

Igala Language Spell Checker

Sani Felix Ayegba¹, Musa Ugbede², Benson Jessica³, Onoja Abu³

Department of Computer Science,
Federal Polytechnic Idah, Kogi State, Nigeria.
Department of Mathematics/Statistics
Federal Polytechnic Idah, Kogi State, Nigeria.

felixsani@yahoo.com, musa.ugbede@gmail.com, jessben100@gmail.com, abuonoa2008@yahoo.com

Abstract

Spell checkers are important components of Natural language processing applications such as office automation systems, machine translation systems and other similar tools. In this paper, we modeled a language processor for Igala which can detect and correct non word or typographic errors. Lexicon look up approach was adopted for determining validity of words while a variant of similarity key technique based on number of common characters was used to generate the list of correct alternatives for misspelled words. Test and evaluation results of the system indicated that it can handle various types of Igala language spelling errors.

Keywords: spell checker, Natural Language Processing, Igala language, non-word errors, typographic errors, similarity key technique, office automation systems, machine translation systems.

1 Introduction

In Natural Language Processing systems such as machine translation systems, speech recognition systems, retrieval systems, text to speech systems etc. spelling errors are almost unavoidable. Spelling errors refers to human-generated writing errors. Error detection is the procedure for identifying incorrectly spelled words in a text. Error correction is the procedure of correcting an error once it has been detected. It means replacing the incorrectly spelled word with the intended word. Manually detecting and correcting errors is a herculean task. Because of the prevalence of spelling errors and the cumbersomeness of manually detecting and correcting errors, techniques for automatic detection and correction of errors has been evolved. Spell checking refers to automatic spelling error detection and correction and the software for performing automatic spelling error detection and correction is called a spell checker. Automatic spelling error detection and correction systems are classified into two categories namely: fully automatic systems and interactive systems. Fully automatic systems detect and correct errors without any intervention by the user. Interactive systems detect errors, then suggest a number of possible corrections and allow the user to choose the word that should replace the erroneous word. Spell checkers have been incorporated into a number of computer software such as web browsers, text processors, and search engines to achieve high level of accuracy [14].

Igala history dates back to Egypt between 36 and 841 AD. The invasion of Egypt by Arabs from Southern Europe led to their migration from Egypt to their present abode in Kogi State [17]. Igala is the language of the ethnic group located at the eastern flank of the confluence of rivers Niger and Benue. They are the ninth largest linguistics group in Nigeria [10]. Geo-politically, they are described as belonging to the middle belt or north-central of Nigeria. They are bordered on the north by Benue and Nassarawa States, on the West by

River Niger, on the East by Enugu State and on the South by Anambra State [3]. Igala land is 120 Kilometres wide and 160 Kilometres long. It is located approximately between latitudes 6° 30' and 8° North and longitudes 6° 30' and 7° 40' East and covers an area of about 13,665 square kilometers. The population of the Igala people is estimated at two-million in the late 1990s [3]. Historically, they are said to be linked to the Yoruba, the Jukuns and the Binis (Edo) and the northern Ibos. Owing to their central location, they have mutually interacted and lived with the Idomas, Bassa-Nkomo, Nupe, Igbirra and Hausa people. The Igala ethnic group is densely populated in their settlements around the major towns such as Idah, Ankpa and Anyigba. They are also found in Edo, Delta, Anambra, Enugu, Nassarawa, Adamawa and Benue States. However, the bulk of them are indisputably found in Idah, Ankpa, Dekina, Omala, Olamaboro, Ofu, Igalamela/Odolu, Ibaji, Bassa (and even Lokoja and Ajaokuta) Local Government Areas of Kogi State [3].

1.1 Features and Morphological Processes in Igala language

Igala is a register tone language with 3 basic tonemes: high, medium and low (H,M,L). The language has 30 phonemic sounds with 23 consonants and 7 vowels. There are two main parts of speech in the language: nouns and verbs, others are derivatives of these two main types. Tone plays a very significant part in realizing parts of speech. It is isolating with agglutinative features, mark aspects but no grammatical tense. The language is fixed word order (SVO).

