Machine Learning Research
2022; 7(1): 8-14
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20220701.12
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

Enhanced Machine Learning Algorithm for Translation of English to Igbo Language

Orji Ifeoma Maryann[1, *], Sylvanus Okwudili Anigbogu[1], Ekwealor Oluchukwu Uzoamaka[1], Chidi Ukamaka Betrand[2]

[1]Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria

[2]Department of Computer Science, Federal University Technology of Owerri, Owerri, Nigeria

Email address:

oma4goodness@gmail.com (O. I. Maryann), so.anigbogu@unizik.edu.ng (S. O. Anigbogu), ou.ekwealor@unizik.edu.ng (E. O. Uzoamaka), chidiberand@futo.edu.ng (C. U. Betrand)

[*]Corresponding author

To cite this article:

Orji Ifeoma Maryann, Sylvanus Okwudili Anigbogu, Ekwealor Oluchukwu Uzoamaka, Chidi Ukamaka Betrand. Enhanced Machine Learning Algorithm for Translation of English to Igbo Language. *Machine Learning Research*. Vol. 7, No. 1, 2022, pp. 8-14.
doi: 10.11648/j.mlr.20220701.12

Received: March 31, 2022; **Accepted**: April 20, 2022; **Published**: May 12, 2022

Abstract: The research has designed a system that has done a morphological analysis of noun phrase and compound verb. Also, the system designed will translate a whole sentence indicating which words are noun and verb in it. Clustering was an unsupervised technique which was used to translate from English to Igbo language. In order to obtain our desired motives, object oriented analysis and design methodology were used. The system has been developed to make Igbo populaces to communicate well with most spoken English country along the global and strengthen the Igbo's pole position in terms of research excellence. Furthermore, it will remove barriers to international trade that will keep Igbo small and medium companies from obtaining their complete economic standard by making ways into markets in other continents beyond our own. These goals lead us to develop a machine learning algorithm for translation of English into Igbo language. Machine learning algorithm for translation of English to Igbo language is the missing puzzle that will bring businesses to the people's doorsteps. Besides, people that refused to acquire Igbo language are denying themselves pleasure of direct and unfiltered communication with others and thereby imprisoned themselves with the thrown of language.

Keywords: Machine Learning, Algorithm, Clustering, Igbo, English, Compound Verb, Noun Phrase, Translation

1. Introduction

Nowadays, Igbo language has about 24 million speakers; many of them reside in Nigeria. The Igbo language was put in writing in Latin script to satisfy scientific wants by British colonialists. The language has more than 20 dialects. It was owned by the Benue-Congo group of the Niger-Congo language family. In 1972, Standard literary language was formed based on Owerri and Umuahia dialects [16]. Researchers claimed that the language originated around the 9th century B. C. [1] Niger and Benue rivers are said to be origin of Igbo language. Igbo is one of the official languages of Nigeria. It is spoken in the Southern Delta States, Abia, Anambra, Ebonyi, Enugu, and Imo, likewise in the North East of the Delta State and South East of the Rivers State. In the South – East States: Abia, Anambra, Ebonyi, Enugu, and Imo, Igbo is seen as the major language of trade and commerce. It is used in mass media communication such as radio and television in the South East region. Igbo is one of the oldest surviving language spoken by gods and men alike. Actually, there is over-weighing evidence in the Adam Trilogy that every language keeps evidences of cultural and historical experiences it has lived through in the course of millennia. There are many evidences that Igbo was the language spoken by God when he 'spoke' create on into being and that it was the language spoken by the first Homo sapiens family – Adam's.

Multilingualism is at the heart of the Igbos idea and a true remedy to assist fight obstacles, promote quislingism and generate more cultural awareness for a strong Igbo region in its

diversity. The disintegrated disequilibrium between English and the smaller language like Igbo languages is precisely where the machine learning and language technology community sees the future's largest obstacle. Many experts in AI perceive in cracking human language to be the next obstacle and also the goal for next generation of AI technologies. Language is the most natural and dominant form of communication, and debatably one of the main apparent signals of higher intelligence. Simultaneously, language is mussy, equivocal and ever-changing so to understand it you need a high level of cultural, common-sense and contextual knowledge. To satisfy our vision of close intelligence where intelligent devices communicate seamlessly with us, we need to substantially ameliorate existing technology and methods that solve machine learning algorithm for translation of English to Igbo language problems.

