Journal of Library and Information Studies 12:1 (June 2014) p.1-19
doi: 10.6182/jlis.2014.12(1).001

Effects of Diacritics on Web Search Engines' Performance for Retrieval of Yoruba Documents

Toluwase Victor Asubiaro[1]

Abstract

This paper aims to find out the possible effect of the use or nonuse of diacritics in Yoruba search queries on the performance of major search engines, AOL, Bing, Google and Yahoo!, in retrieving documents. 30 Yoruba queries created from the most searched keywords from Nigeria on Google search logs were submitted to the search engines. The search queries were posed to the search engines without diacritics and then with diacritics. All of the search engines retrieved more sites in response to the queries without diacritics. Also, they all retrieved more precise results for queries without diacritics. The search engines also answered more queries without diacritics. There was no significant difference in the precision values of any two of the four search engines for diacritized and undiacritized queries. There was a significant difference in the effectiveness of AOL and Yahoo when diacritics were applied and when they were not applied. The findings of the study indicate that the search engines do not find a relationship between the diacritized Yoruba words and the undiacritized versions. Therefore, there is a need for search engines to add normalization steps to pre-process Yoruba queries and indexes. This study concentrates on a problem with search engines that has not been previously investigated.

Keywords: Information Retrieval; Information Retrieval Evaluation; Diacritics; Search Engines; Yoruba Language

1. Introduction

The World Wide Web (WWW) contains resources in and about diverse languages. The World Internet Statistics shows that English is the most popular language on the Internet (in content and usage) and probably the most popular for Internet searches; however, there has been an increase in the presence of non-English users on the information superhighway (Internet World Stats, 2010). The different computer encoding systems, based on orthography of some of these non-English languages, have

[1] E. Latunde Odeku Medical Library, College of Medicine, University College Hospital Campus, University of Ibadan, Ibadan, Nigeria
E-Mail: toluwaase@yahoo.ca

brought about language-dependent problems in Information Retrieval (Alpkocak & Ceylan, 2012). Yoruba for instance, whose orthography involves heavy use of diacritical marks (sub dot and tone marks), uses computer encoding system that accommodate diacritized versions of some America National Standard Institute (ANSI) characters. The base characters are diacritized to indicate tonality of Yoruba and to cater for the need to represent speech sounds that are beyond the range of the basic ANSI characters: á, à, è, é, ẹ, ẹ́, ẹ̀, í, ì, ò, ó, ọ, ọ̀, ọ́, ù, ú, ṣ, ń, and ǹ.

The conventional computer keyboard is based on the ANSI convention. Languages like Yoruba, whose orthography involve the use of characters beyond the ANSI scope, face the problem of inputting texts of the language into the computer with this keyboard. This has led to the development of language specific keyboards. However, Yoruba is a resource scarce language and its language dependent keyboards are not commonly available. Therefore, most writers either do not or partially (i.e. on choice words) append diacritics. This culminates in inconsistent adoption of the standard orthography of the language.

German and Finnish are European languages whose orthography also involves the use of diacritics. In these languages diacritics provide some morphological information. For instance, in German, *schon* means 'already', while *schön* means 'beautiful' and *Apfel* means 'apple', while Äpfel means apples. Therefore, omission or non-usage of diacritics in necessary words and on necessary characters amounts to some loss of information. In Yoruba the use of diacritics also provide morphological and lexical information. For instance, a Yoruba word such as *ogun* has four distinct variants that may be obtained with the use/non use of diacritics; they are the following: *ogun* 'war', ògùn- 'a river', ògún- '*orisha* of iron' and *ogún* 'inheritance'. Italian and French languages also use diacritics, but they do not carry morphological or lexical information.

For the languages that use diacritics, ignoring diacritics (using base characters to represent the diacritized versions) is viewed as a normalization process. In Information Retrieval, normalization is regarded as a text preprocessing, in order to allow more (and hopefully better) matches between query terms and document expressions. On the other hand, autodiacritization is performed on base words in order to give specific and contextual meanings to a word in the sentence/context in which it occurs. Although it appears that Information Retrieval community has not reported research on autodiacritization for query matching, it is probably a modality for achieving more precision.

