Corpus-size Quantification for Computational Morphological Analysis of Igbo Language

O.U. Iheanetu, W.E Nwagwu, T. Adegbola, M.C. Agarana

Abstract - Data-driven approaches to morphology learning have gained popularity over rule-based approaches. This development favours languages with rich electronic linguistic resources because that is a major pre-requisite for data-driven models. However, due to lack of abundant electronic texts in Igbo, and other resource-scarce languages hardly benefit from data-driven approaches

In this study, we seek to quantify the actual corpus size required for morphology induction using a modest Igbo corpus. The impetus for this study being that morphological analysis may not require as much words as would other levels of linguistic analysis.

We used Word Labels (WL) which is a representation of individual words in the corpus using *C*s for consonants and *V*s for vowels. This approach helped to compress the corpus from 29191 words to 2292 unique WLs out of which were found 81 unique Igbo Morphological Structures (MS). This implies ample morphological information in the modest corpus. The unique MS found in new sets of 1000 words approached the zero mark with 6000 words, indicating the neighbourhood of exhaustion of Igbo morphology.

This study shows that electronic corpora scarcity does not constrain computational morphology studies as it would other levels of linguistic analysis.

Key words: Igbo language, Morphology Induction, Resource Scarcity, Corpus size Quantification, Data-driven learning.

I. INTRODUCTION

Computational studies of Igbo language have been plagued with the lack of availability of electronic corpus or computer readable text in the language. Lack of a gold-standard corpus for the computational study of the language has been a challenge until very recently.

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. Michael Agarana is with Mathematics Department, Covenant university, Ota, Nigeria. michael.agarana@covenantuniversity.edu.ng. Williams Nwagwu is Head of CODICE, CODESRIA, Senegal. Williams.nwagwu@codesria.sn. Tunde Adegbola is the Executive Director African Languages Technology Initiative (Alt-i), Ibadan, Nigeria. taintransit@hotmail.com. African Languages Technology Initiative (Alt-i), Ibadan, Nigeria. taintransit@hotmail.com.

Although there have been efforts to subject the language to computational studies in the past, these efforts are sparse and do not aggregate. [13], [5], [6] are some of the efforts in computational Igbo studies. Igbo language has been classified among the less studied or under studied or resource scarce languages. With an approximate 30 million native speakers, Igbo computational studies are still at its fledgling stage. As earlier stated, the lack of large amounts of electronic data in the language is a major factor contributing to this challenge. The available linguistic resources in Igbo are either not electronic form, or the electronic form is only available in very sparse quantities, lacking the necessary diacritizations. The absence of these diacritizations presents ambiguities especially with homonyms, which are prevalent in the language.

Resource scarcity is an amorphous concept when it comes to the scientific study of natural languages. The term is usually employed interchangeably with other terms such as resource-starved, under-resourced, resource scarce, less studied, least developed, under developed, under resourced, and so on. These terminologies have been used to describe languages in which insufficient or no electronic text in written or spoken form are readily available for computational studies in that language. [10] described data sparsity or resource scarcity as the unavailability of monolingual as well as cross-language resources in an electronic format, for a particular language. According to [10], scarcity of linguistic resources can be attributed to language diversity and the emergence of new communication media and stylistic trends. In this paper, an effort is made towards quantifying the amount of electronic text needed for the induction of Igbo morphology.

II. LITERATURE REVIEW

Morphological analysis is the lowest level of linguistic analysis and as such, gives impetus for other linguistic abstractions. The Greek root word for morphology is *morphe* which means shape or form. It the arrangement of the parts of an *objec*t and how these parts come together to create a whole; where the objects may either be physical (organism, ecology), social (an organization) or mental (linguistic forms, systems of ideas) [16]. The study of

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

morphology is not only relevant to linguistics but also to such disciplines as geology, physics, botany , biology [15], including cytology and anatomy [9]. [20] offered a more elaborate definition of morphology when he defined morphology as "….the study of more abstract structural interrelations among phenomena, concepts, and ideas, whatever their character might be". (Zwicky, 1966, p.34).