Machine learning is a device for converting information into knowledge. For the past 50 years, there has been a data outburst. This mass of data is unhelpful unless we analyze it and discover the hidden models within [14]. Machine learning techniques are used to automatically discover the valuable fundamental models within composite data that we would otherwise strife to find. The hidden models and knowledge about a problem can be used to forecast future events and carry out all kinds of composite decision making. Most of us are oblivion that we already communicate with Machine Learning each day. Each time we Google something, pay attention to a song or even snap a shot, Machine Learning is becoming part of the major device behind it, constantly learning and enhancing from every conversation.

Prior to machine learning, computers could only execute conscious-thinking works since those were the only ones we knew how to program. Unconscious-thinking works were beyond the reach of computers since they could not be programmed. Besides, IT systems are only just beginning to finger the definition, aim and sentiment behind our trillions of written and spoken words. Language produces a very big part of data treasure incessantly. [12] Hence, computers cannot decipher texts and questions well enough to give high-quality translations, accurate summaries or suitable answers in all languages.

2. Related Literature

Machine learning algorithms are the engines of machine learning. This means it is the algorithms that turn a data set into a model. Goyal et al [6] proposed Hindi to Punjabi Machine Translation System. This system is on the basis of word-word translation method. It comprises of morphological theory, word sense clarification, transcription and post processing.

Hana, [7] had noted that Amharic to English Language Translator for iOS was developed to utilize a Translation system which was based on Microsoft Interpreter Center point and utilized Microsoft Interpreter Programming interface to make a Translation of writings from Amharic to English.

Odejobi, et al., [10] proposed method that can aid the instruction and studying of Hausa, Igbo, and Yoruba languages. The investigation accepted human body parts ID, plants distinguishing proof, and creatures' names. The English to Yoruba machine translation and Yoruba number tallying systems were a piece of the fundamental theory. The theory was created to develop a system for the learner of the three languages.

Agbeyangi, et al., [2] developed rule-based policy for English to Yoruba Machine Translation System. This platform provided three ways to deal with Machine Translation process. The creators investigated these methodologies and accepted principle based methodologies as the Translation procedure. As indicated by the authors, there was constrained corpus that was accessible for Yoruba language, which illuminated the standard based methodology.

Chinenyeze et al [4] developed Natural Language Processing System for English language to Igbo Language Translation in Android. The aim was to build a translator Android based application that can be useful to some people who wishes to translate English texts into Igbo. The group applied Statistical Machine Translation to employ statistical translation theories gotten from monolingual analysis and bilingual training data. Crucially, this theory used computing power to develop sophisticated data models to translate one source language into another. Statistical Machine Translation comprises of Language Model (LM), Translation Model (TM) and Decoder.

Ifeanyi et al [9] proposed a theory and representation of Igbo text document for a text based analysis accepting its compounding nature and defined its representation with the Word-based N-gram model so as to get it completely ready for any text-based application. Their result showed that Bigram and Trigram n-gram text representation models give more semantic information as well addressed issues of compounding, ordering of word and juxtaposition which are the major language peculiarities in Igbo.

Iheanetu et al [8] developed Hidden Markov-based Part-of-Speech Tagger for Igbo Language. They built an Igbo POS tagger using 19 tokens developed from a corpus comprising the first chapter of the Igbo Bible and a translation of the same using Google API. The resulting translation was contradictory with the official Igbo orthography which is the Onwu orthography and at times, the accepted morphology of Igbo. The Penn Treebank tag set used did not capture all word formed in Igbo and as such, may need to be transformed in order to make the morphological peculiarities of Igbo acceptable.

Alon et al [3] built NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua. The system focused on producing basic NLP resources that have enough quality to be put to any number of users from edifying all manner of NLP tools to strongly aid linguists in better knowledge of indigenous culture and language. For both Mapudungun and Quechua, disjoin work on morphology analysis and on a transfer grammar modularized the problem in a way that permitted rapid development. In

addition, a spelling-checker for Mapudungun, the AVENUE team has created computational lexicons, morphology analyzers and one or more Machine Translation systems for Mapudungun and Quechua into Spanish.