The Yoruba language whose native name is "*ede Yoruba*" is a local African language that is spoken in West Africa. The native speakers of the Yoruba language occupy the southwestern part of Nigeria, the southern Benin Republic, and the southern Togo. A variety of the language called "*Lucumi*" or "*Nago*" is spoken as the sacred language of Santeria in Cuba, Puerto Rico, and the Dominican Republic. There are traces of the language in Sierra Leone where it is called "*oku*".

The main goal of the research work is to evaluate the performance of web search engines in retrieving Yoruba documents based on the use or non-use of diacritical marks on Yoruba queries. Three research questions were stated to guide this research. First, is there a significant difference between the number of hits returned by a search engine when diacritics are applied and when they are not applied on search queries? Second, is there a significant difference among the precision values of a search engine when diacritics are applied and when they are not applied on search queries? Third, is there a significant difference between the precision of all the search engines? It appears research work has not been carried out on the possible effects of diacritics on the performance of Information Retrieval Systems (IRS).

2. Related Works

There exist few studies that have adopted the use of search queries in languages other than English. Griesbaum (2004) is one of the few studies whose search queries are in a language other than English. The search queries of the study were in German. The orthography of German also includes the use of characters that are beyond ANSI, just like Yoruba. The study was on three German search engines: Google.de, Altavista.de, and Lycos.de. Like the Yoruba language, German also possesses characters that are beyond ANSI. Griesbaum observed that Google returned the highest values of search results, followed by Lycos, and then Altavista. Google also returned the highest number of relevant search results across all fifty queries used for the evaluation, with the top twenty precision values for nearly half of the queries. However, the differences between all the three search engines were low. The sign test was used to test for the significance of the differences; it indicated that Google performs significantly better than Altavista, but there was no significant difference between Google and Lycos. Lycos returned better search results than Altavista, but the differences between these two engines were not significant. Griesbaum concluded that the search engines from position one to twenty have very similar effectiveness when answering search queries, with the exception of Google, which seems clearly better.

Tawileh, Mandl and Griesbaum (2011) also carried out studies using Arabic search

queries. Five search engines, which included three international and two Arabic search engines, were evaluated. The test used fifty randomly selected queries from the top searches on the Arabic search engine Araby. The relevance of the top ten results and their descriptions retrieved by each search engine for each query were evaluated by independent jurors. Evaluations of results and descriptions were then compared to assess their conformity. The core finding was that Google performed better than the other engines almost all the time. The difference with Yahoo!, however, was not statically significant, and the difference to MSN, the third ranked engine, was significant to a low degree. The Arabic search engine Araby showed performance on most of the evaluation measures, while Ayna was far behind all other search engines. The other finding was the big differences between search results and their descriptions for all tested engines.

3. Test

3.1 Search engines selection

The most recent list of top search engines was collected from the search Engine Watch website; the four top search engines are AOL, Bing, Google, Yahoo! were used for the study (comScore.com, 2011; Rampton, 2011). All of the search engines are English and international. AOL is owned by Time Warner, Bing and Yahoo! are owned by Microsoft, while Google is owned by Google Inc.

3.2 Query formulation and research design

Text Retrieval Conference (TREC) style evaluation suggested the use of search logs for the purpose of creating topics for Information Retrieval evaluation (Harman et al., 2001). This was also supported by Lewandoski (2012), which recommended the use of collecting topics from the search logs of commercial search engines for evaluation. Therefore, the most searched keywords from Nigerian region and the most searched topics on Yoruba from Nigerian region in the last twelve months were collected with the use of Google search logs accessed through the Google Insight beta search engine services. Google Insight was used because it is an international search engine that is free and available to provide a list of most searched keyword according to region at the time this research was conducted. The list of the keywords produced thirteen unique and useful topics after the removal of reoccurring keywords and names of websites or organization whose direct interpretation may not exist in Yoruba. Search keywords like nollywood, borko haram, 2go, facebook, Yahoo!, Google and bbc fell into such category. Since the keywords were approximated and too general, using them for this study would have been inappropriate

because they are not informational queries. Therefore, they were expanded and constructed to informational search queries. At least two Yoruba informational search queries were constructed from each of the keywords. A total of thirty informational search queries were created from the most searched topics. Lewandoski (2012) reported that most newer research in search engines' evaluation use between twenty five and fifty search queries.

