INTERNATIONAL JOURNAL OF

COMPUTATIONAL LINGUISTICS (IJCL)

INTERNATIONAL JOURNAL OF
COMPUTATIONAL LINGUISTICS (IJCL)

VOLUME 2, ISSUE 2, 2011

EDITED BY
DR. NABEEL TAHIR

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

CSC Publishers, 2011

EDITORIAL PREFACE

The International Journal of Computational Linguistics (IJCL) is an effective medium for interchange of high quality theoretical and applied research in Computational Linguistics from theoretical research to application development. This is the third issue of second volume of IJCL. The Journal is published bi-monthly, with papers being peer reviewed to high international standards. International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches.

IJCL give an opportunity to scientists, researchers, and vendors from different disciplines of Artificial Intelligence to share the ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development. This journal is helpful for the researchers and R&D engineers, scientists all those persons who are involve in Computational Linguistics.

Highly professional scholars give their efforts, valuable time, expertise and motivation to IJCL as Editorial board members. All submissions are evaluated by the International Editorial Board. The International Editorial Board ensures that significant developments in image processing from around the world are reflected in the IJCL publications.

IJCL editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Scribd, CiteSeerX Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCL provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computational Linguistics (IJCL)

EDITORIAL BOARD

TABLE OF CONTENTS

Volume 2, Issue 2, October 2011

Pages

A Computational Model Of Yoruba Morphology Lexical Analyzer

Aladesote Isaiah   *iomjovic@yahoo.com*
Department of Computer Science
Rufus Giwa Polytechnic,
P.M.B 1019
Owo, Ondo State, Nigeria.

Olaseni Olatunde E   *talktopastor@yahoo.com*
Department of Computer Science
Rufus Giwa Polytechnic,
P.M.B 1019,
Owo, Ondo State, Nigeria.

Adetunmbi A O
Department of Computer Science
Federal University of Technology,
Akure, Ondo State.

Akinbohun Folake
Department of Computer Science
Rufus Giwa Polytechnic,
P.M.B 1019
Owo, Ondo State,Nigeria.

Abstract

Morphological analyzers are essential parts of many natural-language processing system such as machine translation systems; they may be efficiently implemented as finite state transducers. This paper models a Yoruba lexical analyzer using a rule based approach to computational morphology. This analysis relies solely on one source of information: a dictionary of the valid Yoruba language words.

Keywords: Morphology, Yoruba, Transducer

1. INTRODUCTION

Natural Language Processing (**NLP**) is a field of Computer Science and linguistics concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language. Natural language understanding systems convert samples of human language into more formal representations such as parse trees or first order logic that are easier for computer programs to manipulate. Many problems within NLP apply to both generation and understanding; for example, a computer must be able to model morphology (the structure of words) in order to understand an English sentence, and a model of morphology is also needed for producing a grammatically correct English sentence.

NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish human languages (such as Spanish, Swahili or Swedish) from formal or computer languages (such as C++, Java or LISP). Although NLP may encompass both text and speech, work on speech processing has evolved into a separate field.

A language is a system for encoding information. In its most common use, the term refers to so-called "natural languages" the forms of communication considered peculiar to humankind. In linguistics the term is extended to refer to the human cognitive facility of creating and using language. Essential to

Aladesote Isaiah, Olaseni O. E, Adetunmbi A. O & Akinbohun Folake

both meanings is the systematic creation and usage of systems of symbols, each referring to linguistic concepts with semantic or logical or otherwise expressive meanings.

Morphology, as a branch of linguistics which is concerned with the study of how words are formed, has had a chequered history. Although everybody knows the importance of words in human language, a separate branch of linguistics which is devoted to the study of the internal stucture of words did not emerge until the early part of the nineteenth century (cf. Katamba 1993:3)

The Yoruba language belongs to the West Benue-Congo of the Niger-Congo phylum of African languages (Williamson and Blench 2000: 31). Apart from Nigeria with about 30 million Yoruba speakers. Yoruba is still spoken in Togo, Republic of Benin, Ghana, Sudan, Sierra-Leone and Cote D'Ivoire. Outside Africa, a great number of speakers of the language are in Brazil, Cuba, including Trinidad and Tobago. Yoruba is regarded as one of the major languages of Nigeria. The effective speakers of the language in the country are about 35% of the country's total population. According to the International African Institute (1980: 60), the Yoruba language "is used by the media i.e. the Press, Radio and Television. It is also used as a language of formal instruction and a curriculum subject in the primary school. In the secondary school and post-secondary level (including University); it is a curriculum subject. It has a standard orthography" the Yoruba language occupies a privileged place within the entire range of African studies. A relatively literate exists on the language both in the European languages and in the Yoruba language itself.