The language has a number of diacritics such as (‘ ’ ~ . etc.) that must be added to its orthographic or phonetic characters for correct meaning and pronunciation to be realized, for example *leta* and *eta* means “three” and “to sell” respectively.

The process of negating verbs is another feature that distinguishes the language from others. To negate a verb in Igala, a space is provided after the verb followed by *n*, e.g. *gękiwa n*, means do not come, it is the letter *n* placed after the verb “*wa*” that negated the sentence.

Igala language exhibits a number of morphological processes such as affixation (prefixation and suffixation), compounding and derivation. For example the root word *kęche* can be prefixed with *e* to form *ekęche* which means learning. It can also be prefixed with *a* to form *akęche* to mean learner.

Morphological analysis is the identification of a word-stem from a full word-form and sometimes also the identification of the syntactic category of the stem. It is the decomposition of words into their uninflected forms, performed at word levels. It is a very important subtask in Natural Language Processing. According to [11] there are two techniques for dealing with morphological analysis: i. full – form lexicon: in this technique all the inflected variants of the word are listed in the lexicon. ii. Morphological analysis component: this is a rule based module that analyzes a word and relate it to its root or base form.

This study adopts the full form lexicon technique which lists all words and their derivatives in the dictionary. The reason for this choice is that Igala is an isolating language which means that it is also poor in morphology [16].

1.2 Classification of Spelling Errors

Spelling errors are common occurrences in Natural Language Processing systems especially in text processing. A number of studies have been carried out to analyze the types and trends of spelling errors. [9, 5] identified two types of spelling errors namely: Typographic and Cognitive errors. Typographic error occurs when a word is mistyped. Cognitive error occurs when the correct spelling of the word is not known due to the similarity in the pronunciation of the mistyped word and the intended correct word. [8, 13, 15, 2] further classified spelling errors into two categories viz: non word errors and real word errors. Non word error occurs when a word have no meaning and therefore not included in the lexicon. In real word error, words are lexically correct but syntactically and semantically incorrect.

In order to ascertain the frequency of occurrence of the different types of errors, news articles and minutes of local town hall meetings were collected from Nigeria Television Authority (NTA) Idah and Igalamela/Odolu Youth Development Association respectively for study. Table 1 shows the result of the study.

Table 1: Error types and frequencies

Document Source	Number of words	Non word errors	Real word errors
News Article (NTA) Idah	2,182	1,251	286
Minute of Meeting	1,865	976	157

Analysis of the result showed that 55.02% of the errors were non word or typographic errors while 10.9% were real word errors. This shows that non word error is more prevalent. Although it was also found out that the level of education and years of experience of the typist is a factor that determines the frequency of occurrence of each type of error, this result is a pointer to the fact that non word errors occurs frequently when performing word processing operations in Igala language, hence the need for a spell checker. We therefore concentrate on the development of a spell checker for correction of non word errors.

1.3 Rationale for the Study

Large volumes of documents are produced in Igala language daily, examples include: news articles, Igala language lesson notes, sermons, customary court proceedings, minutes of local town hall meetings etc. Detecting and correcting errors in these documents is done manually which is both cumbersome and time consuming leading to delay in the release of relevant information to those who need it. Currently there is no spell checker for Igala language. This study is undertaken to provide a spell checker for Igala language. Furthermore the study will also provide a platform for the development of other Natural Language Processing applications.

1.4 Limitation of Study

As mentioned in section II, there are two types of spelling errors, non-word and real word errors. Detecting and correcting real word errors requires both syntactic and semantic analysis. Although semantic analysis can be performed by measuring the semantic distance between words, there is currently no parser for Igala language to carry out the syntax analysis

aspect. Therefore detecting and correcting real word error is not within the scope of our present study.

1.5 Task Description

Let $\mathbf{W} (w_1, w_2 \dots w_n)$ be the set of words in an Igala text $\mathbf{T}$. For each word w_i in $\mathbf{W}$, verify that w_i is misspelled by looking it up in the Igala lexicon. If w_i is not found in the lexicon then it is misspelled. If w_i is misspelled then generate from the lexicon a list of words $\mathbf{Cw}$ ($cw_1, cw_2 \dots cw_n$) which are closeto w_i . Similarity of w_i to cw_i in $\mathbf{Cw}$ is computed using a similarity algorithm. cw_i in $\mathbf{Cw}$ whose similarity to w_i is greater than or equal to 65% is considered a candidate word and is placed in similar word list $\mathbf{S}(s_1, s_2 \dots s_n)$. Any s_i in $\mathbf{S}$ considered closest to w_i is used to replace w_i .