3. English and Igbo Language

There are much significant difference between Igbo in morphology, syntax and semantics. In Igbo languages, we have two main parts of speech - Nouns and verbs, the others are seen as derivatives of these two main types.

3.1. Compound Verb

The Oxford Modern English Grammar, Aarts [13] classified the following compound verbs into types:

Table 1. *Classification of compound verb.*

Compound verb types	Examples
verb + verb	blow-fly, drink-link, freeze-frame, crab boil, stir shit, stir bar, fry cook
noun + noun	backdate, backdoor, backfield, bombshell
noun + verb	backdrop, back-crawl, back-draught, back-fill, butterfly, bloodstain
adjective + noun	backend, blabber-mouth, blue-tooth, cool-ant
adjective + verb	cork-screw, deep-fry
preposition + noun/verb	Bargeboard, bargepole, blackmail, blacksmith

It seems that compound verbs in English form a well-developed lexical class and need the attention of word-formationists, typologists and the linguistic community at large. Acceptably, compound verbs are not essential for forming general language typology, but a discussion of compound verbs can lead to the rising interest towards the typology of derivational morphology (for a definition of the broad term derivational morphology and the typological approach to its study see Štekauer, et al [15].

Oha [11], on the basis of Uwalaka's classification, pinpoint the following eleven Igbo compound verb types:

Table 2. *Igbo compound verb types.*

Compound verbs types	Examples
Causative	Kpo-wa, pi-wa
Multi event	Gburi, siri
Motion	Gba-fu, gba-fe
Change of ownership	Nye-fe, zu-nye
Occurrence	Cha-ru, kpo-chu
Surface contact	Ta-kwu, ma-do
Placement	Gbu-nye, do-nye
Experiencer	Le-ba, hu-ju
Mental exertion	Sọ-pụ, ru-be
Communication	Bụ-nye, kwu-hye
Emission	Gbọ-pụ, nyụ-chi

Short passages from Oha [11] observe the combining relations of the components of the V-V compounds including their problems in terms of the operative relations that hold between them. Compound verbs are not freely formed in Igbo but under some problems. There is an unyielding connection between the conceptual properties of the compound verbs and their compound types; one which according to Gamerschlag [5] derives the argument structure from such conceptual properties.

3.2. Noun Phrase

A noun Phrase (NP) is a phrase in which a noun or pronoun is the governor or head word, facultatively accompanied by a modifier set. NP can be pre-modified or post modified. The NP is said to be pre-modified only when the modifier is placed before the noun while then the NP is said to be post-modified when the modifier is placed after the noun. English accepts both forms of modification. Igbo accepts only post-modification, modifiers are placed after nouns. A noun phrase consists of three parts, the head which is the principal part and other two optionally occurring parts.

4. Proposed System Architecture and Modules

The process used for translating from English to Igbo in this dissertation was carried out using machine learning algorithm. Unsupervised learning was used for the translation. Clustering is an unsupervised technique which was used to translate from English to Igbo. Clustering is an unsupervised learning task. Hence, given a data set of unlabeled data of bilingual languages a clustering algorithm tries to translate and group them to more significant clusters. Then label will be assigned to each cluster, providing a means for generalizing over the data objects and their features, in contrast to supervised learning (i.e., classification), where for the data set a label or target is already given to the patterns (training set). Clustering is very important for large and high dimensional datasets since it gives a simplification of the underlying data distribution, and it also helps to disclose obscure structure and information. Hence, there is no teacher to supervise the learning process. The purpose is to gather the examples into groups based only on their observable characteristics, such that each group contains objects that share some essential characteristics. Once the new system has become tuned to the statistical regularities of the input data, it creates the ability to form internal representations for encoding characteristics of the input data and thereby develop new categories automatically. The following steps were followed to achieve the new system;

1) Data set collection
2) Parser
3) Data Preprocessor, tokenization, postprocessor
4) Feature extraction
5) Clustering

4.1. Data Set Collection

The dataset was gathered from the online oxford dictionary and wordnet. Wordnet is the lexical database for English language.

4.2. Parser

It is an algorithm that yields a syntactic framework for a given input. The first component of the rule based machine translation system is the parser and it is used on the source (English) side. The Parser establish in Natural Language Processing Toolkit was downloaded and used. The parser is used to check the grammatical correctness of the English input.