MOTIVATION OF THE PAPER

Languages with large number of speakers like Yoruba can nonetheless be in danger. Brenzinger (1998: 93) had earlier noted this when he said "even Yoruba, with 20 over million speakers, has been called 'deprived' because of the way it has come to be dominated by English in higher education". The section 53 of the 1999 constitution of the Federal Republic of Nigeria recognizes English as the official language. Moreover, the suppressive effects of English over the Yoruba language and other Nigerian languages are too overwhelming and suicidal.

The Global Information Capitalism has already sanctioned the versality and dynamism of English as the only thriving language. And, since English has captured the Nigerian nation, implementing any educational policies on mother tongue like that of UNESCO, will continue to be an exercise in futility. This will lead to endangerment, then to moribund and finally to total extinction. Until this happens to Yoruba and other indigenous African languages, the suppressive tasks before the English language and her few allies, would be not completed. The task is still going on because as of today, a good percentage of the products of the Nigerian educational system are, according to Bamgbose (1973: 7), "neither competent in the use of English nor in that of their mother tongue".

Global Information Technology should not necessarily be an avenue towards a total annihilation of the Yoruba language. Instead, the IT and the Internet should give Yoruba language a public profile. It is therefore no doubt that Yoruba language can be available to Information Technology (IT).

OBJECTIVE OF THE STUDY

The aim of this paper is to develop morphology Lexical Analyzer for Yoruba language using computational model finite automata, which is one of the essential parts of natural language processing systems.

METHODOLOGY

There are two approaches to Computational morphology: rule-based and data-based approaches. The former involves the use of grammatical rules to construct computational morphology while the latter uses statistical information to develop computational morphology. Rule-based approach is adopted in constucting a morphological lexical analyzer. One of the most efficient approaches to morphological lexical analysis and generation uses Finte - State Tranducers (FST) (Mohri 1997a; Oncina et al 1993). FST is composed by a finite set of states and a set of transitions between pairs of states.

DISCUSSION

A.   Analysis of the Proposed System

The system functions by taking input (sentence, in form of text) in Yoruba language from the user, removes punctuation marks like comma, semi colon, colon and blank space(s), brakes the input into individual words with corresponding part of speech (in form of text).