Among linguistic scholars, it is generally agreed that morphology concentrates on the rules of word formation or internal word structure [12], [18], [7], [16] and [19. In addition, morphology studies morphemes, which are the building blocks that constitute a word and the rules of combining these morphemes to form words [2]. In general, the study of morphology is fundamental to linguistic analysis and involves the assessment, investigation or description of the morphological processes or concepts in question.

In Human Language Technology (HLT), morphological analysis gives a foundation to any computational analysis [1]. HLTs such as automatic speech recognizers, automatic speech synthesizers, machine translators, spelling checkers, automatic abstracting, information retrieval, and so on. In this vein, [18] defined morphology as the discovery and description of the mechanisms behind the infinity of words produced from a finite collection of smaller units. [17] and [8] conclude that morphological generalisations include information about sound patterns, and phonological generalisations include information about morphology. This conclusion further shows the significance and applicability of morphological studies.

[4] stated that with as low as 5000 word corpus, morphological analysis can be performed on *Linguistica* - an unsupervised language model based on Minimum Description Length (MDL). This suggests the question; would resource scarcity drastically affect Igbo computational morphological analysis using data-driven approaches?

Theoretically, resource scarcity should not have as drastic an effect on morphology as it should have on syntax and other levels of linguistic analysis. The reason being that morphology is morpheme-based, and the number of morphemes of any language is not only finite but also relatively limited. Hence a modest corpus can produce useful results because morphological rules are limited. In this study, a modest corpus of approximately 30,000 unique words is employed. No study, known to the author has tried to quantify the corpus size for data-driven modeling of the morphology of Igbo language.

III. THE PROBLEM

Computational studies of resource-scarce languages like Igbo have been deterred by lack of large amounts of electronic linguistic data in such languages. A greater challenge is faced when such languages are to be subjected to data-driven computational approaches. Data-driven approaches to learning require very large amounts of electronic linguistic data because as the name implies, such models or approaches are heavily dependent on data, from which learning is made possible. The three approaches to data-driven learning include (i) Supervised learning (ii) Unsupervised learning and (iii) Reinforced learning [14]. There is no known literature to the author where reinforced learning has been applied to natural language learning or understanding, but according to [3], it is an approach to learning that is best suited for game applications; offering a positive reinforcement or a negative reinforcement for every right or wrong performance respectively. This enables the system to *learn* the path that will yield the goal.

Supervised learning, based on [11], [3] and [14] position, provides the opportunity for a system to learn some unique features of a data set from a pre-classified or annotated training data. The system (or classifier) uses the knowledge of these *learnt* features to classify *unseen* data.

The unsupervised learning approach is best described as the approach to language acquisition, manifested by a child. A child does not learn a language by learning the grammatical rules of the language. Rather, a child learns from the many examples which she is able to pick from her environment. Unsupervised learning models are based on the common behaviours of natural language. A major pre-requisite for adopting this approach of learning is the availability of large amounts of data, which provides the many examples from which the system learns some unique features before it is able to make predictions when presented with *unseen* data. Compared to rule-based language models, unsupervised learning models are scalable to other languages, void of any human errors, eliminates the cumbersome task of text annotation and its associated costs, both financial costs and time costs. Unsupervised approaches have become more popular due to these advantages. However, adopting such an approach for computational studies of Igbo, a resource-scarce language is elusive. However, because morphology is a lower level of language analysis than syntactic and semantic analysis, we make an assumption in this study that: Computational morphological studies may not require as much data as would other higher levels of linguistic analysis. This assumption provides the impetus for this study.

