Chapter 11
Addressing the Challenges of Igbo Computational Morphological Studies Using Frequent Pattern-Based Induction

Olamma U. Iheanetu and Obododimma Oha

Abstract Computational studies of Igbo language are constrained by non-availability of large electronic corpora of Igbo text, a prerequisite for data-driven morphological induction. Existing unsupervised models, which are frequent-segment based, do not sufficiently address non-concatenative morphology and cascaded affixation prevalent in Igbo morphology, as well achieving affix labelling. This study devised a data-driven model that could induce non-concatenative aspects of Igbo morphology, cascaded affixation and affix labelling using frequent pattern-based induction. Ten-fold Cross Validation (TCV) test was used to validate the propositions using percentages. An average accuracy measure of 88% was returned for the developed model. Ten purposively selected Igbo first speakers also evaluated samples of 100 model-analysed words each and the mean accuracy score of 82% was recorded. We conclude that morphology induction can be realized with a modestly sized corpus, demonstrating that electronic corpora scarcity does not constrain computational morphology studies as it would other higher levels of linguistic analysis.

Keywords Computational morphology · Frequent pattern-based morphology
Igbo computational morphology · Igbo morphology · Rule-based learning
Morphology induction

11.1 Introduction

Statistical morphological analysis of natural languages is a more viable option than traditional rule-based approach, especially for resource-scarce languages [2] due to the fact that these set of languages are less studied or under-studied. Statistical approaches to learning are largely data-driven. A very favoured statistical learning

O. U. Iheanetu (✉)
Department of Computer and Information Sciences, Covenant University, Ota, Nigeria
e-mail: olamma.iheanetu@covenantuniversity.edu.ng

O. Oha
Faculty of Arts, Department of English, University of Ibadan, Ibadan, Nigeria

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-13-2191-7_11

method is the Unsupervised Learning Method (ULM). ULMs require large volumes of data to train the learner in other to make accurate classifications of unseen objects. Unfortunately, large volumes of electronic data are a scarce commodity for resource-scarce languages such as Igbo, and thus a major set-back in adopting data-driven approaches for computational studies [15].

The underlying fact is that most of the existing models of morphology induction are designed based on inherent behaviour of language which are arbitrary occurrence of word segments and frequent occurrence of word segments. This results in somewhat universal models although some uncommon peculiarities of some languages can render such models inappropriate/ insufficient for those languages. However, [5] argued that existing ULMs are not sufficient for Bantu languages. Igbo may not be a Bantu language but [3] believes that it belongs to the same language phylum as Bantu languages, causing it to share some similarities with them. Creutz [4] are of the opinion that ULMs are unsuitable for languages with sparse linguistics data in computer readable form.

Although some of the existing ULMs correctly analyse inflected Igbo words, results show that these models cannot sufficiently breakdown words produced from reduplication, circumfixation, compounding and interfixation morphological processes in Igbo. A further complication results from the highly agglutinative nature of Igbo, according the language the capability of having as high as four cascades of affixes, thus posing a challenge for the analyses of all affixes [15]. These situations and circumstances frustrate the morphological induction of Igbo and other resource-scare languages, which are mostly African languages. Proffered methods, as suggested by [16], should take into cognizance the challenge of scarcity as well as other peculiar characteristics of the Igbo morphology.

The underlying language behaviour for inflection is frequent word segments prevalent in a given corpus, hence most existing ULMs are frequent segment-based. As established in [16], frequent pattern-based ULMs could work better for most derived words, especially if they share some similarities with Igbo.

11.2 Previous Literature

Most unsupervised morphology models carter for languages having orthographic inconsistencies which necessitate adjustments in the heuristics used. Such adjustments may not be necessary for Igbo because of the high level of regularity in its orthography [15]. A consensus in ULM is that affix classification is still elusive [13]. Arbitrary Character Assumption (ACA) and Frequent Flyer Assumption (FFA) was postulated by [12] for the extraction of morpheme boundaries, on the premise that words are normally arbitrarily occurring segments. The frequently occurring segments which are equi-length with the segments that occur arbitrarily have a high probability of being affixes. Goldsmith [11] identified morpheme boundaries using the principle of Minimum Description Length (MDL) which describes the morphology that offers the least minimum description length. Harris [14] is a foundational

study in unsupervised morphology induction that bootstraps the problem of identifying morpheme boundaries using a heuristic approach; *successor frequency*. These ULMs are only representative of the general approaches to the unsupervised learning of morphology which are mostly frequent-segment based.