Yoruba is an SVO (Subject Object Verb) language. The following tree diagrams explain the techniques for implementing the proposed system. These diagrams illustrate the structure of the tokens of the source language (Yoruba sentence). Classification of languages on the basis of the basic order of the verb, the subject and the object in a sentence into several types: SVO, VSO.
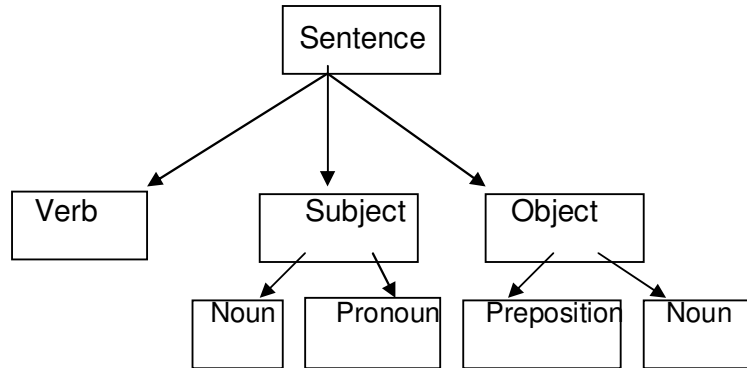
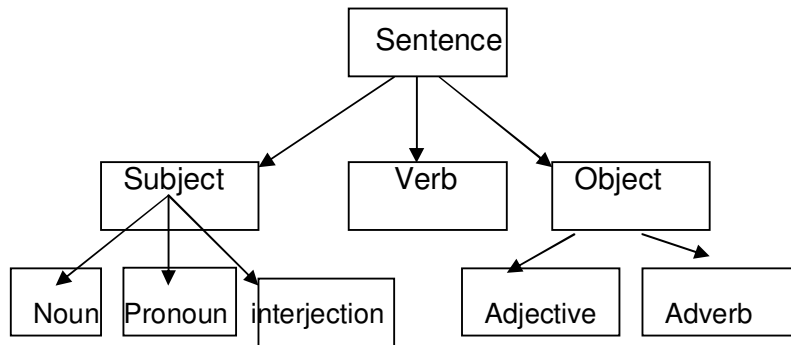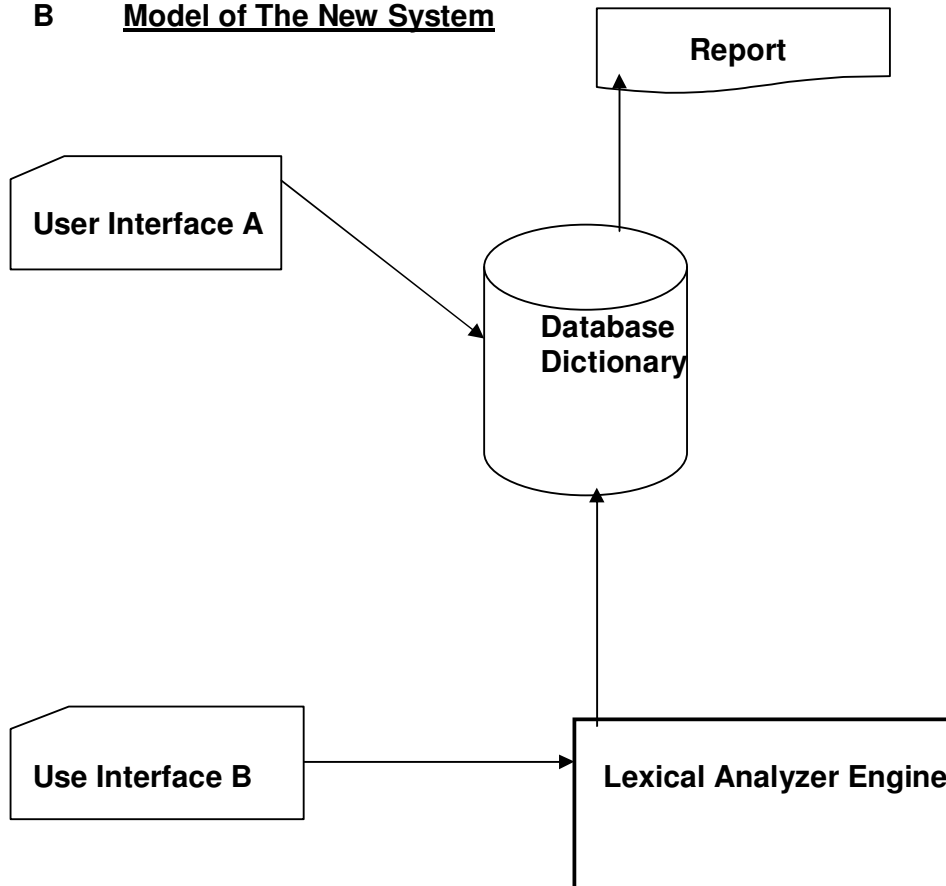**FIGURE 1:** Examples are Riri nimo ri, Gbigba ni mo gba ile.

FIGURE 2: Examples are Mo jeun lana, Oh! O na Ade

B   Model of The New System

Report

User Interface A

Database Dictionary

Use Interface B

Lexical Analyzer Engine

User Interface A: This is a fundamental part that serves as an entrance that users can visit. Here, the word in Yoruba Language with their corresponding part of speech is entered via the Yoruba keyboard on the module named "entrypart". The entrypart links to partsubmit which is the engine that makes entrypart works.

Database dictionary: This is named "eyaede" which is a necessity for storing Yoruba words. There is a link between the user interface and the dictionary, at the click of store button, the word (oro) will be stored in the dictionary.

User Interface B: This is the interface where the sentence (gbolohun) to be analyzed is entered on the module named "analyzer".

Lexical Analyzer Engine: This takes input (gbolohun) through the interface B, reads through the input characters, it stops whenever it encounters a space, which signifies the end of a word (oro), eliminates the space and also the Analyzer checks through the dictionary (Eyaede) to determine the part of speech of such word (oro). If such word is not found in the dictionary, it can be re-entered in the user interface A, otherwise, the word will be eliminated automatically. It repeats this process for all the words (oro) in that sentence (gbolohun) until it encounters full stop (ami idaduro) which indicates the end of the sentence (gbolohun).

C   Specifications of Hardware and Software

The proposed system can functions its operation as expected if hardware and software are included.

