphology of a corpus.

The above are only representative of the general approaches to the unsupervised learning of morphology

A. Background on Igbo Language

Igbo is a tonal language spoken mainly in the South-eastern part of Nigeria by an approximate of thirty million speakers. Igbo belongs to the Niger-Congo language family; having two tones and a down step that give different meanings to similarly spelt words. Many dialects of Igbo exist but Standard Igbo is widely spoken and understood, and as well generally accepted among Igbo speakers. Igbo has thirty six alphabets, including eight vowels which are grouped into two and nine diagraphs.

Igbo language features a wide variety of highly productive concatenative and non-concatenative morphological processes. Cascaded affixation is a common occurrence in Igbo morphology owing to the agglutinative nature of the language and it is also highly productive in the language.

Igbo is highly agglutinative and exhibits verbal inflections. Furthermore, [26] reported that Igbo verbs also inflect for aspect, which includes progressive, perfective, durative and inchoative which are realized via prefixation and suffixation. Plurals in Igbo are marked with the plural markers *unu*, *ndị* and *umu*.

Most Igbo morphological processes can be generalized as affixation. Affixation is a morphological process that entails the concatenation of a root word and an affix [1], where an affix is a lexical component that yields either the inflected or derived form of the word it is concatenated to.

III. THE PROBLEM

As earlier stated, the cost of rule-based morphology models is very high in terms of data annotation cost and the time involved. In addition, RBMs are subject to human error. Unsupervised learning methods automatically learn the morphology of a language without prior knowledge of the morphological rules thereby excluding any human intervention. However, the demand of this approach is unattainable for resource scarce languages like Igbo. ULMs require many examples to learn from. These examples are offered by the very large amounts of linguistic data of the target language. Igbo at present has very sparse linguistic data available for computational studies.

Igbo is a highly agglutinative language. Simple affixations may be discovered by known approaches to unsupervised learning; however, these approaches may not work well for multiple affixations, which is prevalent in some Igbo morphological processes. In this vein, [9] stated that *Linguistica;* developed by [14, 15] based on the MDL approach will not work well for Bantu languages, of which Igbo is a member. An experiment on *Linguistica* using Igbo words proved that *Linguistica* like many existing unsupervised learning models do not sufficiently capture multiple affixes.

Non-concatenative processes like reduplication and compounding which are derived morphological processes pose a challenge for existing unsupervised learning models. This is because the affixes or in words realized from these morphological processes *may not occur as frequently* in the corpus as the relevant threshold may demand. It is possible therefore that some relevant segments may record lone occurrence in the corpus and therefore many not be properly identified as valid morphemes. It is pertinent therefore to devise new methods that can cater for Igbo words that feature compounding and reduplication.

Furthermore, a supposedly rare morphological behaviour occurs in Igbo. According to [11], only Igbo and Dghwede are languages known to exhibit interfixation, even though [4] identified similar morphological process in Yoruba. Interfixes may not feature frequently occurring segments and therefore may not be captured by these frequency-based morpheme boundary identification schemes. This owes to the fact that typical Igbo interfixes are merely Igbo phonemes represented in writing as *l, m, r,* and so on, which stand between two copies of a stem. For example, *ogo**l**ogo*, *egwu**r**egwu*, *anu**m**anu*. Figure 1 below represents six

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

affixation procedures that yield four Igbo morphological processes namely; nominalization, verbal inflections, circumfixation and reduplication.