A brief background on Igbo language shows that Igbo is a tonal language spoken in South-east part of Nigeria [16] having an approximate thirty dialects [24, 26] but Standard Igbo is widely spoken and understood, and as well generally accepted among Igbo speakers. Igbo language is a member of the West Benue-Congo languages [3], formerly classified under the Niger-Congo Kwa language family; a language family characterized by two tones (high and low) and a down step that apply different meanings to the same set of phones [9]. Igbo exhibits a rich agglutinative morphology [23, 26] and Igbo verbs also inflect for aspect [26]. Igbo features a wide variety of highly productive concatenative and non-concatenative morphological processes. Owing to the agglutinative nature of Igbo morphology, cascaded affixation, a highly productive morphological process, becomes a common occurrence.

Most Igbo morphological processes can be generalized as affixation. An affix can come before the root word (prefix), or in between (infix or interfix, depending on where it splits the root word) or after the root word (suffix). Ndimele [22] noted that the position and function of an affix when it is attached to a root is definitive of the category of that affix. Hence there exist prefixes, suffixes, infixes, interfixes, circumfixes, superfixes and suprafixes in the positional classification of affixes. For Igbo language, prefixation, suffixation, interfixation [7, 20], superfixation and circumfixation apply [20]. Non-concatenative morphological operations involve root modification [25] truncation, subtractions, conversion and transfixation, which according [18] cannot be resolved by recurrent partials. Although most Igbo morphological processes are concatenative, some manifest majorly as non-concatenative morphology. These include interfixation, circumfixation and compounding.

11.3 The Problem

The cost of rule-based morphology models is very high in terms of data annotation cost, labour and the time involved. In addition, RBMs are subject to human error. The demands of UML approach are unattainable for resource-scarce languages like Igbo. ULMs require many examples to learn from. These examples are offered by the availability of very large amounts of electronic data or corpora. Igbo at present has very sparse linguistic data in computer readable form available for computational studies.

Simple affixations in Igbo 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. While experimenting with *Linguistica*, it was observed that *Linguistica* could not accurately analyse derived words like compounds and words having multiple extensional

suffixes. This was quite explainable judging that *Linguistica* was tested on English and French corpora; languages that rarely exhibit multiple suffixation.

Non-concatenative processes like reduplication and compounding pose a challenge for existing unsupervised learning models. This is because the affixes of 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 may 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.

Igbo, Dghweɗe [6] and Yoruba languages [1] exhibit interfixation; a rare morphological behaviour. Interfixes may not feature as frequently occurring segments and therefore may not be captured by frequency-based morpheme boundary identification methods. Typical Igbo interfixes are merely Igbo phonemes represented in writing as *l, m*, *r*, and so on, stand between two copies of a stem. Some morphological processes of interest in this study include the following

Extensional suffixation	Where *wa, zie, nụ and kwa* are extensional suffixes that extend the meaning of the verb. See examples below
kpachapụkwaranụ	kpa-cha-pụ-kwa-ra-nụ
laghachikwaara	la-ghachi-kwa-a-ra

Reduplication—full reduplication is achieved by duplicating the stem word, although some scholars have argued that because the words do not undergo any morphological processing, it is mere repetition. Another way of realizing gerunds in Igbo is by reduplicating verb stems having the—CV structure and prefixing a harmonizing ọ or ò vowel and appropriate reduplicating vowel—i, i, u, or ụ. This is known as partial reduplication

ọ + lụ (marry) + lụ → ọlụlụ (marriage)
ò + zù (steal) + zù → òzùzù (act of stealing)

Partially reduplicated word is described as a word having a prefix and two identical roots, or a word having a prefix and two roots that differ only in the vowel of one of the roots, only if such differing vowel is a member of the list of valid reduplicating vowels—*i*, ị, *u*, and *ụ*. $v \in \{i, ị, u, ụ\}$.