Hardware Specifications
(i)   Pentium IV or higher motherboard
(ii)   Minimum of 100MHz clock speed processor
(iii)   512 RAM
(iv)   2 or 3 GB of Hard disk.
(v)   Colour Monitor
(vi)   Mouse
(vii)   Keyboard
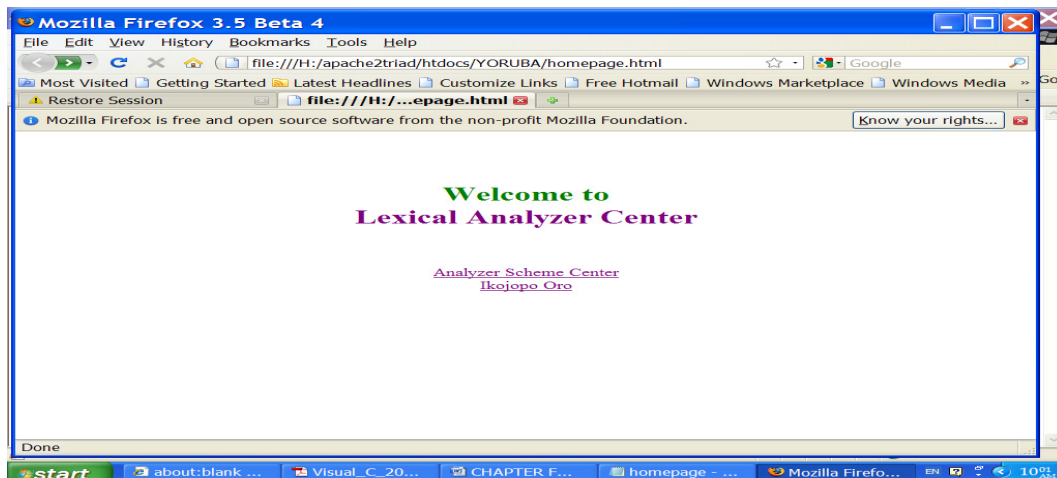(viii) Printers

Software Specifications: Windows Operating System, MYSQL, PHP, Macromedia Flash, Macromedia fireworks and Microsoft Access.

D   The Systems Interface
To launch the application for the Yoruba alphabets, follow the steps below
go to the start menu, click on run
a.   type in the box http://localhost/yoruba/homepage.php

FIGURE : D1

b. click on analyzer scheme center
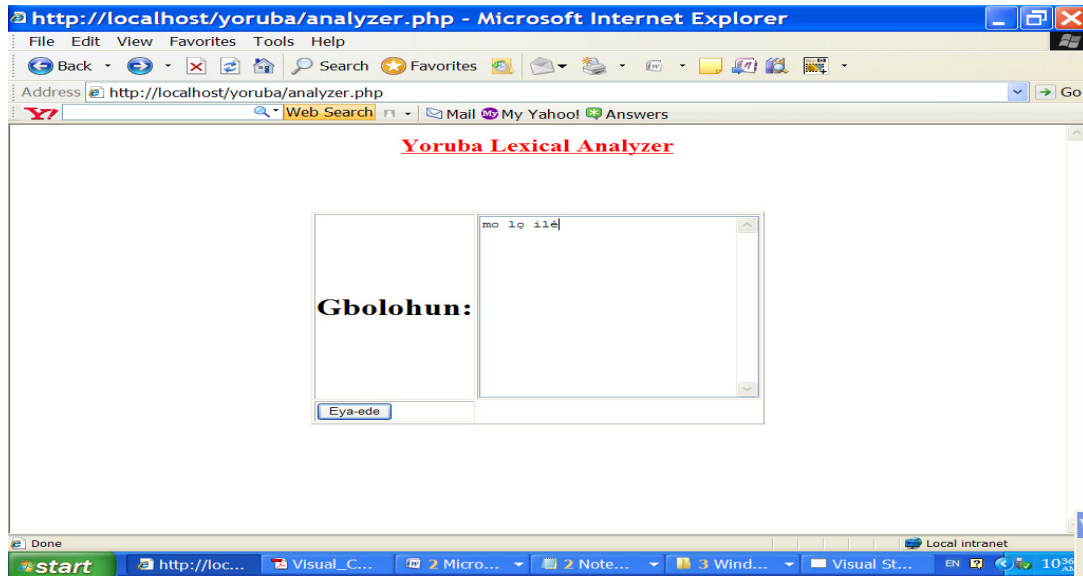
FIGURE : D2

To launch the application for the translation of Yoruba language dictionary, follow the steps below:

a. Type in your word, sentence or Yoruba word (gbolohun) in the text box
b. Click the 'Eya-ede' button
c. The translation is then displayed.
d. The following window appears.

FIGURE : D3

To launch the Yoruba Dictionary Corner.

a. From Homepage,: press Ikojopo Oro
b. Type the word into the text box.
c. Select the part of speech for the word
d. Click Store.