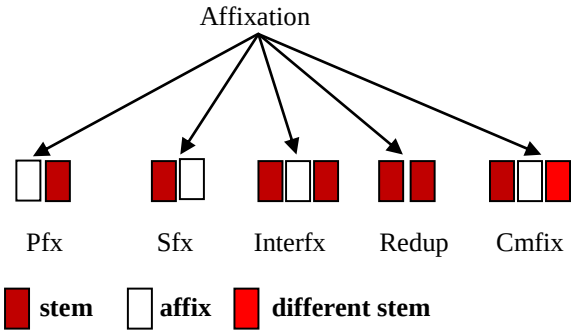

Figure 1 – Morphological realization processes in Igbo

In figure 1 above, Pfx = prefixation, Sfx = suffixation, Interfx = interfixation, Redup = reduplication and Cmfix = circumfixation

A. Some Igbo Morphological Processes

1) *Verbal Realizations*: verbal realization processes include a. Imperatives – Stem + Suffix

ga + a →gaa (go)

b. Infinitives – Prefix + Stem

ị + ga → iga (to ga)

c. Past tense – Stem + r V (where V = last vowel of a monosyllabic verb or disyllabic verb)

ga + r+ a → gara (went)

d. Participle – Prefix + Imperative form + Suffix

a + gaa + la → agaala (has gone)

stem + Extensional suffix

ga + wa → gawa (begin to go)

ga + wa + zie → gawazie (begin to go now)

ga + wa + zie + nụ → gawazienụ (go, for heaven's sake!)

ga + wa + kwa → gawakwa (proceed!)

where *wa, zie, nụ and kwa* are extensional suffixes that extend the meaning of the verb. *nụ* presents a sense of pleading while *kwa* sounds a warning. As agglutinative as *gawakwa* is, it can still take a prefix *a* and suffix *la* to realize the participle form of the word. This yields *agawakwala* meaning "has begun to go" in the literal sense.

In all the highlighted processes above, vowel harmomy is taken into cognizance.

2) Nominalization

The key factor to realizing nominalizations in Igbo is the construct of the verbs in question. Generally, either any of these characters form the prefix during the nominalization process. According to [20] the prefixes include [(a, e, i, ị, o, ọ, n, m), (ò, ọ̀, ù, ụ̀)]. Examples:

ọ + ṅụ → ọṅụ (joy)

ù + re → ùre (rotteness)

n + chekwube → nchekwube (hope)

An important thing to point out here is that not all prefixes go with every stem all of the time. Verbal constructs and vowel harmony imposes a restriction on morpheme combinatory patterns here.

3) Reduplication

Igbo reduplications are achieved by duplicating the stem word. Hence we have words like *kọịọị, ngwangwa, ikeike, gbùrùgbùrù* and so on.

4) Compounding

Compounds are gotten from the coexistence of two stems, base forms or morphemes as a single word. In general terms, a compound comprises two or more stems that come together to form a word, and according to [27] must not contain verbs with extensional or inflectional suffixes. Most Igbo names are products of compounding processes. Igbo compounds can be simple or complex. Examples are

Stem + stem

obi + ọma → obiọma (good heart)

elu + igwe → eluigwe (heaven)

chi + nwe + ndụ → chinwendụ (God owns life).

Igbo interfixation process has already been described above. The obvious challenge identified by this study is that existing ULMs is may not be able to capture the peculiarites of Igbo morphological processes as described above. Most of these ULMs are based on frequency of occuring semants or characters which is a fundamental language behaviour. However, this behaviour may be restricted in Igbo morphology by concepts like vowel harmony, verbal constructs, and tones making it less likely to encounter affixes with a high frequency of occurrence, as is the case with most languages.

IV. SOLUTION

Frequent pattern mining (FPM) was first proposed by [3] in 1993 for market basket analysis for finding associations between different items in a customer's shopping basket [18]. Since then, frequent pattern mining has been applied to some other domains of knowledge like indexing, web and stream data mining, and so on. However, to the knowledge of the authors, frequent pattern mining has not been applied to morphological induction. Frequent pattern mining has been applied in data mining to identify strong associations between item set and also for capturing underlying semantics in data [17]. The choice of adopting FPM for the unsupervised learning of morphology as a bootstrapping step is mainly based on the fact that existing unsupervised learning models do not cater for some of the productive morphological processes which are non-concatenative in nature.