Given an Igbo root verb $R = r_1, r_2, \ldots, r_n$, a partially reduplicated word $w = pr_1vR$, such that $p \in \{o, ọ\}$, $\{i, ị, u, ụ\} \in v$ and $R \in VocR$ where $VocR$ is a set of Igbo verb roots.

ọ-gb-ụ-gbọ	reduplicating vowel = ụ
o-gb-u-gbu	reduplicating vowel = u
o-l-i-lo	reduplicating vowel = i
ọ-g-ị-ga	reduplicating vowel = ị

Interfixation—This is a word that has an affix (interfix) which splits the word into two identical segments.

Given a vocabulary Voc and words $w \in Voc$, if w can be broken into segments such that $w = w_1, s_1, w_1$ there is a probability $P_{infx}(w)$ that s_1 is an interfix between

two segments of w_1 if $w_1 = w_1$ and $s_1 \in \{da, gh, l, m, r\}$. Where $P_{infx}(w)$ = Probability that w is an interfixed word [15]. In a way, interfixation can be viewed as full reduplication having an interfix between the words, if the interfix belongs to the list of true interfixes comprising *l*, *m*, *r*, *da* or *gh*.

ede-m-ede interfix = m
ogo-l-ogo interfix = l
egwu-r-egwu interfix = r

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 together to form a word.

obi + ọma → obiọma (good heart)
chi + ma → chima (God knows).

Given a vocabulary *Voc* and words $w \in Voc$, if w can be broken into segments such that $w = w_1, w_2, \ldots, w_n$, there is a probability $P_{(cpd)}(w)$ if $w_1, w_2, \ldots, w_n$ and $w_1 \neq w_2 \neq w_3, \ldots \neq w_n \in Voc$. Where $P_{cpd}(w)$ = Probability that w is a compound word. [15].

Circumfixation—The phenomenon of Igbo circumfixation morphological process is still very controversial and hence not well studied. It involves two disparate affixes one at each end of stem [21]. The vowel-syllabic nasal circumfix and the vowel-incorporated preposition circumfix exist [21]. However, no rule(s) that governs the choice of the initial vowel or syllabic nasal was presented by [21]. Given a vocabulary *Voc* and words $w \in Voc$, if w can be broken into segments such that $w = w_1, s_1, w_2$ there is a probability $P_{cfx}(w)$ if $w_1, w_2 \in Voc$ and $s_1 \in \{m, n\}$. Where $P_{cfx}(w)$ = Probability that w is a circumfixed word [15].

n-dụ-m-ọdụ circumfix = n-m → (advice)
e-zu-m-ike circumfix = e-m → (rest/holidays)

The notion of resource scarcity in the scientific study of a language is yet to acquire a clear definition [15]. The term is used interchangeably with other terms such as resource-starved, under-resourced, resource scarce, less studied, least developed, under developed, under resourced, and so on. In computational linguistics, these terminologies have been used to describe languages that have insufficient or no electronic texts in written or spoken form which are readily available for computational studies in that language [15]. The concept of resource scarcity needs to be formally addressed and properly re-addressed given that linguistic resources have become incredibly valuable and data availability plays a fundamental role in NLP applications and technologies [17]. Any language that cannot boast of large amounts of appropriately diacritized electronic spoken or written texts, which is easily available for use, is a resource-scarce language. For such languages, computational studies become cumbersome because computer readable texts have to be first generated and developed. The general impression is that resource-scarce languages, like Igbo, cannot be subjected to data-driven morphological analysis, as earlier echoed by [4].