Each query was posed to the four search engines on the same day in order to minimize errors due to the rapidly changing nature of the WWW. The queries were first posed to the search engines with tone marks and later without tone marks. The first fifty hits were assessed for relevance. For queries that returned less than fifty hits, all the results were assessed for relevance.

3.3 Defining Yoruba documents?

A web search engine is meant to retrieve documents that match searchers' queries. Since queries are users' expression of information need and the users of search engines is "anybody", users are not restricted to experts or speakers of a particular language, neither is there a literacy level criterion imposed. Therefore, a user can best pose a query in his/her native language when the information need revolves around the language. Yoruba information is regarded as the information contained in retrieved document, if it is relevant to the searchers' query, regardless of its language of composition.

3.4 Evaluation criteria

The only evaluation criteria considered in this study is relevance. Most of the performance evaluation studies have used relevance as the basic evaluation criteria, although it has been criticized from different quarters because relevance was described to be "ambiguous" (Tawileh et al., 2011). To really measure relevance and reduce bias, cautions were taken. Search results were categorized using relevance scales based on Kumar and Pavithra (2010) methodology; the scales used are ***"relevant", "somehow relevant", "irrelevant", "relevant links",*** and **"*pages not found*"**.

A web page was considered **"*relevant*"** if its content was found to "match" or discuss the subject matter of the query and was assigned a weight of three. A web page was considered ***"somehow relevant"*** if the content of the web page was not wholly related to the subject matter of the search query, but the content contained some related information that could satisfy information needed about the subject of the query. ***"Somehow relevant"*** pages were assigned a weight of two. ***"Irrelevant"*** web pages were assigned weight of zero and refer to

the web pages whose contents were not related to the subject matter in the search query. The pages that contain links were also assessed for relevance; pages that contain at least one or two links to relevant documents were classified as ***"relevant link"*** and were assigned a weight of one, while a search result that contain links to documents that do not provide relevant information to the query are regarded as ***"irrelevant"*** web pages and were assigned a weight of zero. Also, pages that are not found, but were returned by the search engines are regarded as not found page and are assigned a weight of zero. If two web pages are extracted from the same website, they are counted as two.

Documents that were perceived to be useful but written in other languages that are not understood by the researcher, especially documents in languages other than Yoruba and English, were translated to English by Google translator before they were assessed for relevance.

3.5 Evaluation metric

The retrieval performance of the search engines were evaluated based on the precision of the search engines. Precision is regarded as one of the two classical retrieval performance measures in most retrieval performance studies (Griesbaum, 2004; Kumar & Pavithra, 2010).

3.5.1 Precision

Precision is the test of exactness. Shafi and Rather (2005) defined precision for web search engines as a "fraction of a search output that is relevant for a particular query". The calculation of the absolute precision of web search engines by Shafi and Rather (2005), Kumar and Pavithra (2010) that is used for this study is:

$$\text{Precision} = \frac{\text{Sum of the weights of relevant docs retrieved from a search engine}}{3\text{x(Total number of selected documents from document retrieved by a search engine)}}$$

Other evaluation metrics in other recent studies include recall, expected reciprocal rank (ERR), Mean Average Precision (MAP), Precision at rank k, Recall at rank k, Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG), mean average generalized precision (MAgP) for structured documents retrieval, inex_Eval metric, inex_eval_ng metric, XCG metric, T2I, HiXEval metric, EPRUM and GR metrics, Logistic Average Misclassification Rate (LAMR), Time Biased Gain (TBG), U-measure, and Retrieval Status Value (RSV) (Clarke, Craswell, & Voorhees, 2012; Harris & Srinivasan, 2012; Magdy & Jones, 2010; Pehcevski & Piwowarski, 2007; Radlinski & Craswell, 2010; Sakai & Dou, 2013). The metric employed for this study is deficient in

that it ignores ranking, and therefore is not appropriate for ranked retrieval systems. For instance, MAP makes it possible to track the precision of results considering the order of the returned results or the documents that come earlier. In this case, the probability of relevance of retrieved documents is not independent of the document that was retrieved earlier. Furthermore, other metrics that are rank based and useful in retrieval evaluation cannot be further elicited using this metric. Despite its shortcomings, precision is the most consistently used retrieval evaluation metric.