4.3. Preprocessor

The preprocessor counts the number of words in the English noun phrase input, verb and compound verb then explicitly state three arrays of the size of the number of words for use by the other modules.

4.4. Tokenization

A basic text processing operation is tokenization which is the breaking up of raw text or sentence into words. This function is carried out by the tokenizer. The input sentence is broken up at this point into words. It accepts a word whenever a space is encountered which shows the end of the word. It then puts each of the tokens (words) into one of the arrays developed by the preprocessor.

4.5. Postprocessor

The Postprocessor gives access to the full-form bilingual dictionary for each of the tokens in the array, regains its part of speech (pos tag) and Igbo equivalent. It saves the regained pos tags and Igbo equivalents in the remaining two arrays respectively. Thus arrays of English word (the tokens), pos tags and Igbo equivalents are produced by the postprocessor. Tokens not seen in the full–form bilingual dictionary are enumerated and shown by the postprocessor.

4.6. Feature Extraction

Feature extraction is a dimensionality reduction process where the dataset is reduced to more manageable characteristics for processing while still correctly and fully defining the original dataset. It is intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps in some cases leading to better human interpretations. One of the frequently used method to get the characteristics from textual data is checking the frequency of words/tokens in the document/corpus. The mapping from textual data to real-valued vectors is called feature extraction. One of the easiest methods to numerically represent text is Bag of Words (BOW). In BOW, there is a list of unique words in the text corpus called vocabulary. Then each sentence or document can be represented as a vector, with each word represented as 1 for presence and 0 for absence.

4.7. Clustering

Clustering is the task of dividing the dataset into groups called clusters. The purpose is to separate the data in such a way that points within single cluster are very comparable and points in different clusters are diverse It ascertains grouping among unlabeled data. Cluster was built from Wordnet which is the lexical database for English language. K-means algorithm was used to get lexical similarity and semantic comparability. Words are alike lexically if they have a common feature sequence. Lexical comparability can be taken using string-based algorithms that operate on string sequences and character composition. Words are alike semantically if they have the same meaning, and opposite of each other used in the same way, used in the same context or one is a type of another. Illustrating how K-means algorithm work during clustering.

1) Initialize each mean's values based in the range of the feature values of the part of speech (POS), min and max range of noun and verb and we inculcate it to derive the noun phrase and compound verb from it.
2) Euclidean distance was used as a metric for the similarity of the data set depending on the verb and noun we are working with.
3) Use the distance found in the prior step to allot each data point to closest centroid.
4) Take the average of the points in each cluster group to locate the new centroid.

Repeat 2 to 4 steps for a fixed number of iteration, forming a sentence and classifying the two POS to their nearest cluster and update them until the centroid refuses to change then we stop the process as the algorithm has found the optimal solution.

The data used for the translation should be transformed into the single flat table needed by the machine learning algorithms. Data is commonly acquired by regaining it from a relational database using some form of query language (e.g., SQL). This often includes taking data from a variety of sources or tables in the database and joining them into a single relation.

5. System Implementation

In order to create the database, a relational database design was used. RDBMS is an important device used for arranging huge number of data and describing the connection among the data sets in a more compatible way. A RDBMS creates a framework that's sufficiently pliable to house different set of data. Connections among tables were described by developing unique columns (keys) that include similar set of values in each table. The tables can be combined on various combinations to get the necessary data. A RDBMS also gave pliability which assisted in redesigning and restricting of reports from the database without requiring any means to retype the data. Data dictionaries were used to give clear meaning of data used; these involved the final data frameworks for the different

tables and their relate data fields, description and sizes. The user application programs and interface were created using Php with the assistance of structured query language (SQL) and MySql Database.

6. Math Specification

Mathematical computation involved in this work is the breaking of sentence into tokens. The formula used is as shown below;

$$\text{tokens} = \frac{sentence}{total\ noof\ words}$$

7. Input / Output Format

We have three input forms in the translation of English to Igbo. They include the login form, the vocabulary building form and the translation form.

7.1. The Admin Login Form

Figure 1 contains the login specification for admin user on the platform which includes the username and the password. Once the specification is entered, clicking on the LOGIN button will validate the data before launching the user on the assigned subsystem, while CANCEL will stop the process.