IV. PROBLEM SOLUTION

A corpus of 29,191 unique words was extracted from five Igbo texts namely: *Baibụl Nsọ* (Nhazi Katọlik), *Baibụl Nsọ* (Bible Society of Nigeria), *Jụọchi* (novel), *Ogene* newspaper and *Odenigbo* lecture transcripts. We borrow from a linguistic phenomenon of representing all consonants and vowels in our wordlist as *Cs* and *Vs*

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

respectively. We go a step further by appending indexes to every unique consonant or vowel that appears in a word, repeating a particular index if such a consonant or vowel have been earlier encountered in that particular word. To determine the amount of Igbo morphology contained in a unit set of 1000 words, the unique morphological structures rather than the unique word labels will be focused on. For this test, the words in the study corpus were partitioned into 30 subsets of 1000 randomly chosen words each, with the last subset having only 191 words. The words in each subset were converted into word labels and the word labels would be sorted in order to identify unique word label from each cluster of word labels. This was done for all 30 subsets of 1000 words with the aid of a simple Visual basic script on Microsoft Excel spreadsheet.

Table I gives the results of determining unique word labels in the 30 word subsets while Table II shows the results of the estimation of the morphology of Igbo contained in 30 word subsets of 100 words in each, using the morphological structure.

From Table I, all the unit sets of 1000 words gave at least 300 word labels, with only one unit set giving 292. This is a uniform distribution. The least number of unique word labels that can be got from a unit set of 1000 Igbo words is 292 while the highest number that can be got is 336. The 30th unit set of words did not have as much as 1000 words; hence the number of unique word labels is 107, which is far from the mean number of unique word labels, 310.47.

Likewise in Table II, the highest number of morphological structures that can be realized from a unit set of 1000 Igbo words is 48, while the least is 38. The average number of morphological structures from a set of 1000 unique Igbo words is 41.17. Out of the individual sets of 1000 words, 24 of had unique morphological score above 40. Five of the scores were at least 38. The last came from the 30th set which did not have as much as 1000 words

In a second experiment, the wordlist was partitioned into 30 subsets of 1000 randomly chosen words each but the last subset had only 191 words. The words in each subset were automatically converted into word labels and the word labels were sorted. This was done in order to identify unique word labels available in each cluster of 1000 word labels irrespective of the morphological structures of the word labels. The unique word labels found in each unit subset were then accumulated by adding the newly encountered word labels from each subset to the stock of already encountered word labels. At each stage, the number of newly encountered word labels was recorded against the size of the corpus in order to determine the rate at which new word labels emerge in relation to the size of the wordlist. It then became possible to determine the number of word labels that would be contributed by an arbitrary size of wordlist to the existing stock by extrapolation based on the size of wordlist. The graph of word labels was plotted against the number of words and this is shown in Figure I

Table I: No. of unique word labels in 30 sets of 1000 randomly chosen Igbo words

Word Subset	No. of Unique Word Labels
1	319
2	335
3	308
4	313
5	327
6	321
7	316
8	309
9	311
10	324
11	327
12	315
13	306
14	292
15	315
16	315
17	336
18	322
19	318
20	319
21	313
22	328
23	309
24	320
25	312
26	320
27	320
28	308
29	329
30	107
Mean	**310.5**

Table II: No. of unique morphological structures in 30 sets of unique word labels

Word subset	No. of Unique Morphological Structure
1	39
2	42
3	42
4	43
5	42
6	42
7	39
8	48
9	40
10	40
11	38
12	41
13	44
14	41
15	40
16	39
17	40
18	39
19	41
20	40
21	42
22	44
23	48
24	41
25	42
26	40
27	47
28	45
29	40
30	26
Mean	**41.17**