4. Results

4.1 Number of hits returned

Tables 2 and 3 show the number of hits returned by each search engine to queries posed. More hits were returned to undiacritized queries than diacritized queries. The diacritized query constituted 17.37% of the total number of retrieved sites by the four search engines in response to both the diacritized and undiacritized queries. Figure 1 presents the average number of hits returned by the search engines. In response to undiacritized, AOL outperformed all other search engines by returning 55.41% of the total hits of the four search engines with 21,773,155 hits, and 1,249 sites were selected for evaluation. Google followed with 13,475,357 retrieved sites, 34.29% of hits returned by the four search engines, while 1,355 sites were selected for the assessment. Though AOL returned more hits, more sites were used from Google's result because Google was more stable and retrieved

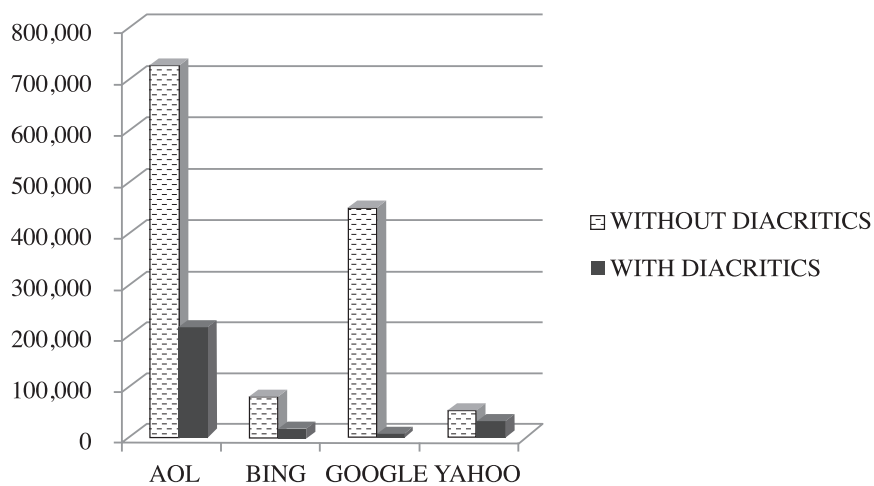

Figure 1. Average Number of Hits by Search Engines

Table 2. Number of Hits Returned for Diacritized and Undiacritized Queries

QUERY No.	AOL		BING		GOOGLE		YAHOO	
	Diacritized Queries	Undiacritized Queries	Diacritized Queries	Undiacritized Queries	Diacritized Queries	Undiacritized Queries	Diacritized Queries	Undiacritized Queries
1	272,000	216,000	413,000	406,000	15,600	254,000	418,000	401,000
2	280,000	366,000	105,000	289,000	11,500	270,000	106,000	281,000
3	1,040	89,000	101,000	92,000	256	150,000	101,000	163,000
4	25	19,700	449	30,500	8,820	22,700	148	10,500
5	1	4	8	8	6	29	8	8
6	5,770,000	19,900	44	33,100	20,300	22,600	45	10,700
7	4	14	28	28	1,720	240	28	28
8	0	57,400	0	39,400	0	10,800	0	15,600
9	5	9,290,000	10	47,800	985	9,250,000	0	8,250
10	131,000	430,000	8,420	480,000	130,000	679,000	239,000	239,000
11	1,360	131,000	9	118,000	235	11,000	9	120,000
12	4	34	1	46	22	228	0	46
13	0	4,800	0	9,540	0	4,480	0	8,260
14	4	7,460	26,000	26,500	14	7,740	7,540	25,800
15	0	34,200	0	5	0	20	0	5
16	7	10,600,000	6	85,500	235	1,990,000	6	80,300
17	5,270	405,000	28,100	748,000	5,610	737,000	12,400	235,000
18	0	272	2	437	3	1,600	2	1,890
19	19,700	88,400	384	28,400	18,600	26,000	61	9,420
20	0	26	0	35	0	337	0	36
21	2	38	3	26	45	381	3	26
22	1	34	3	14	2	26,100	13	13
23	25	853	8	54	66	14	11	121
24	12	36	6	564	55	430	6	25
25	4	12	0	10	12	37	0	9
26	1	694	4	473	3	847	4	127
27	10	445	34	34	1,930	504	34	35
28	0	11,500	0	24	0	9,170	0	25
29	0	332	1	81	3	95	1	22
30	0	1	0	2	0	5	0	2