Lack of diacritized texts—Most of the available computer-readable texts in Igbo are not diacritized, and when they are, the tone marks are either incomplete, incorrect, or irregular with the standard tone-marking convention. Some claim it is easier to publish text without the necessary diacritics, while others submit that the technology to include those diacritics do not exist or are not available. This challenge presents ambiguity issues. For morphological studies, the impact of this may not be as drastic, but for syntactic, semantic and higher levels of linguistic analysis, it may be catastrophic. Homonyms are words that share the same spellings but have different meanings based on the respective tone-marks on the words. Hence, without the appropriate tone-marks, it would be almost impossible to analyse homonyms appropriately. The following nouns are known homonyms in Igbo.

akwa → (cry) [High – High]; àkwà → (bed) [Low – Low]
àkwa → (egg) [High – Low]; akwà → (cloth) [Low – High]

11.4 Proferred Solution

Data Source—Data for this study was got from *Baibụl Nsọ* (*Nhazi Katọlik*); a current Igbo version of the holy Bible imprimatured by Bishop A. K. Obiefuna in 2000, an electronic version of Baibụl Nsọ by Bible Society of Nigeria, a story book; *Jụọchi*, one edition of the now defunct *Ogene* newspaper, and two years of Odenigbo lecture transcripts. The hardcopy story book and the newspaper were typed in order to have an electronic copy while *Baibụl Nsọ* (*Nhazi Katọlik*) was scanned with Fine reader, an optical character recogniser so as to make available to the study, an electronic copy of the text. Although not all of these Igbo texts were written in standard Igbo, all texts were consistent with the Ọnwụ orthography; the standard orthography. The resulting wordlist after data cleaning and processing was a 29,191 wordlist. The words in the wordlist were converted to strings of Cs and Vs and then these *word label* strings were manually segmented according to their respective inherent morphological structures.

Frequent Pattern Theory—[16] have established that non-concatenative and cascaded affixation in Igbo language can be analysed by ULM that are frequent pattern-based as against frequent segment-based. They developed the Frequent Pattern Theory (FPT), on the premise that the segments of some derived Igbo words do not manifest as frequent segments in a wordlist. This study applied frequent pattern mining for morphological analysis of especially non-concatenative aspects of Igbo morphology. The introflected segments of words emanating from morphological processes like partial reduplication, circumfixation, and interfixation cannot be identified by any morphological analysis model that is frequent segment-based [16]. A linguistic phenomenon was adopted, where words in a wordlist are represented as a combination of Cs and Vs, having number subscripts for unique identification of the letters. C represents consonants and V represents vowels. A Python program was written to automate the transcription of our wordlist into strings of Cs and Vs. Hence, C_0, C_1, C_2, is a substitute for the first, second and third consonants of a given word.

Table 11.1 Patterns of identified Igbo morphological processes

S/no	Morphological process	Pattern
1	Unbound morpheme	Root
2	Prefixation	Prefix-Root
3	Suffixation	Root-Suffix
4	Interfixation	Root-Interfix-Root
5	Circumfixation	Prefix-Root1-Cmfx-Root2
6	Partial reduplication	Prefix-MRoot-Root
7	Full reduplication	Root-Root
8	Compounding	Root1-Root2

MRoot—Modified Root, Cmfx—Circumfix

Likewise, V_0, V_1, V_2 stands for the first, second and third vowels that occur in a given word [16].

The identified morphological processes in this study and the associated patterns embedded in the words that they produce are shown in Table 11.1.

The first test carried out tried to determine if there are embedded features in Igbo words which can be used to induce the non-concatenative aspects of Igbo morphology. A PROLOG program was used to extract words from the wordlist according to the individual patterns of the identified Igbo morphological process.

Results showed that identified Igbo morphological processes clearly manifest distinguishable patterns which are stamped on the words produced by these morphological processes. It was observed that when the pattern of a given morphological process was used to extract words from the wordlist, most of the words extracted conformed to the pattern of the appropriate morphological process being tested. However, this observation was not uniform. Compound words was the most challenging word pattern to describe because the pattern description was rather fluid, extracting verbs like *abanye*, made up of two different words.

This test gives credence to the fact that Igbo words have embedded patterns, owing to the morphological process that produced such a word, which can be deployed as a feature for inducing that word. Some of the confused words from this test are shown in Table 11.2.

In another test, a comparison of the *word label* clusters and the word patterns which words belonging to the same cluster manifest was undertaken. A maximum of twenty words each were randomly selected from each class of word labels to verify if word labels as textual proxy of patterns were productive for capturing word patterns. Standard accuracy measure based on true positives and false positives was calculated.