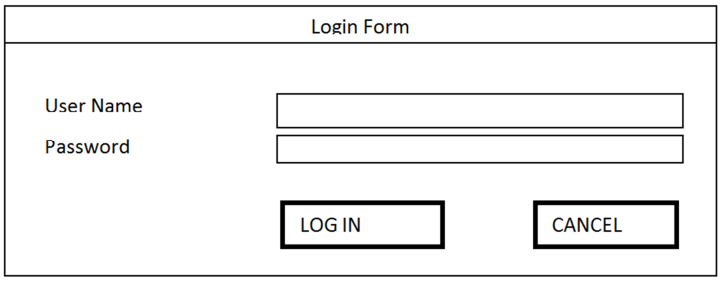

Figure 1. *Admin Login Form.*

7.2. The English – Igbo Build Vocabulary Form

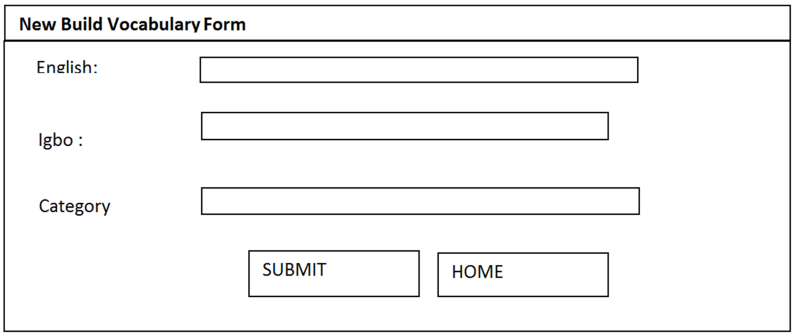

Figure 2. *Build English – Igbo Vocabulary Form.*

Figure 2 is used to create a new word vocabulary on the database. The user enters the English word with the Igbo translation and uploads it to the database. Once the specifications are entered, clicking on the SUBMIT button will validate the data before submitting the record to database.

7.3. The Translation Form

Figure 3 allows the user to enter the English word or phrase to search and the system will display the meaning of the word in Igbo once the word is found in the dictionary.