Table 3. Number of Answered Queries by the Search Engines

Search Engines	Undiacritized Queries	Diacritized Queries
AOL	27	21
BING	26	23
GOOGLE	26	25
YAHOO	26	22

sites for queries rather consistently while AOL retrieved very high number of sites for certain queries. Bing and Yahoo! returned the least number of hits with 2,435,581 and 1,610,248 that constituted 6.2% and 4.1% respectively of the total number of hits returned by the four search engines. 1,182 and 1,130 sites were selected for the study from the hits returned by Bing and Yahoo! respectively.

For diacritized queries, AOL also outperformed the other search engines by returning 78.4% of the hits by of the four search engines with 6,480,475 hits; 505 sites were selected for study. After AOL, Yahoo! retrieved 10.70% of the results, with 884,319 hits; 570 results were selected for evaluation. Bing came third and returned 682,520 hits, which constitute 8.25% of the result by the four search engines; 567 sites were selected for evaluation. Though Google returned the least number of results, 216,022, just 2.61% of the total number of hits returned by the four search engines, it is noteworthy that more sites, 881, were selected for assessment from the retrieved sites by Google. It confirms the earlier observation that Google is more stable and consistently answered queries, unlike other search engines that returned very high results for certain queries and failed in others.

Research Question: *Is there significant difference between the number of hits returned by a search engine when diacritics is applied and when it is not applied on search queries?*

Wilcoxon signed rank test performed on the data show that there is no significant difference between the number of hits returned by AOL (p=0.084, Z=-1.731) and Google (p=0.187, Z=-1.321) when diacritics is applied and when it is not applied on search queries. Whereas there is a significant difference between the number of hits returned by Bing (p=0.001, Z=-3.473), and Yahoo! (p=0.001, Z=3.997) when diacritics is applied and when it is not applied on search queries.

4.2 Number of useful results and number of answered queries

It is possible for a search engine to return more hits to search queries, but the hits may be less useful to the searcher. The number of useful results refers to the number of results that at least contains links to relevant documents. The nature of the web has made the usefulness of links imperative in evaluating web search engines. Figure 2 shows that Google returned more useful results for the diacritized and undiacritized queries. For the undiacritized queries, Google returned a total of 343 useful results; AOL was the second best search engine with a total of 337 useful results, while Yahoo! and Bing followed with 287 and 286 respectively. For the diacritized queries, Google was also the best with a total of 240 useful results; Bing was second best with 180 useful results; AOL and Yahoo! followed with 177 and 176 respectively.

The number of answered queries depicts how helpful a search engine could be to a user. The number of answered queries refers to the number of queries that returned at least one relevant result. Table 3 shows AOL answered one query more than all other search engines for queries with diacritics. On the other hand, it shows that Google returned more answered queries for diacritized queries, closed followed by Bing, Yahoo! and AOL.

4.3 Precision

Google provided the most precise search results to the undiacritized Yoruba queries with 0.2 mean precision; Yahoo! followed with 0.183 precision; Bing and AOL followed with a mean precision of 0.180. Figure 3 presents the precision of the search engines for the undiacritized Yoruba queries.