From Table 11.3, results showed that some word label clusters like $V_0C_0C_1V_1$, $C_0C_1V_0C_0V_1C_2V_1$, and $V_0C_0V_1C_0V_1V_2$, for PRt1, Cmfx1 and PRedp2 morphological processes respectively, did not have as many as 20 words. A high level of consistency between word labels and their structures was observed. Some word labels did

Table 11.2 Morphological processes versus corresponding extracted words

S/no	Morphological processes being tested				
	Full reduplication	Circumfixation	Interruption	Compounding	Partial reduplication
1	*ksam-ksam	**i-ds-m-miri	**kwa-do-kwa	**aba-cha	*n-gi-ge
2	*ngwa-ngwa	**n-gwa-mma	**ara-ch-ara	*aba-nti	**i-iu-tu
3	*kpom-kpom	*n-si-m-ike	*ede-m-ede	*nwa-nza	**n-ga-gide
4	*wara-wara	*a-wa-m-anya	*iri-gh-iri	*nwa-mmefu	*o-di-di
5	*bara-bara	*e-nye-m-aka	*oko-m-oko	*odo-mmiri	*o-wu-wu

*True positives, ** False positives

not capture their morphological processes well. For example, $V_0C_0V_0C_1V_0C_0V_0$ clusters interfixed words better than $C_0V_0C_1V_0C_0V_0$, judging from the accuracy scores of 75% and 2% respectively. Prefixation, compounding and suffixation morphological processes each had an accuracy of 100%, 95% and 85% respectively, demonstrating that some word labels cluster words of a particular morphological process better than other word labels.

The third test was undertaken to understand how accurately word labels reflect morphological structure when used as textual proxies of word patterns. This study proposes that the use of word labels as a textual representation of the structural patterns of Igbo words constitutes a productive feature for the induction of the non-concatenative aspects of Igbo morphology. A model of Igbo morphology was developed using a determined textual proxy for the representation of the structural patterns of Igbo words, that is the morphological structure of a word. The model was evaluated in order to verify the accuracy of the model output [15].

According to [27], the ⌊ random i.i.d for this study is 29191. Word labels automatically clustered words according to their morphological structures based on a manually classified and segmented word label table. Using this table, the model learned the morpheme boundaries based on word patterns as captured by the segmented word labels presented in the manually segmented word label table. After the training, the developed model was able to predict segmentation of unseen words based solely on the word labels, thereby demonstrating the ability to generalize. The model also suggests the morphological structure of the segmented words, and this information doubles as the affix label of that word.

The Ten-fold Cross Validation (TCV) test, implemented in Visual Basic was used to evaluate the model's accuracy at predicting morpheme boundaries of unseen words. The TCV test involves four main steps namely:

- Partitioning the study wordlist into 10 data subsets
- Training the model using nine of the ten data subsets and testing it with the remaining one data subset to determine the accuracy of the model

Table 11.3 Accuracy of the use of word labels to cluster words produced by the same morphological process

Morphological processes	Intfx1	Intfx2	PRt1	PRt2	Cmfx1	Cmfx2
Word labels	$C_0V_0C_1V_0C_0V_0$	$V_0C_0V_0C_1V_0C_0V_0$	$V_0C_0C_1V_1$	$C_0C_1V_0C_2V_1$	$C_0C_1V_0C_0V_1C_2V_1$	$C_0C_1V_0C_0V_1C_2V_2$
True positive	4	15	4	20	3	11
False positive	16	5	0	0	4	7
Accuracy	0.2	0.75	1	1	0.43	0.61

Table 11.4 Resuit of the TVC test

Tests	True positives (1)	False positives (2)	Empty cells (3)	Uncertain (4)	Total = {(1) + (2) + (3) + (4)}	Accuracy score = (1)/Total
Test 1	282	14	4	0	300	0.94
Test 2	267	22	11	0	300	0.89
Test 3	270	22	7	1	300	0.90
Test 4	267	24	9	0	300	0.89
Test 5	261	28	10	1	300	0.87
Test 6	273	19	7	1	300	0.91
Test 7	267	24	9	0	300	0.89
Test S	258	30	12	0	300	0.86
Test 9	267	24	9	0	300	0.89
Test 10	243	41	13	3	300	0.81
Mean	265.5	24.8				**0.88**

- Repeating step 2 above nine more times, using different combinations of the training and testing data subsets each time
- Computing the mean accuracy of the ten tests.