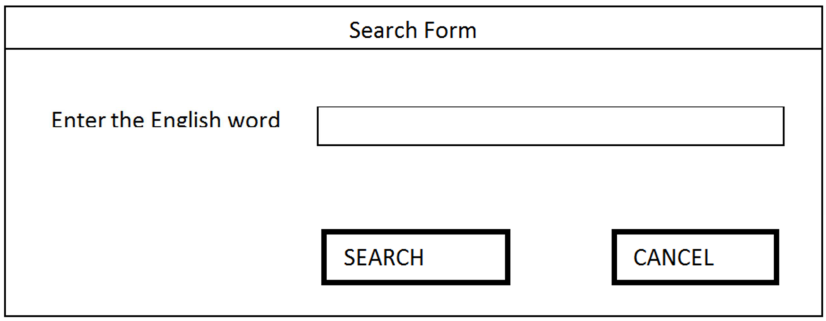

Figure 3. *Translation Form.*

7.4. Algorithm

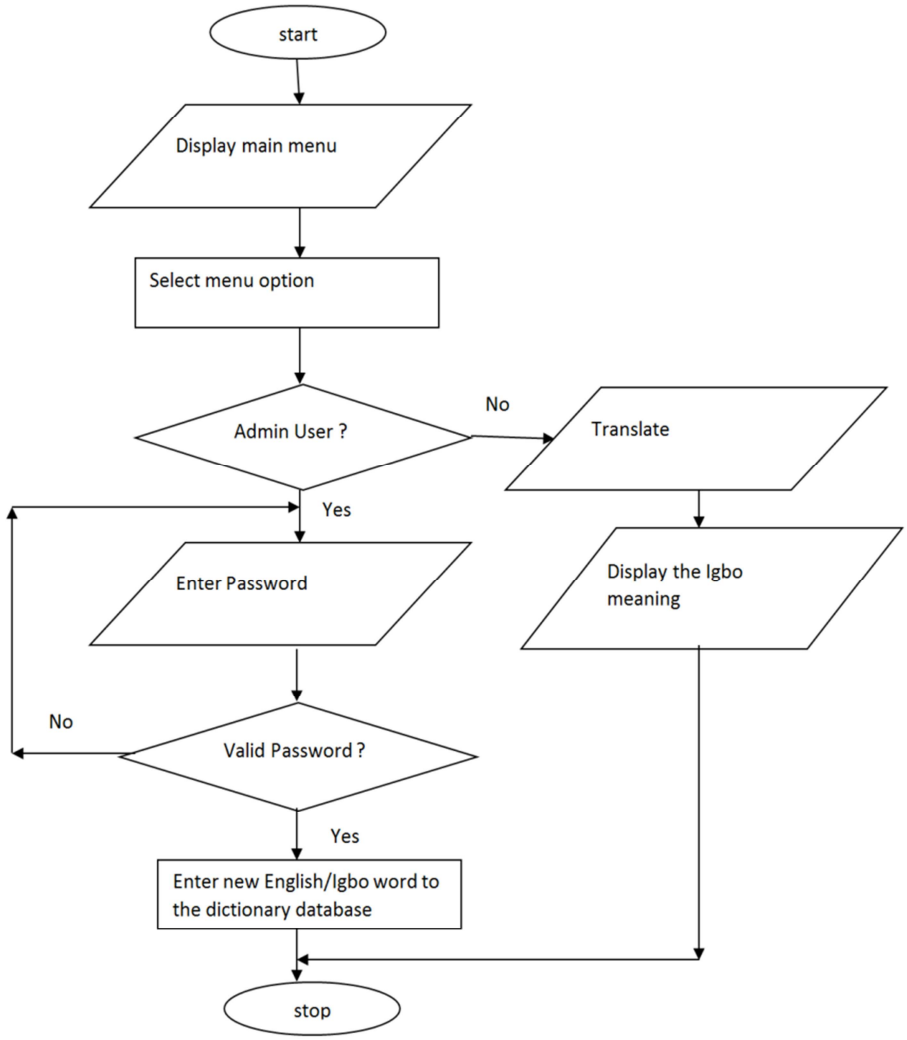

Figure 4. *Program Algorithm.*

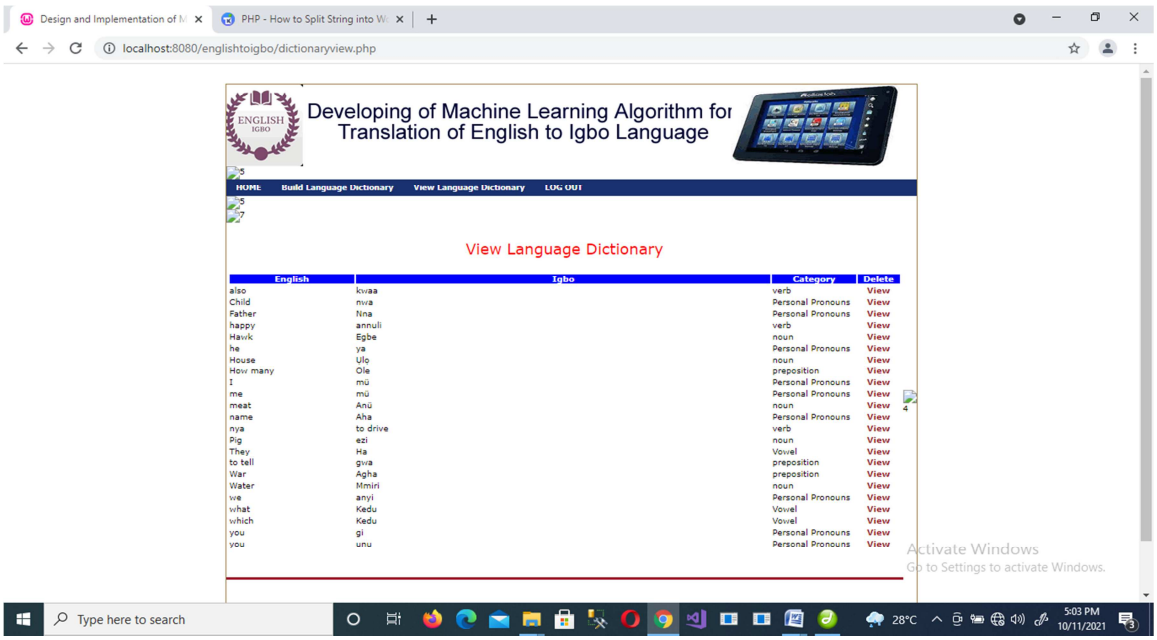

Figure 5. *English / Igbo Translation Database View.*

8. Conclusion

This research implemented machine learning technique for translating English to Igbo language. Machine learning deals with gaining knowledge through data, and is different from traditional applications or programs that generate statistics or engineering output. Machine learning technologies give advantages such as speed, power, efficiency and intelligence through learning without having to explicitly program these features into an application. In other words, Machine learning aids us to determine how we make decisions instead of programming those decision. This gives several opportunities for developers and data science teams to boost product offerings, customer relationships, marketing and advertising, process improvement e.t.c.

9. Recommendation

This research work is recommended to all organizations, companies and individuals that are having difficulty in learning /translating English to Igbo language. This research suggests that future research on English to Igbo translator should be built to integrate voice translation of the language.

References

[1] Abiola O. B, Adetunmbi A. O, Oguntimilehin A (2015) "Review of the Various Approaches to Text to Text Machine Translations" International Journal of Computer Applications. Vol 120 No 18, pp 7-12. ISSN: 0975-8887.

[2] Agbeyangi, A. O., Eludiora, S. I., and Adenekan, D. I. (2015). "English to Yorùbá Machine Translation System using Rule-Based Approach". Journal of Multidisciplinary Engineering Science and Technology (JMEST), Vol. 2 Issue 8, August, Nigeria.

[3] Alon. L, Christian. M. and Roberto. A. (2020) "Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua". Proceedings of the MT2020 workshop at MT Summit.

[4] Chinenyeze C. E, Bennett E. O. and Taylor O. E. (2019) A Natural Language Processing System for English to Igbo Language Translation in Android. International Journal of Computer Science and Mathematical Theory ISSN 2545-5699 Vol. 5 No. 1. www.iiardpub.org