A. Frequent Pattern Theory – Brief Introduction

For the purposes of this study, we have chosen to identify frequent patterns instead of frequent segments in a wordlist. As earlier discussed, some Igbo words do not manifest frequent segments in their morphology. These include words that result from morphological processes like reduplication, circumfixation, and interfixation in Igbo. The segments of such introflected words is not repeted in the wordlist and so cannot be captured. This situation rules out the possibility of identifying such segments with existing frequency-based methods. Our approach proposes a representation of the words in a wordlist as a combination of Cs and Vs with number subscripts to identify unique letters, where C = consonant and V = vowel.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In other to achieve the above, a python algorithm was written to convert all the words in our Igbo corpus as a string of Cs and Vs. C_0, C_1, C_2, etc is used to represent the first, second and third consonants respectively. Likewise V_0, V_1, V_2 represents the first, second and third vowels respectively. See example in table 1 below.

Table I – Igbo word pattern representation

S/No	Igbo Words	Pattern Representation
	oriri	$V_0C_0V_1V_0V_1$
	egwuregwu	$V_0C_0V_1C_1V_2C_0V_1$
	chịkọtachaa	c0v0c1v1c2v2c0v2v2
	ọgiga	v0c0v1c0v2
	akpamokwu	v0c0v0c1v1c2v2

With the word to pattern representation, pattern repetitions within a word are evident, regardless of the length of the word. These repetitions suggest the presence of certain morphological activity, and as such, would contain some morphological information. Such patterns occur more frequently in the corpus than some others. After the word patterns are generated, the patterns are then clustered based on similarity of the morphological process each pattern represents. Patterns within a cluster share some commonalities which could be a pattern segment. This pattern segment could be as short as a single character or as long as three or four characters. This method captures words realized from reduplication, interfixation and circumfixation morphological processes very well while inflected words appeared to have no regular pattern(s).

V. LIMITATIONS AND CONCLUSION

This paper is an overview of an on-going study, highlighting the problems which the study seeks to solve and an introduction to a proposed solution to these problems. A major setback encountered in this study is the unavailability of computer readable Igbo text as well as a fully diacritized Igbo corpus. Based on the study's preliminary findings on the Igbo language, we conclude that frequent pattern identification is a viable bootstrapping option for unsupervised learning of Bantu languages that share basic similarities in their morphology with Igbo for the morphological induction of some derived morphological processes like reduplication, interfixation and circumfixation. Our new method can aid existing unsupervised learning models to discover certain obscure morphological processes of the Igbo language which may not easily present themselves to the present popular schemes.

REFERENCES

[1] Abubakre, S. O. 2008. Affixation in Hausa and Eggon: a comparative analysis. *Journal of Linguistics, Literature and Culture* 1: 35-45.

[2] Adegbola and Odilinye. (2012). Quantifying the Effect of Corpus Size on the Quality of Automatic Diacritization of Yoruba texts. *Proc of the third workshop on Spoken-languages Technologies for Under-resource languages (SLT-U).*

[3] Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining Association Rules between Sets if Items in a Database. *Proc ACM SIGMOD* 1993, pp. 207-216

[4] Awobuliyi, O (2008). Eko Iseda- Oro Yoruba. Akure, Ondo state. Montem Paper Backs.

[5] Ayogu, I., Adetunmbi, A., Alese, B., and Kammelu, N. (2011). Igbomorph: A finite State-Based Morphological Analyser for Igbo. *Proc AICTTRA* 2011, pp. 1-12.

[6] Bardia, T., Egea, A. and Tuells, A. (n.d). CATMORF: Multi Two-level Steps for Catalan Morphology. Retrieved from http://www.sciweavers.org/publications/catmorf-multi-two-level-steps-catalan-morphology

[7] Beesley, K. R and Karttunen, L. (2003). Finite State Morphology. Stanford, United States of America. CSLI Publications.

[8] Bosch, S., Pretorious, L. and Fleisch, A. (2008). Experimental Boostrapping of Morphological Analysers for Nguni Languages. Nordic Journal of African Studies, 17(2), pp.66-88.