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

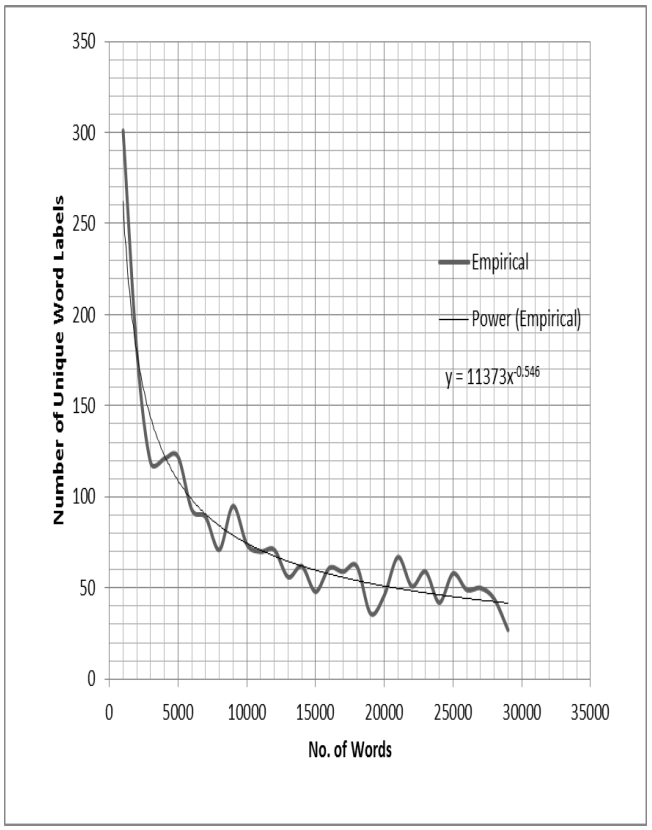

Figure I: Plot of unique word labels contained in sets of 1000 randomly selected Igbo words.

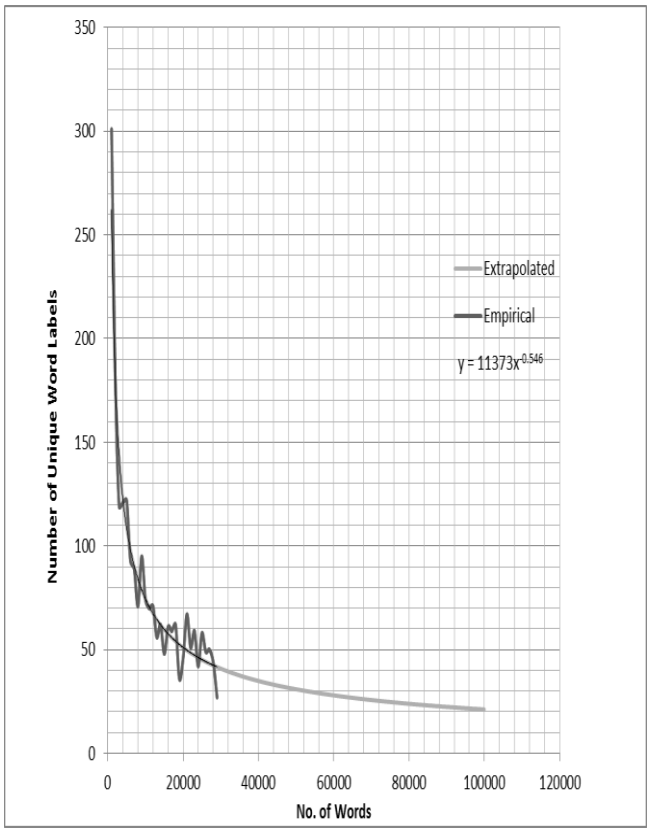

Figure II: Extrapolation of the curve for 100,000 Igbo words

We observed from Figure I that the number of unique word labels contained in a subset of 1000 words decreases as the number of word batches increases. In this test, the first subset of 1000 words produced 301 unique word labels, the second subset of 1000 words produced 180 unique labels, the third subset gave 119 unique word labels and the fourth, 112 unique word labels. By the time the thirteenth subset of 1000 words was analyzed, only 56 unique labels were produced. The twenty-ninth and last subset had only 27 unique word labels. The implication is that the number of unique word labels will continue to decrease until no more unique word labels can be produced. When this state is reached, it could then mean that Igbo morphological rules might have been exhausted. The power function given as: $Y= 11373x^{-0.546}$ was derived as a good representation of the relationship between corpus size and newly encountered word labels.