[5] Gammerschlaag, T. (2000) Deriving Argument Structure in Japanese V-V Compounds. Working Paper of the SFB Theorie Des Lexikons No. 282, University of Düsseldorf.

[6] Goyal, V., and Lehal, G. S. (2010). "Hindi to Punjabi Machine Translation System". Department of Computer Science, Punjabi University, Patiala, India.

[7] Hana, B. D. (2016). "Amharic to English Language Translator for iOS". Department of Information Technology, Helsinki Metropolia University, Finland.

[8] Iheanetu. O., Michael. K. and Ojo. S. O (2019) "Hidden Markov-based Part-of-Speech Tagger for Igbo Language". Aclweb.org/anthology/nsurl-1.18.

[9] Ifeanyi. R. N, Ugwu. C. and Adegbola. T. (2017) "Analysis and Representation of Igbo Text Document for a Text-Based System". International Journal of Data Mining Techniques and Applications Volume: 06, Issue: 01, Page No. 26-32 ISSN: 2278-2419.

[10] Odejobi. O. A., Ajayi. A. O, Lukman. A. and Safiriyu. I. E (2015) "A Web Based System for Supporting Teaching and Learning of Nigerian Indigenous Languaage". Faculty of Technology Conference at Obafemi Awolowo University, Ile-Ife, Nigerian.

[11] Oha, A. B. (2010) Verb Compounding in Igbo: A Morpho-Syntactic Analysis. Unpublished PhD Thesis, University of Nigeria, Nsukka.

[12] Olufemi. D. N., Abimbola. R. I., Isacc. O. E and Olamide. E. O. (2017) "Computational Analysis of Igbo Numeral in a Number-To-Text Conversion System". Journal of Computer and Education Research Decemeber. Volume 5. Issue 10.

[13] Oxford Modern English Grammar (2011) Aart edition.

[14] Sangeetha. J, S. and Jothilakshmi, R. N. (2014), "An Efficient Machine Translation System for English to Indian Languages Using Hybrid Mechanism" International Journal of Engineering and Technology (IJET), pp 1909-1919, Vol 6 No 4, ISSN: 0975-4024.

[15] Štekauer, P., Valera, S. and Körtvélyessy, L. (2012). Word-Formation in the World's Languages: A Typological Survey. Cambridge: Cambridge University Press.

[16] UCLA (2014). Language materials project: Igbo. http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=13.