The wordlist was divided into ten subsets, giving an approximate of 2919 words in each subset. The morphological analyser model was then trained ten times with nine different data subsets and tested ten times with one data-subset, which was different for each test. The training required the model to read a word, convert it to a corresponding word label, and segment the word label based on the manual classification table. The output of the training is a series of segmented word labels. For the testing, the model reads in a word, determines an appropriate segmentation of the word based on the segmented word label which is the output of the training, and automatically segments the word based on a determined word label segmentation from the training.

The accuracy of each of the ten tests was calculated based on true positives and false positives. The average accuracy score of the ten tests was then determined and used as the overall accuracy of the analyser model. Table 11.4 shows the test output from the developed analyser model, including the morphological structure (affix label) and correctness of the segmented words.

The mean score of the TVC test gave 88% accuracy for the developed model, an indication that the developed model could segment Igbo words based on their embedded patterns.

To strengthen this claim, 10 Igbo first language speakers consisting of 3 Igbo primary school teachers, 3 Igbo linguists and 4 Igbo postgraduate students, who were purposively selected, evaluated the output of the model. The assessors manually cross-checked 100 samples each of the model output which were randomly selected. Based on the number of "Yes" from each assessor, an accuracy score was calculated. Table 11.5 captures this evaluation. More assessors would have been sought but for

Table 11.5 Model accuracy score as assessed by 10 Igbo native speakers

Assessors	Yes	No	Accuracy
1	95	5	0.95
2	76	24	0.76
3	74	26	0.74
4	74	26	0.74
5	74	26	0.74
6	94	6	0.94
7	92	8	0.92
8	89	11	0.89
9	63	37	0.63
10	89	11	0.89
Average			**0.82**

the challenge of finding native speakers who understand morphological segmentation in a non-Igbo speaking location.

Thus, the mean accuracy of the model as evaluated by 10 native Igbo speakers was 82%, further buttressing the fact that pattern segments can be used to induce non-concatenative aspects of Igbo morphology.

According to [10], *Linguistica* achieved an accuracy score of 72% on the first 200,000 and 300,000 words of the Brown corpus; Syromorph developed by [19] had an accuracy score of 90% while [8] model had 98% accuracy for Arabic words and 85.3% for English words. Based on [27], a measure of the analyser model's accuracy is implied in the loss or discrepancy measure between the response y of the supervisor and that of the learning machine using the same input as shown below.

$$l(y, (f(x, \alpha))) = \left\{ \begin{array}{l} 0 \text{ if } y = f(x, \alpha) \\ 1 \text{ if } y = f(x, \alpha) \end{array} \right\}$$

where y = response of the supervisor, x = a certain input, and $f(x, \alpha)$ = response of the learning machine.

If $f(x, \alpha) = y$, then, the output of the model is same as the linguist (supervisor) and invariably, the accuracy score of the model is quite high. If $f(x, \alpha) \neq y$, then the guess or prediction of the learner model is not very close to that of the supervisor, in this case the linguist, and may imply a low accuracy score. This is shown in Table 11.5. Therefore, every instance ("Yes") an assessor agrees with the segmentation offered by the analyser model $f(x, \alpha)$, the output of the developed analyser model equals to zero because y, which is the word segmentation as suggested by Igbo linguists, is the same as the word segmentation that the analyser model predicted. That is, $f(x, \alpha) = y$, implying that the discrepancy between $f(x, \alpha)$ and y is zero. However, for every 'no' the implication is that the response of the analyser model, $f(x, \alpha)$ equals one because y is different from the word segmentation predicted by the analyser model.

That is $f(x, \alpha) \neq y$, implying that there is some discrepancy between $f(x, \alpha)$; the guess of the analyser model and y; the prediction of Igbo native speakers.

By representing words with their patterns, 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, could possibly contain some morphological information. Such patterns that are driven by a morphological activity(ies) occur more frequently in the corpus than some others. Hence, patterns within a cluster share some commonalities like similar segment patterns. This could be as short as a single character or as long as three or four characters.