Aladesote Isaiah, Olaseni O. E, Adetunmbi A. O & Akinbohun Folake

FIGURE : D4 The following windows appear if the word is exiting or already saved
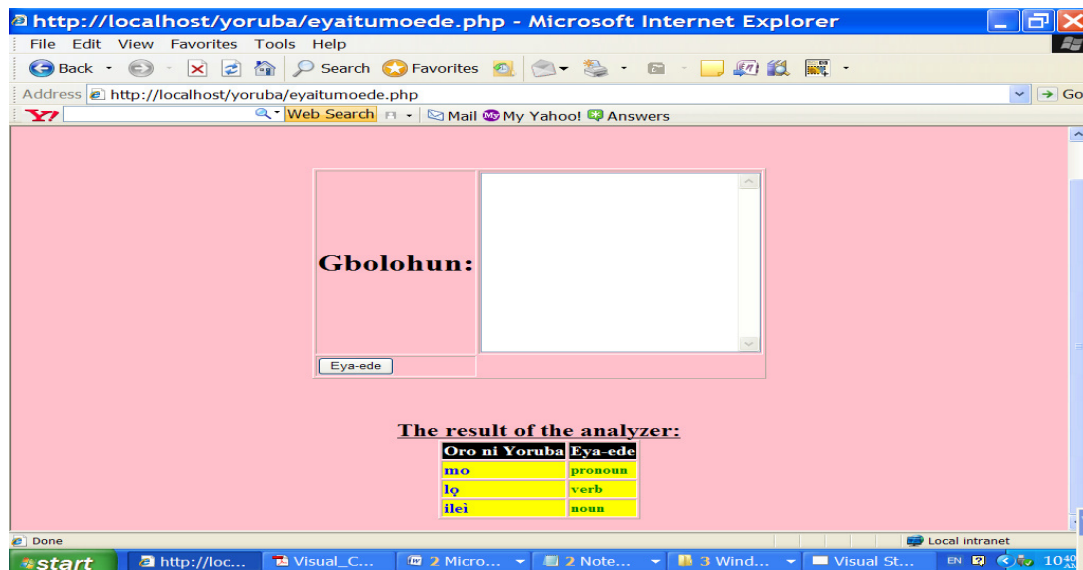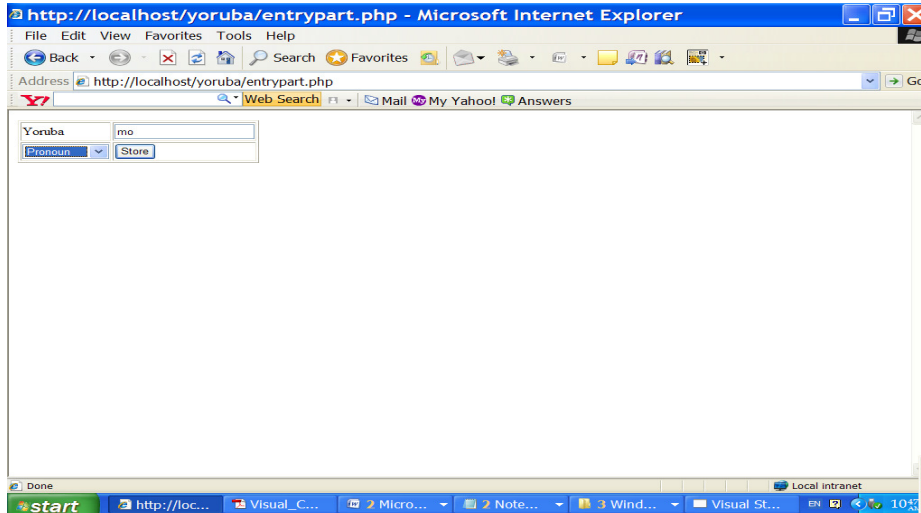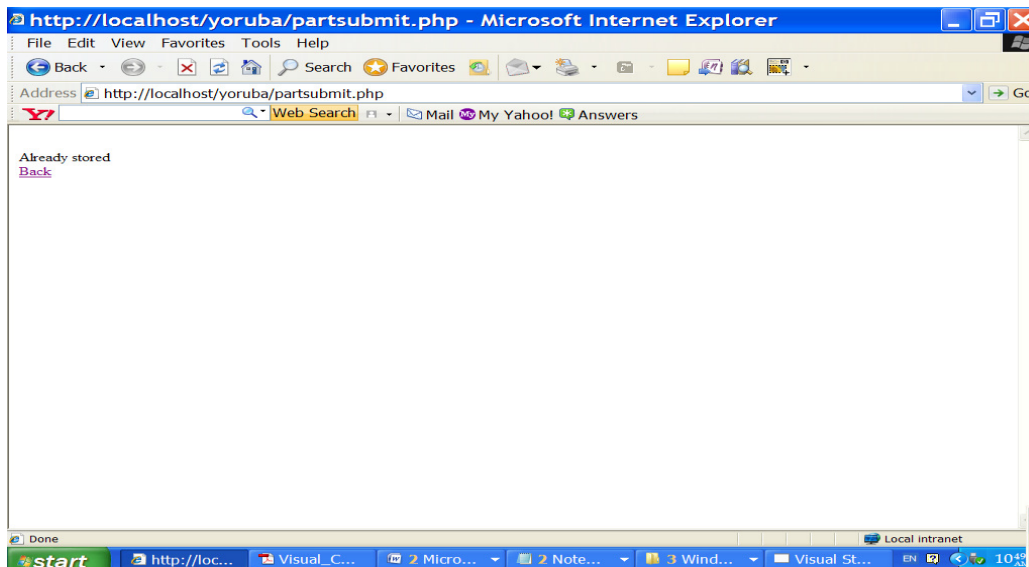