2 Summary of Literature

[13] identified spelling error as major hindrance to real world application of Natural Language Processing. Researchers discovered very early that the full potentials of Natural Language Processing systems will not be fully realized if they are not integrated with applications for automatic detection and correction of spelling errors. Therefore various technologies for detecting and correcting errors were evolved. Application program for detecting and correcting spelling errors is called a spell checker. Error detection involves looking up each word in the dictionary. Two common techniques for detecting errors include dictionary look up and N-gram analysis [15]. In addition to detecting spelling error the spell checker provides suggestions for the incorrectly spelled word in the text. Generating a list of correct alternatives is based on calculating the similarity between the misspelled word and the words contained in the lexicon. Several algorithms for computing similarity between words have been proposed. These include minimum edit distance technique (Levenshtein algorithm, Hamming algorithm, Longest Common Subsequence algorithm), rule based technique, Probabilistic Techniques, Similarity key technique, N-gram Based Techniques, Neural Net Techniques. Details of these techniques are contained in the study carried out by [9, 15]. Others are noisy channel model [4, 6], finite state automaton [1].

3 Materials and Method

The methodology and approach adopted in this paper are described below. In this section, the algorithm used for the generation of correct alternatives was highlighted, architecture of the proposed spell checker and its implementation were presented. Furthermore the techniques and materials used for test and evaluation were discussed.

3.1 Algorithm

A variant of the similarity key algorithm for computing the similarity between two strings was proposed to determine the similarity between the erroneous word and the words in the speller lexicon. Similarity is based on the number of common characters in the words. The number of common characters in each word in the candidate set and the erroneous word is computed. The generation of candidate set is based on the observation made by [19] that in

most cases, the first letter of the misspelled word is always correct. Therefore words in the lexicon whose first letter is equal to the first letter of the misspelled word are retrieved and stored in an array. This forms the candidate set.

For clarity we define the following variables and functions which will be used in the algorithm for the computation of similarity score.

m: misspelled word

f: first letter of misspelled word

Length: function to compute length of array or string

splitChar: function to split a word into its constituent characters and store each character in array

equivalentList: array of words retrieved from the lexicon whose first letter = f

n = equivalentList.Length

mArray = m.splitChar

k = mArray.Length

numCommonChar: number of common characters in two words whose similarity is been computed.

for i = 0 to n - 1

numCommonChar = 0

b = equivalentList[i]

bArray = b.splitChar

p = bArray.Length

for x = 0 to k - 1

for y = 0 to p - 1

if mArray(x) = bArray(y) then

numCommonChar = numCommonChar + 1

end if

Next

Next

similarityScore = numCommonChar * 200 / (k + p)

if similarityScore >= 65 then

b is similar to m

place b in the similarity list

end if

Next

3.2 System Architecture and Modules

The architecture of the proposed system is presented in Figure 1.