11.5 Limitations

A major setback encountered in this study is the unavailability of computer readable Igbo texts. Unfortunately, few electronics Igbo texts that were available lacked the necessary diacritics that would qualify them as valid Igbo words. Baibụl Nsọ (Nhazi Katọlik) was scanned using Fine Reader, an Optical Character Recognizer (OCR). Fine Reader either inconsistently replaced the UTF-8 characters with invalid Igbo characters, some illegible characters and symbols, as well as numerals or it completely deleted the sub dots from their characters. As a result, some of the scanned words were totally incomprehensible. This challenge prolonged the data cleaning stage because the process could not be automated due to inconsistencies and irregularities in the data.

11.6 Conclusion

The results documented in this study give credence to the fact that word labels depict patterns embedded in words and also present as a valuable feature for clustering both concatenative and non-concatenative Igbo words according to their morphological structures without supervision. The FPT can be viewed as a counterpart to the frequent segment postulates and assumptions on which the various approaches to the induction of concatenative morphology found in the literature have so far been based on. Using FPT, the study established a nexus between Igbo words and Igbo morphological processes, via word patterns, word labels, morphological structures and ultimately morphological processes.

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. Formally strengthening [11] claim that a 5,000-word corpus may be adequate for *Linguistica*. Igbo morphology was induced with a wordlist of 29,191 based on frequent pattern-based induction. This refutes [4] assertion that unsupervised learning was unsuitable for languages with sparse data. Theoretically, resource scarcity should not have as drastic an effect on morphology as it should have on syntax and other levels of linguistic analysis.

Hammarström [12] and Hammarström and Borin [13] assert that unsupervised induction of non-concatenative morphology may not be achieved in the near future because such an algorithm runs in exponential time, most unsupervised induction model of concatenative morphology run in quadratic time. In this study, FPT was applied to the induction of non-concatenative aspects of Igbo morphology. The proposed FPT is novel and clusters non-concatenative Igbo words in linear time.

From observations, length and repetition are two key features that interplay in FPT. One affects the other inversely. The longer a label the higher the choice of the symbol to be added which means that such labels have a higher probability of being a unique word label. Conversely, for word labels whose length result from cascaded suffixes, the choice of symbols for the word labels are constrained because there is repetition. Therefore, longer word labels tend to have more frequent flyers. Hammarström [12] ACA fails here. This is because the segments that form a cascaded suffixation word are frequent in the wordlist.

The idea that the distribution of certain characters in a word establishes patterns, stands to reason that the distribution of characters may have morphological significance. However, if a character is repeated in a word, such a repetition may not only be due to morphology, it may also be due to chance. Fortunately, due to the equi-probability of occurrence of events that are due to chance as expressed in the ACA proposed by [12], using the FPT, we expect that the frequency of repetition of characters due to chance is expected to be relatively lower than the frequency of the repetition of characters due to morphology.

Acknowledgements I acknowledge the entire management and staff of Covenant University, Ota, Nigeria for financing the publication of this material.
The tests and results presented here are those contained in an unpublished thesis of Iheanetu (2015). I acknowledge all my supervisors for their immense contributions to this study.
The Catholic Arch Bishop of Owerri, His Grace, Dr. Amarachi Obinna is highly appreciated for the release and permission to use electronic prints of Odenigbo lecture series. Finally, I acknowledge the management and staff of Africana-Fep publishers for the permission to use Baibụl Nsọ Nhazi Katọlik for the purposes of this study.

References

1. O. Awobuliyi, *Eko Iseda-Oro Yoruba* (Montem Paper Backs, Akure, Ondo state, 2008)
2. K.R. Beesley, L. Karttunen, *Finite State Morphology* (CSLI Publications, Stanford, United States of America, 2003)
3. R. Blench, *Atlas of Nigerian languages*, 3rd edn. (2012) Retrieved from 9 June 2015, www.rogerblench.info/Language/Africa/Nigeria/Atlas%20of%20Nigerian%20Languages-%20ed%20III.pdf
4. M. Creutz, Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition. Ph.D. Thesis, Computer and Information Science Department, Helsinki, University of Technology, Espoo (2007), xi+110 pp.
5. G. De Pauw, G. De Schryver, Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. Lexikos Afr. Assoc. Lexicogr. (AFRILEX)-reeks Series **18**, 303–318 (2008)