From Figure II, the extrapolation of the curve of unique word labels against sets of 1000 Igbo words revealed that for a corpus size of 100,000 words, only an additional seven unique word labels would be found. These word labels do not necessarily imply unseen morphological structures or morphological processes.

It was necessary to find out if all morphological rules in Igbo have been exhausted in this present study. In order to achieve this goal, the unique morphological structures in the 30 subset of 1000 word were extracted. Unique morphological structures were then accumulated over these 30 subsets while adding only the yet unseen morphological structures contained in each subset of 1000 words. As we iterated through the 30 word subsets, it was discovered, as Figure I depicts, that the number of morphological structures that are yet unseen was gradually diminishing. Therefore the graph of unique morphological structures in the 30 subsets of word gradually asymptotes following a reverse J curve.

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

V. RESULTS AND DISCUSSION

From Figure II, the reversed J shape of the curve in figure 4.1 indicates that even though the modest corpus for this study may not have exhausted Igbo morphology totally, it manifests an asymptotic behaviour. Hence an extrapolation of the curve can be used to determine how many more unique word labels would be produced for a progressively larger size of wordlist, for example 100,000 words (represented as word labels). The asymptotic behavior of the curve implies that the curve is gradually getting to a very low minimum. The introduction of more words may not change this behavior of the curve much as the curve is terminating, gradually approaching the zero mark.

From Figure II, it was observed that only 7 more unique word labels may be produced if the wordlist had about 70, 000 thousand more words, making it a total of 100,000 words in the corpus. This low number of word labels yet unseen may be due to the sensitivity of the word labels or just a clear indication of the exhaustion of the word labels. It may be possible that these additional unique word labels may not be representative of any new Igbo morphological process since a single morphological process can be implied by more than one morphological structures and word labels.

The asymptotic nature of the curve in Figure I depicts that the modest corpus used in this present study is approaching exhaustion of Igbo morphology. The curve first hit the zero mark at 10,000 words. The implication is that a wordlist of 10,000 words is nearly exhaustive of the morphology of the language it was used to describe. [4] stated that a corpus size of 5000 words could be used on *Linguistica.* At 5000 words, only about four new morphological structures can be elicited. At 6000 words, just one morphological structure is missing. The implication is that a 6,000 wordlist could be used comfortably without much concerns of corpus size, to model a language using unsupervised learning approach.

Applying Hoeffding's inequality; as the theoretical basis of machine learning theory, we substantiate the results of this test. Hoeffding's inequality states that the probability that a certain error is greater than or equal to the mean of the observed minus the mean of the expected outcome of a distribution is less than the negative exponential of that distribution and that certain error. In relation to the present study, it implies that as the words in the wordlist increase, the error reduces.

$$P(|X - E[|X|]| \geq t) \leq 2e^{-2nt^2}$$

where t = a certain error, n = size of sample, $E[|X|] =$ expected value

This inequality of Hoeffding is equal to the reverse J function $Y = 293.35X^{0.612}$. More words in the corpus will diminish the error or strengthen the mean and variance which is bounded at 6000 unique words. As n approaches 6000, number of incorrectly classified words will start to decrease.

VI. CONCLUSION

This study formally showed that resource scarcity does not affect morphological computational studies much as it would other levels of linguistic analysis like syntax or semantics. [4] casually stated that a 5,000 word corpus may be adequate for *Linguistica*. This statement by [4] has been formally strengthened in this present study. The conclusion is that unsupervised morphology induction is possible with a corpus size as low as 29191unique words.

ACKNOWLEDGEMENTS

Sincere appreciations to the Catholic Arch Bishop of Owerri, His Lordship, Dr. Amarachi Obinna for the provision of and permission to use the electronic prints of Odenigbo lecture series.

Appreciations to the entire management and staff of Africana-Fep publishers for granting us the permission to use *Baibụl Nsọ* Nhazi Katọlik for this study and also to Dr. Uchechukwu Chinedu for his support and assistance in making available the electronic version of *Baibụl Nsọ* Nhazi Katọlik available to us.