FIGURE 1: D5

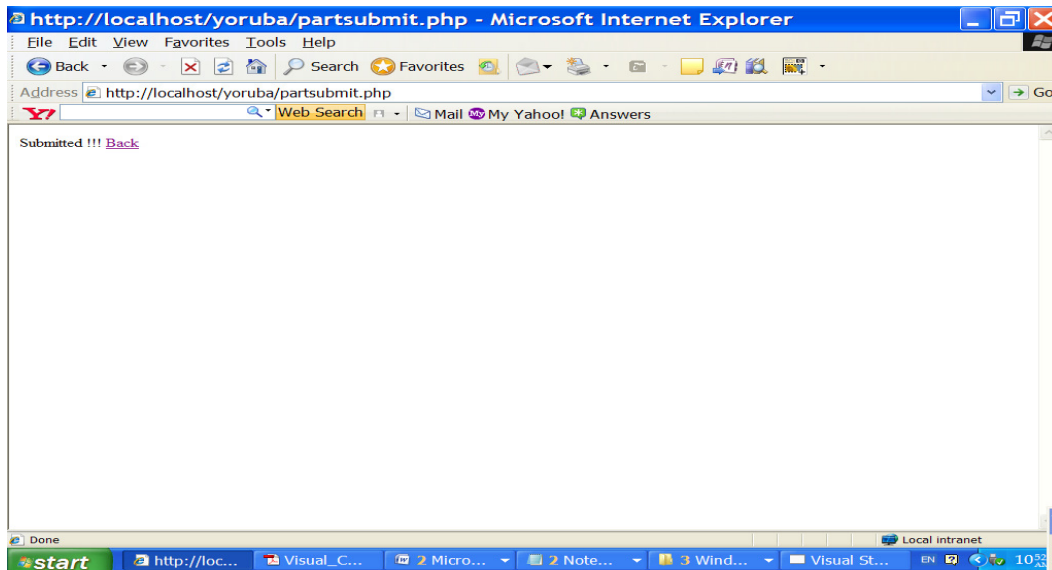The following page appears if not existing or already saved.

Aladesote Isaiah, Olaseni O. E, Adetunmbi A. O & Akinbohun Folake

FIGURE : D6

CONCLUSION

In conclusion, lexical analysis of a language can be done in three ways: by the use of Lexical-analyzer generator such as Lex compiler, by writing a conventional-programming language and by writing a lexical analyzer using assembly language.

The computational model of Yoruba language lexical analyzer was implemented and tested using a conventional programming language, which is a promising technique. The result obtained shows that the method adopted is satisfactory. This would be of great importance to researchers who are working on Yoruba Grammar.

REFERENCES

[1]   A dictionary of the Yoruba Language (2008), University Press Plc.

[2]   Adesuyan, A. "Agbeyawo Sintaasi EdeYoruba", Lekoba Publisher, Ibadan. 2003

[3]   Albert, S. (et al) "A Morphological Analyzer for Machine Translation based on Finite State Transducer", CICYT-FEDER

[4]   Awobuluyi, O. 1978. Essentials of Yoruba Grammar. Ibadan: Oxford University Press.

[5]   Bamgbose, A. "Linguistic in a Developing Country; University of Ibadan. Inuagural Lecture, Ibadan" University of Ibadan Press, 1973.