6. N. Emenanjo, The interfix: an aspect of universal morphology. J. West Afr. Lang. XII **1**(1982), 77–88 (1982)
7. N. Emenanjo, *Elements of Modern Igbo Grammar* (University Press Limited (UPL), Ibadan, 1987)
8. M.A. Fullwood, T.J. O'Donnell, Learning non-concatenative morphology, in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 8 August 2013, Sofia, Bulgaria, pp. 21–27
9. Gale Group Inc., Igbo. Junior Worldmark Encyclopedia of World Cultures (1999). Retrieved from 10 August 2010, Encyclopedia.com: http://www.encyclopedia.com/doc/1G2-3435900354.html
10. J. Goldsmith, Unsupervised learning of the morphology of a natural language. Mass. Inst. Technol. (MIT) Press J. **27**(2), 153–198 (2001). https://doi.org/10.1162/089120101750300490
11. J. Goldsmith, An algorithm for the unsupervised learning of morphology. Nat. Lang. Eng. **1**(1) (2005). Cambridge University Press. Retrieved from, http://hum.uchicago.edu/~jagoldsm/Papers/algorithm.pdf
12. H. Hammarström, 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 (2009), 284 pp.
13. H. Hammarström, L. Borin, Unsupervised learning of morphology. MIT Press J. **37**(2), 309–350 (2010)
14. Z. Harris, *Morpheme Boundaries within Words: Report on a Computer Test*. Transformations and Discourse Analysis Papers (1967), p. 73
15. O.U. Iheanetu, Data-driven model of Igbo morphology. Doctoral thesis, Africa Regional Centre for Information Science, (ARCIS), University of Ibadan, Nigeria (2015) 284 pp.
16. O.U. Iheanetu, O. Oha, Some salient issues in the unsupervised learning of Igbo morphology, in *World Congress on Engineering and Computer Science 2017*. Lecture Notes in Engineering and Computer Science, 25–27 October 2017, San Francisco, USA, pp. 389–393
17. P. Lambert, M. Costa-jussa, R.E. Banchas, Introduction, in *Workshop on Creating Cross-Language Resources for Disconnected Languages and Styles*, 27th May 2012. Istanbul, Turkey (2012)
18. J.J. McCarthy, A prosodic theory of nonconcatenative morphology. Linguist. Inq. **12**(3), 373–418 (1981)
19. P. MacClanahan, G. Busby, R. Haertel, K. Heal, D. Lonsdale, K. Seppi, E. Ringer, A probabilistic morphological analyser for Syriac, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2010), pp. 810–820
20. B.M. Mba, *A Minimalist Theory and Application to Igbo* (Catholic Institute for Development Justice and Peace (CIDJAP) Press, Enugu, 2011)
21. B.M. Mba, Circumfixation: interface of morphology and syntax in Igbo derivational morphology. IOSR J. Humanit. Soc. Sci. (JHSS) **5**(6), 1–8 (2012)
22. O.M. Ndimele, *A First Course on Morphology and Syntax* (Emhai Printing and Publishing Company, Port Harcourt, 1999)
23. B.I.N. Osuagwu, G.I. Nwaozuzu, G.A. Dike, V.N. Nwaogu, L.C. Okoro, *Fundamentals of linguistics* (Colon Concept Ltd, Owerri, 1997)
24. L.M. Paul, G.F. Simons, C.D. Fennig (eds.), *Ethnologue: Languages of the World, Eighteenth Edition* (SIL International, Dallas, Texas, 2015). Retrieved from 20 June 2015, http://www.ethnologue.com/language/ibo
25. A.K. Simpson, The origin and development of nonconcatenative morphology. Ph.D. Thesis. Graduate Division of the Department of California (2009), 194 pp.
26. University of California, Los Angeles (UCLA) Language Materials Project 2009. Igbo. UCLA Language Materials Project. Retrieved from 20 October 2010, http://www.lmp.ucla.edu/Profile.aspx?LangID=13&menu=004
27. V.N. Vapnik, An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)