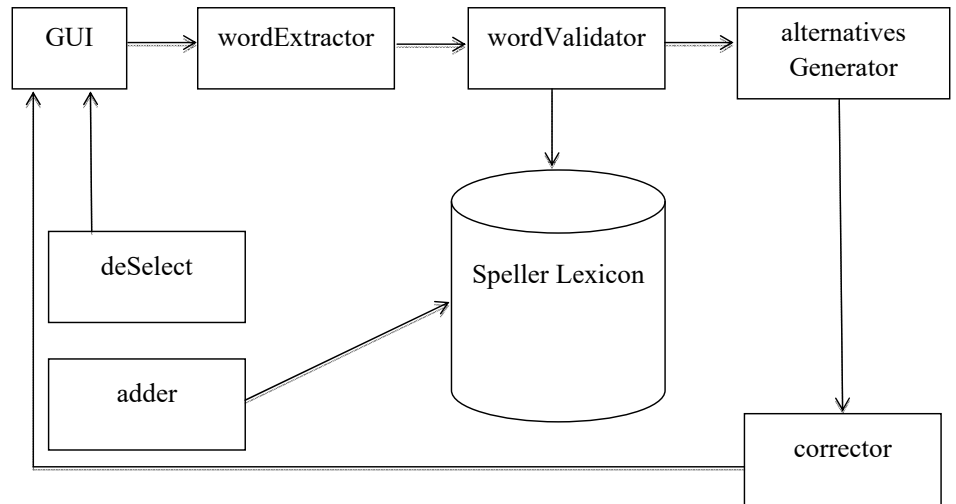


Figure 1: Proposed system architecture

It consists of seven main modules: **GUI**, **wordExtractor**, **wordValidator**, **alternativesGenerator**, **corrector**, **deSelect**, **adder**. **GUI** (for graphic user interface) creates the interface which is displayed for the user to enter the text to be spell checked. The **wordExtractor** determines and extracts the last word typed after pressing the spacebar button on the keyboard. **wordValidator** performs a look up in the speller lexicon to determine if the extracted word is valid or not. If the word is not valid it underlines it. **alternativesGenerator** generates a list of correct alternatives for the erroneous word. **corrector** replaces the underlined erroneous word with the selected word from the alternatives list. **deSelect** removes underline from the word if the user wants to continue with the word without changing it. **adder** adds the underlined word to the speller lexicon and removes the underline.

3.3 System Implementation

The table for the speller lexicon was created in MYSQL. The spell checking engine which automatically detects erroneous words and generate a set of correct alternatives for the erroneous word was developed using vb.net 2010. The interface is shown in figure 2.

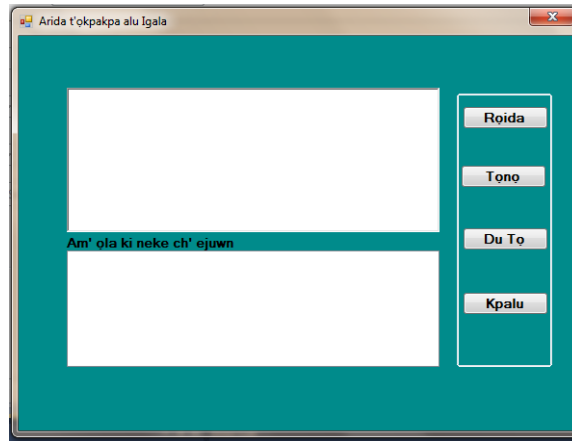


Figure 2: igSpellChecker interface

If a user types a word and presses the spacebar key on the keyboard, the spell checker confirms the validity of the word, if the word is invalid, it is underlined. This is illustrated in figure 3.

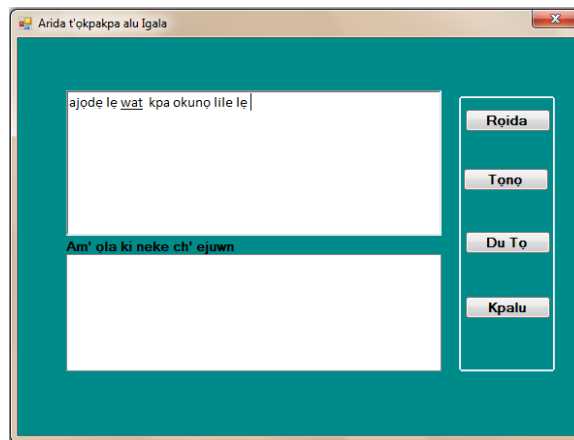


Figure 3: Invalid word detected and underlined

Whenever the cursor is placed on the underlined word and right clicked, a list of correct alternative for the invalid word is generated as shown in figure 4

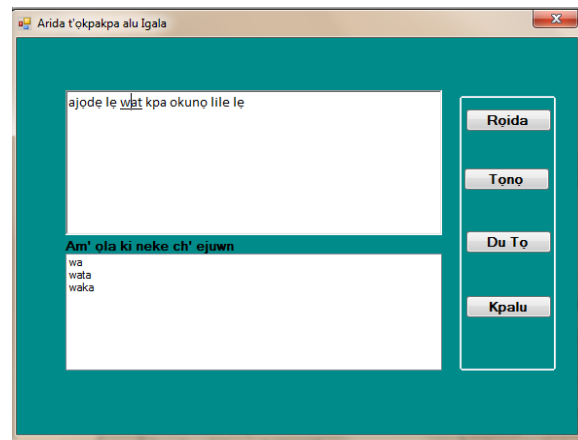


Figure 4: List of alternative words generated for underlined word

The user then selects any of the suggested words and clicks the Roida button, the misspelled word is immediately replaced with the selected word. Figure 5 and 6 illustrates this graphically.