Finally, Dr. Williams Nwagwu, Dr. Tunde Adegbola and Prof. Obododimma Oha are highly appreciated for their unalloyed support and excellent supervisory roles.

REFERENCES

[1] Barton, G. E. 2004. The computational complexity of a two-level morphology. *DSpace@MIT Artificial Intelligence Lab Publications.* Retrieved August 10, 2010, from http://dspace.mit.edu/handle/1721.1/6427.

[2] Blackburn, P. and Striegnitz, K. 2002. Natural language processing techniques in PROLOG. Retrieved June 16, 2009, from http://cs.union.edu/~striegnk/courses/nlp-with-prolog/html

[3] Coppin, B. 2004. *Artificial intelligence illuminated*. Sudbury, Massachusetts: Jones and Bartlett Publishers

[4] Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *MIT Press Journal* 27. 2: 153-198.

[5] Iheanetu, O. and Adeyeye, M. 2013. Finite state representations of reduplication processes in Igbo. *IEEE XPlore Digital Library.* doi:10.1109/AFRCON.2013.6757772. 1-6.

[6] Iheanetu, O. (2015). Data-driven model of Igbo morphology. PhD. Thesis. Africa Regional Centre for Ingormation Science. University of Ibadan. xv+226pp.

[7] Iloene, G. 2007. Igbo phonology. *Basic linguistics for Nigerian languages teachers*. O. Yussuf. Ed. Port Harcourt: M & J Grand Orbit Communications Ltd and Emhai Press. 163-180

[8] Inkelaqs, S. 2009. The morphology-phonology connection. *Workshop on the Division of Labour between Morphology and Phonology*. *7th Jan., 2009*. Retrieved September 20, 2014, from http://www.uni-leipzig.de/~exponet/Slides/Amsterdam/Inkelas.pdf

[9] Kaplan, R., D. 2001. The science of plant morphology: definition, history and role in modern biology. *American Journal of Botany*, 88: 1711–1741.

[10] Lambert, P., Costa-jussa, M. and Banchas, R. E. 2012. Introduction. *Workshop on Creating Cross-Language Resources for Disconnected Languages and Styles*. 27th May, 2012. Istanbul: Turkey.

[11] Mitchie, D., Spiegelhalter, D. J., and Taylor, C. C. 1994. Machine learning, neural and statistical classification. Retrieved August 10, 2010, from http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf.

[12] Mitkov, R. 2004. Anaphora resolution. *The Oxford handbook of computational linguistics*. R. Mitkov. Ed. NY: Oxford University Press. 266–283.

[13] Ng'ang'a, W. 2010. Towards a comprehensive, machine-readable dialectal dictionary of Igbo. Proceedings *of the 2nd workshop on Africa Languages Technology (AfLAT). 18th May 2010*. Valleta, Malta: AfLAT. 63-67

[14] Nilsson, N. J. 2005. An introduction to machine learning. *Proceedings of the 2005 Symposium on Dynamic Languages. 18th October, 2005*. San Diego, USA: ACM Digital Library. 1-10

[15] Pelczar, M., Pelczar, R. and Steere, W. C. 2012. Botany. *Encyclopeadia Britannica*. Retrieved October 20, 2012, from http://www.britannica.com/EBcheked/topic/75034/botany

[16] Ritchey, T. 2011. General morphological analysis. Retrieved October 20, 2012, from http://www.swemorph.com/gma.html.

[17] Spencer, A. 1991. *Morphological theory.* Oxford: Basil Blackwell

[18] Trost, H. 2003. Morphology. *The Oxford handbook of computational linguistics.* R. Mitkov. Ed. New York: Oxford University Press. 25-47.

[19] Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O., Kurimo, M. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues (TAL)* 52.2: 45-*90*.

[20] Zwicky, F. 1996. Discovery, invention, research - Through the morphological approach. New York: Macmillian company.

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