Bing provided the most precise search results to the diacritized Yoruba queries with

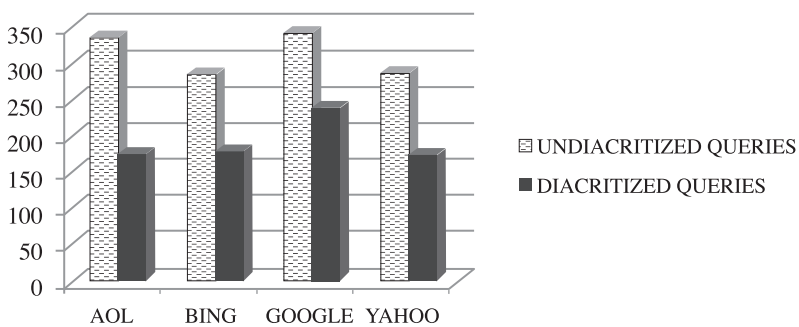

Figure 2. Useful Results for All the Queries

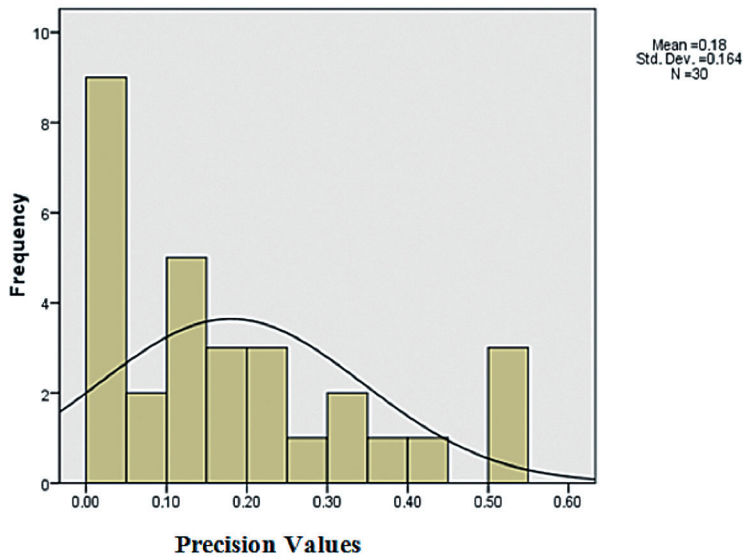

Figure 3. Precision Values of AOL for Undiacritized Queries

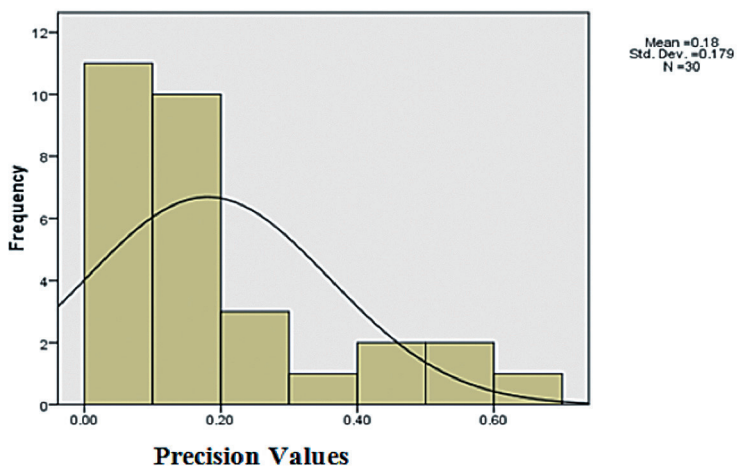

Figure 4. Precision Values of Bing for Undiacritized Queries

mean precision of 0.136, Google followed with a mean precision of 0.131 mean precision, Yahoo! with 0.127 mean precision, and AOL with 0.118 mean precision. Figure 4 presents the precision of the search engines for the keyword categories of the diacritized Yoruba queries.