[6]   Bamgboye, A .1966. A Grammar of Yoruba. Ibadan: Cambridge University Press. Bamboye, A. 1967. A Short Yoruba Grammar. Ibadan: Heinemann Educational Books.

[7]   Bamgboye, A. 1990. Fonoloji ati Girama Yoruba. Ibadan: University Press Limited.

[8]   Brenziger, M. "Contribution on Endangered Language" **latiku** 1-5. New Hampshire 1998.

[9]   Bottler C.S (1990). Language and Computation. In N.E Collinge (ed): An Encyclopedia of Language. Routledge. London.

[10]   Bynon, T. 1977. Historical Linguistics. Cambridge: Cambridge University Press.

[11]   Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge: MIT Press.

[12]   De Gaye, J. A. and W. S. Beecroft. 1922. Yoruba Grammar. London: Routledge and Kegan Paul.

[13]   Hornby "Oxford Advanced Learner's dictionary of current English", Oxford
       University Press, 6th edition. (2000)

[14]   Hyman L.M. (2000). How to Become a *Kwa* Verb. *Symposium on areal Typology of west African
       Languages. Leipzig, 1-18*

[15]   Hacken P. Word Formation in Computational Linguistics. Proc. of TAN; Nancy.

[16]   International African Institute "Provisional Survey of Major Languages in the
       Independence States of Saharan Africa, P. Baker(ed.)." UNESCO: International
       African Institute.1980.

[17]   Jan. D. (1998), : "Finite State Automata " [Online] Available at http://www.pg.p1/~jandac/fsa.html
       Kirsten M. (1991). *The Linguistic Encyclopedia* (Kirsten Eds. London:Routledge).

[18]   Oluseye, A. (2003), " Yoruba: A Grammar Sketch" [Online} Available at
       www.igbimoedeyoruba.org

[19]   Owolabi, Kola. 1976. "Noun-Noun construction in Yoruba: A Syntactic and Semantic Analysis",
       PhD Dissertation, University of Ibadan, Nigeria.

[20]   Ozo – Mekuri Ndimele "A first course on Morphology & Syntax" Linguistc
       & Communication Studies, University of Port Harcourt, Nigeria. 1999

[21]   Williamson, K. and Blench, R. "Niger-Congo in Heine. African Languages: An
       Introduction, B. and D. Nurse (eds.). Cambridge: Cambridge University Press. 2000.

PROGRAMMING CODES

Analyzer.php

```php
<?
echo "<head><center><font color=red><u><h2>Yoruba Lexical
Analyzer</font></u></h2></head><form action='eyaitumoede.php' method='post'>
<br><br><br><table border=1><tr><td><h1>Gbolohun:</td><td><h1><textarea name=gbolohun
cols=30 rows=15></textarea></td></tr><tr><td><input type = submit name=submit value = Eya-
ede></td></tr></table></form>";

?>
```

Dbsubmitpart.php

```php
<?
$dbhost = "localhost";
$dbname = "eyaede";
$dbuser = "root";
$dbpass = "aladesote";
$dbtab = "ede";

$link = mysql_connect($dbhost, $dbuser, $dbpass) or die ("Could not connect to $dbname on
$dbhost with $dbuser@".$_SERVER[REMOTE_ADDR]);
//echo "Connected successfully to $dbname on $dbhost with
$dbuser@".$_SERVER[REMOTE_ADDR]."<br>\n";
mysql_select_db ($dbname) or die ("Could not select database named : $dbname" . mysql_error());
//echo "$dbtab";
?>
```

Storagepart.php

```php
<?
```

Aladesote Isaiah, Olaseni O. E, Adetunmbi A. O & Akinbohun Folake

```
include "dbsubmitpart.php";

$query = "SELECT * FROM $dbtab";
$result = mysql_query ($query) or die ("Query failed for table : $dbtab . " . mysql_error());
echo"<table border=1><tr><td>Yoruba</td><td>Part of Speech</td></tr>";
while($row=mysql_fetch_array($result))
{

echo "<tr><td>$row[0]</td>";
echo "<td>$row[1]</td></tr>";

}
echo"</table>";
?>
```

Entrypart.php
```
<?
echo"<form action='partsubmit.php' method='post'><table border=1><tr><td>Yoruba</td><td><input
type=text name=yoruba></td></tr><tr><td><select name=part><option value =
verb>Verb</option><option value =noun>Noun</option>
<option value = pronoun>Pronoun</option>
<option value = verb>Verb</option>
<option value = conjuction>Conjuction</option>
<option value = interjection>Interjection</option>
<option value = adverb>Adverb</option>
<option value = preposition>Preposition</option>
<option value = adjective>Adjective</option>
</select></td><td><input type=submit name=submit value=Store>
</tr></table></form>";
?>
```