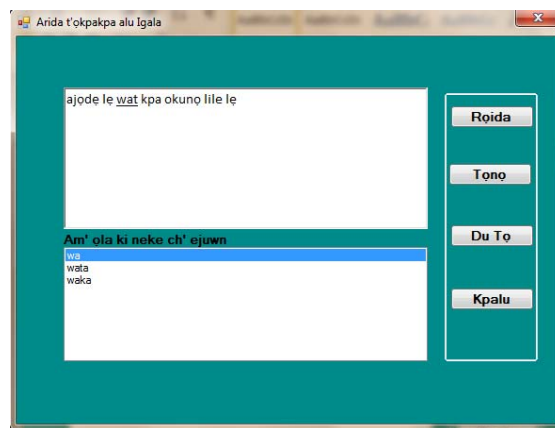


Figure 5: User selects a word from suggestion list

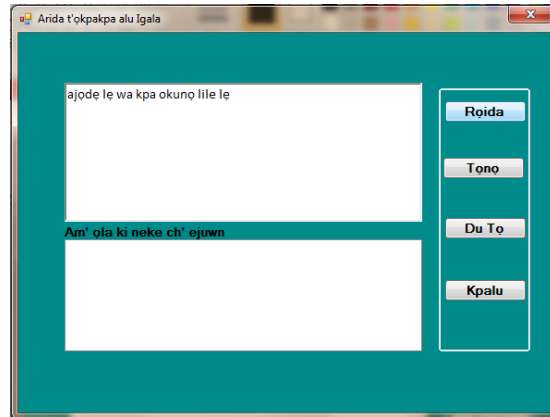


Figure 6: Selected word from suggestion list replaced underlined word

3.4 Test and Evaluation

According to [18, 12, 7] the quality or effectiveness of a spell checker is judged along three key dimensions namely recall (lexical and error), precision and suggestion adequacy. [5] also identified lexicon size, test set size, correction accuracy and error type as evaluation metrics for spell checkers. In this research lexical recall, error recall, precision and suggestion adequacy were used as evaluation metrics for **igSpellChecker**.

Lexical recall refers to the number of valid words in the text that are recognized by the spell checker. Error recall is the number of invalid words in the text that are flagged as invalid by the spell checker. Precision is the ratio of the number of correctly flagged invalid words to the number of words flagged. Suggestion adequacy is a measure of the extent to which the spell checker provides correct alternatives for the erroneous word to the user.

Soft copies of unedited news articles written in Igala language were collected from Nigeria Television Authority Idah branch. The soft copies were merged into one to obtain reasonable text size to avoid repeating the test on small text sizes. The metrics discussed above were then measured after pasting the text in the spell checker control.

4. Results and Discussions

The combined softcopy of the unedited news article was printed out and manually spell checked by two Igala native speakers. The data set was found to contain 1,876 valid words and 422 misspelled words. In this sample, 1,782 out of the 1,876 words were correctly recognized as valid words by the spell checker while all the 422 misspelled words were flagged as misspelled. The result obtained from measuring each metric and computing the percentage is presented in Table 2.

Table 2: Evaluation Results

Metrics	Value		
	No of valid words	No of words recognized as valid	%
Lexical Recall	1,876	1,782	$1782/1876 \times 100 = 94.9$
Error Recall	No of misspelled words	No of misspelled words flagged	%
	422	422	$422/422 \times 100 = 100$
Precision	No of words correctly flagged	No of words flagged	%
	422	582	$422/582 \times 100 = 72.5$
Suggestion Adequacy	81.2%		

From the result presented in the table 2, it is obvious that **igSpellChecker** performed satisfactorily.