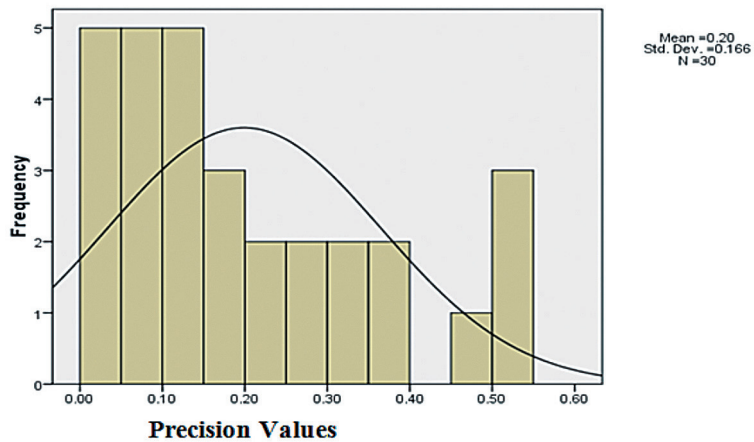

Figure 5. Precision Values of Google for Undiacritized Queries

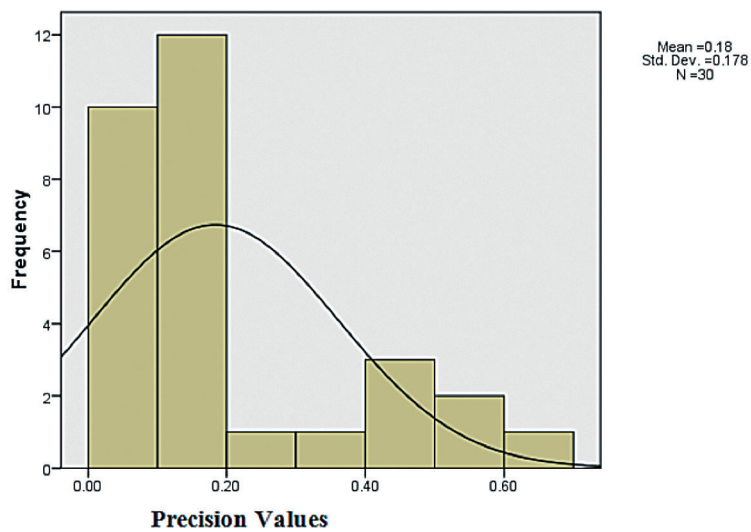

Figure 6. Precision Values of Yahoo for Undiacritized Queries

Research Question: *Is there a significant difference between the precision of any two of the search engines?*

The Friedman's test (p=0.759, Z=1.175) between the four search engines depicts that there is no significant difference between the effectiveness of any two the four search engines for the undiacritized search queries.

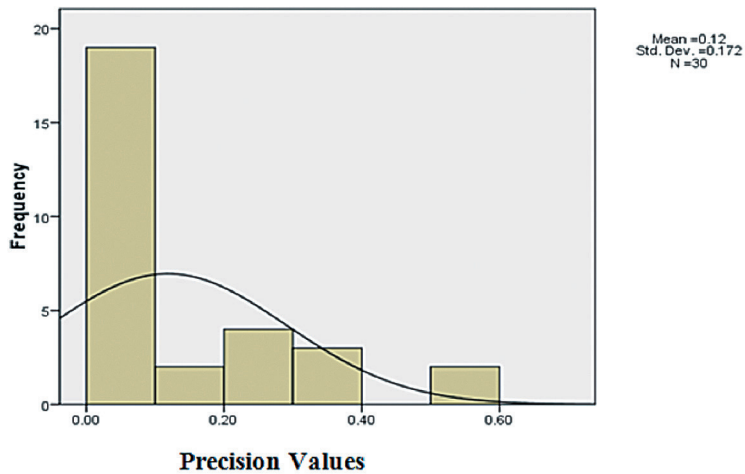

Figure 7. Precision Values of AOL for Diacritized Queries

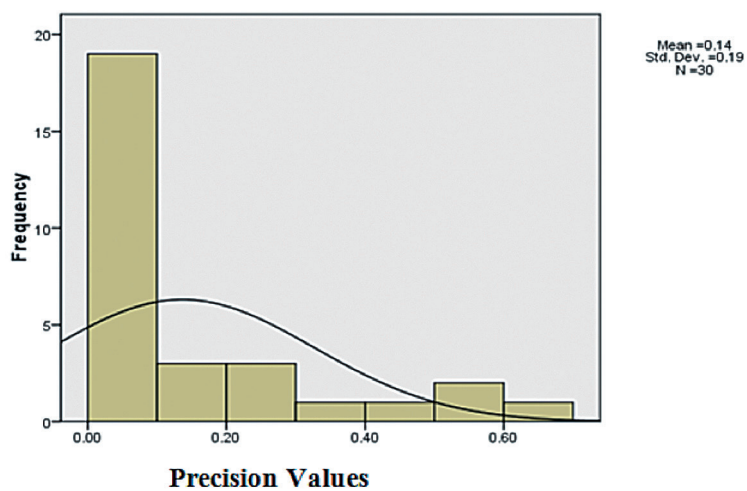

Figure 8. Precision Values of Bing for Diacritized Queries

The Friedman's test (p=0.679, Z=1.516) between the four search engines depicts that there is no significant difference between the effectiveness of any two of the four search engines for the diacritized search queries.