Eyaitumoede.php
```
<?
include "dbsubmitpart.php";
include "analyzer.php";

if( $submit == "Eya-ede")
{
$a[]=array();
$sentence=$gbolohun;

                if($gbolohun == "" )
                {
                        echo "Ko eyi ti o fe lati tunmo";

                }
                else if($gbolohun != "")
                {

                        $a=split(" ",$sentence);

                        //echo $a[0].','.$a[1].','.$a[2].','.$a[3].','.$a[4];
                        //echo count($a);
                        echo"<body bgcolor=pink><br><br><h2><u>The result of the
analyzer:<table                                        border=1><tr><td bgcolor=light><font
color=white><h3>Oro ni Yoruba</td><td bgcolor=light><font color=white><h3>Eya-ede</td></tr>";
                        for($i=0;$i<=count($a);$i++)
                        {
                                $query[$i] = "SELECT eya FROM $dbtab where yoruba = '$a[$i]'";
                                $result[$i] = mysql_query ($query[$i]) or die ("Query failed for
        table : $dbtab . " .                                  mysql_error());
```

```php
                while ($row = mysql_fetch_array($result[$i]))
                {
                        $r= $row["eya"];
                        echo "<tr><td bgcolor=yellow><font color=blue><h3>$a[$i]</td><td
bgcolor=yellow><font color=green><h4>$r</td></tr>";
                }

        }
        echo"</table>";

        }

    }

?>
```

Homepage.php

```html
<html>
<head><h1><br><br><center><font color=green>Welcome to<br><font color=purple size=6> Lexical
Analyzer Center</font></h1></head><br><br>
<center><font size=4 color=blue><a href="analyzer.php"> Analyzer Scheme Center</a><br>
<a href="entrypart.php">Ikojopo Oro</a>
</html>
```

Partsubmit.Php

```php
<?
//echo "$part,$english,$yoruba";//
if ($submit == "Store")
{
        include "dbsubmitpart.php";

        //$query= "SELECT yoruba FROM $dbtab where yoruba = '$yoruba'";

$query= "SELECT * FROM $dbtab where yoruba='$yoruba'";
                        $result= mysql_query ($query) or die      ("Query failed for
        table : $dbtab . " .                         mysql_error());

while($row=mysql_fetch_array($result))
{
//echo "$row[0]<br>";
$temp=$row[0];
}
if( $temp == $yoruba)
        {echo "<br>Already stored<br><a href='entrypart.php'>Back</a>";}
else
        {
        $query = "INSERT INTO $dbtab values ('$yoruba','$part')";
        $result = mysql_query ($query) or die ("Query failed for table :    $dbtab . " . mysql_error());
        echo "Submitted !!! <a href='entrypart.php'>Back </a>";
        }

//}

}
else
{echo "Not submitted";}
?>
```

INSTRUCTIONS TO CONTRIBUTORS

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. Today, computational language acquisition stands as one of the most fundamental, beguiling, and surprisingly open questions for computer science. With the aims to provide a scientific forum where computer scientists, experts in artificial intelligence, mathematicians, logicians, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists can present research studies, International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches. IJCL is a peer review journal and a bi-monthly journal.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 2, 2011, IJCL appears in more focused issues. Besides normal publications, IJCL intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCL List of Topics:
The realm of International Journal of Computational Linguistics (IJCL) extends, but not limited, to the following:

- Computational Linguistics
- Computational Theories
- Formal Linguistics-Theoretic and Grammar Induction
- Language Generation
- Linguistics Modeling Techniques
- Machine Translation

- Models that Address the Acquisition of Word-order
- Models that Employ Statistical/probabilistic Gramm
- Natural Language Processing
- Speech Analysis/Synthesis
- Spoken Dialog Systems

- Computational Models
- Corpus Linguistics
- Information Retrieval and Extraction

- Language Learning
- Linguistics Theories
- Models of Language Change and its Effect on Lingui
- Models that Combine Linguistics Parsing

- Models that Employ Techniques from machine learnin
- Quantitative Linguistics
- Speech Recognition/Understanding
- Web Information

CALL FOR PAPERS

Volume: 3 - **Issue:** 1 - February 2012

i. Paper Submission: November 30, 2011   **ii. Author Notification:** January 01, 2012

iii. Issue Publication: January / February 2012

CONTACT INFORMATION