6 Conclusion and Future plan

This study has concentrated on the identification, design and implementation of an important linguistic resource, the spelling checker whose absence has greatly inhibited advancement in the development and deployment of Natural Language Processing applications for Igala language. The design was successfully implemented. Test and evaluation results for the proposed system showed it performed satisfactorily. Text editing which is usually a herculean task can easily be carried out with igSpellChecker. We therefore recommend its integration in all Natural Language Processing applications for Igala language. In future work we plan to develop a context sensitive or context aware spelling checker that can handle real word errors for the language.

References

- [1] Aye Myat Mon, ThandarThein: Myanmar Spell Checker, International Journal of Science and Research, volume 2 issue I, 2013, PP 333-340
- [2] BiswajitSarma, DebashreeGoswami, GeetoshreeGoswami, Assamese Spell Checker Design and Implementation: International Journal of Modern Trends in Engineering and Research 2016, PP 44-47
- [3] Egbunu, F. E. 2013. Education and Re-orientation of Igala Cultural Values, African Journal of Culture, Religious, Educational and Environmental Sustainability (AJCREES), Vol. 1, No. 2. Pp. 66 – 82. Dec., 2013.
- [4] Eric Brill and Robert C. Moore, An improved error model for noisy channel spelling correction, in Proceedings of 38th Annual Meeting of the Association for Computational Linguistics., Hong Kong, 2000, pp. 286-293.

- [5] GaddisaOlani, DidaMidekso, Design and Implementation of morphology based spell checker, International Journal of Scientific and Technology research, volume 3 issue 12, December 2014, PP 118-125
- [6] Kristina Toutanova and Robert C. Moore, Pronunciation modeling for improved spelling correction in Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 2002, pp. 144-151
- [7] Marianne Starlander, Andrei Popescu-Belis, A. 2002. Corpus-based evaluation of a French spelling and grammar checker. In: Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain. pp. 268-274.
- [8] Meenu Bhagat Types of Non-word errors in Punjabi Typed Text and Its Comparison with Bangla Text: International Journal of Emerging Technology and Advanced Engineering. (ISSN 2250-2459, ISO 9001:2008, certified journal). Volume 4, issue 7, 2014
- [9] Neha Gupta, Pratish Mathur, Spell Checking Techniques in NLP: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering, volume 2 issue 12, 2012 PP 217-221.
- [10] Omachonu G.S. 2012. Igala Language Studies and Development: Progress, Issues and Challenges, Text of a paper presented at the 12th Igala Education Summit held at Kogi State University, Anyigba- Kogi State, Nigeria. (Dec. 2012).
- [11] Omar Shirko, Nazlia Omar, Haslina Arshad and Mohammed Albared: *Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach*, Journal of Computer Science 6 (3): pp 350-356, 2010, ISSN 1549-3636
- [12] Patrizia Paggio, Nancy L. Underwood, 1997. Validating the TEMAA LE evaluation methodology: a case study on Danish spelling. Natural Language Engineering. 1(1): 1-18.
- [13] Pratip Samanta, Bidyut B. Chaudri, A Simple Real-word Error Detection and Correction using local word bigram and trigram. Proceedings of the twenty-fifth conference on computational linguistics and speech processing (ROCLING 2013) PP. 211-220
- [14] Renato Cordeiro de Amorim, Narcos Zampieri Effective Spell Checking methods using Clustering Algorithms. Proceeding of Recent Advances in Natural language processing, Hissan, Bulgaria, PP 172-178, 2013
- [15] Ritika Mishra, Navjot Kaur: A survey of spelling error detection and correction techniques: International Journal of Computer Trends and Technology, volume 4, issue 3, 2013. PP 372-374
- [16] Sani F. A. 2016. English to Igala Machine Translation System. PhD Dissertation, Universidad Azteca, Mexico.

- [17] Sani Rita I. 2013. Adaptation of the Staff of Office of Attah Igala into Textile Design Forms and Products. Master's Thesis, University of Nigeria, Nnsukka.
- [18] TEMAA. 1997. Evaluation framework for Spelling Checkers: Final Report.
- [19] Yannakoudakis E. J, Fawthrop D. An intelligent spelling error corrector. Information Processing and Management. 19:1, pp 101-108, 1983.