[9] Chinnakotla, M. K., Damani, O. P. and Satoskar, A. (2010). Transliteration for Resource-Scrce Languages. Journal of ACM Transactions on Asian Language Information Processing (TALIP). Vol 9 (4). Doi:10.1145/1838751.1838753

[10] De Pauw, G. and De Schryver, G. (2008). Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. *Proc of the 13th International Conference of the African Association for Lexicography.*

[11] Emenanjo, N (1982). The Interfix: An Aspect of Universal Morphology. Journal of West African Languages. XII, 1 (1982) 77 - 88.

[12] Girish, N. J., Muktanand, A., Subash, Sudhir, K. M., Diwakar, M., Diwakar, M., Manji, B., Surjit, K. S. (n.d). Inflectional Morphology Analyser for Sanskrit. Retrieved from http://sanskrit.inra.fr/Symposium/DOC/Girish.pdf

[13] Glatz, S. (2010). The role of New Technologies/ICT in Language Education and Teaching. Available at http://www.google.com.ng/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CDkQFjAA&url=http%3A%2F%2Fwww.arts.monash.edu%2Flanguage-and-society%2Fpostgrad-conferences%2Fglatz.ppt. Accessed on July 12, 2013

[14] Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. Massachusetts Institute of Technology (MIT) Press Journal, 27 (2), 153-198. doi: 10.1162/089120101750300490

[15] Goldsmith, J. (2005). An Algorithm for the Unsupervised Learning of Morphology. Natural Language Engineering Vol 1 (1). Cambridge University Press. Retrieved from http://hum.uchicago.edu/~jagoldsm/Papers/algorithm.pdf

[16] Hammarström, H. (2009). *Unsupervised Learning of Morphology and the Languages of the World* (Doctoral thesis), Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Sweden, 284p

[17] Han. J.; Cheng, H.; Xin, D.; and Yan, X (2007). Frequent pattern mining: current status and future directions. DOI 10.1007/s10618-006-0059-1

[18] Harris, Z. (1967). Morpheme Boundaries within Words: Report on a Computer Test. *Transformations and Discourse Analysis Papers* 73.

[19] Koskienniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production* (Doctoral thesis), Department of General Linguistics, University of Helsinki, Finland.

[20] MacClanahan, P., Busby, G., Haertel, R., Heal, K., Lonsdale, D., Seppi, K., and Ringer, E. (2010). A Probabilistic Morphological Analyser for Syriac. *Proc of the Conference on Empirical Methods in Natural Language Processing.* pp. 810 - 820.

[21] Maduagwu, G. O. (2006). *Lexicalization Strategies in Ọ̀gbahụ Dialect in Igbo, Nigeria.* (Doctoral thesis), Department of Linguistics and African Studies, University of Ibadan, Nigeria, 135p

[22] Megerdoomian , K. (2003). Finite-State Morphological Analysis of Persian. Retrieved from http://www.zoorna.org/papers/Coling04workshop-PersianFSM.pdf . Accessed on 20th August 2010.

[23] Pretorius, R. Berg, A. Pretorius, L. and Viljoen, B. (2009). Setswana Tokenization and Computational Verb Morphology: Facing the Challenge of a Disjunctive Orthography. *Proc of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* pp. 66-73.

[24] Schmid, H., Fitschen, A. and Heid, U. (2002). SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. Retrieved from http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/LREC04/smor.pdf

[25] Snover, M. G. and Brent, M. R. (2001). A Bayesian Model for Morpheme and Paradigm Identification. *Proc of the 39th Annual Meeting for Computational Linguistics (ACL),* pp. 482–490.

[26] University of California, Los Angeles (UCLA) Language Materials Project. 2009. Igbo. *UCLA Language Materials Project.* Retrieved October 20, 2010, from http://www.lmp.ucla.edu/Profile.aspx?LangID=13&menu=004.

[27] Zsiga, E. C. (1992). A Mistmatch Between Morphological and Prosodic Domains: Evident from Two Igbo Rules. Phonology, 9(1), 101-135. Doi: 10.1017/S0952675700001512

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)