Research Question: *Is there a significant difference between the precision value of a search engine when diacritics is applied and when they are not applied on search queries?*

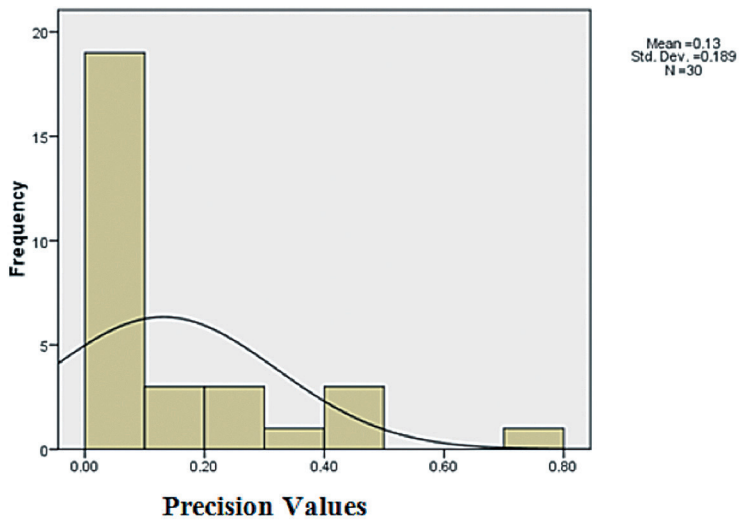

Figure 9. Precision Values of Google for Diacritized Queries

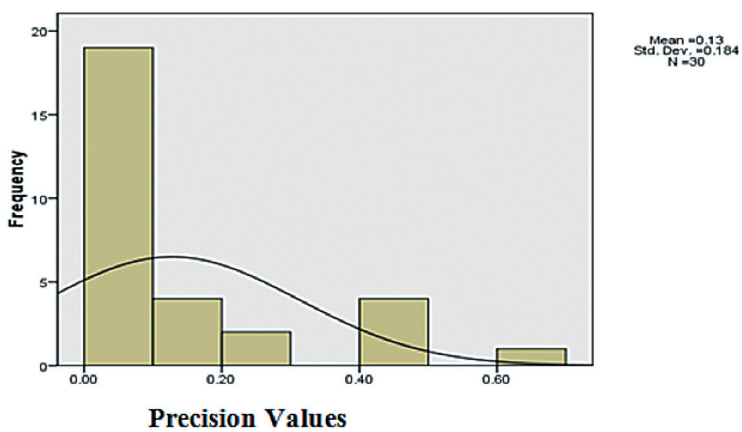

Figure 10. Precision Values of Yahoo for Diacritized Queries

Wilcoxon signed rank test show that there is significant difference in the effectiveness of AOL (p=0.0307, Z= -2.090) and Yahoo! (p=0.010, Z=-2.570) when diacritics is applied and when it is not applied. There is no significant difference in the effectiveness of Bing (p=0.178, Z=0.072) and Google (p=0.0072, Z=-1.802) when diacritics are and are not applied.

5. Conclusion and Recommendation

The results of the study show that Google is more effective than the other search engines. Although it did not return the highest number of hits--in fact it returned the lowest number of hits for diacritized queries-- it showed more consistency and stability in the distribution of the number of sites retrieved for assessment than the other three search engines. Furthermore, it provided more sites for assessment than any of the other search engines. The use of diacritics had an effect on the four search engines, the number of retrieved documents reduced with the queries with diacritics.

AOL retrieved more sites than other search engines for the diacritized and undiacritized queries, but it was found less useful and it returned less relevant results. Google was the second best in retrieving documents in response to Yoruba queries. Bing and Yahoo! were third and fourth in the rank. AOL answered just one query more than other three search engines for queries with diacritics, but Google answered two more queries than Bing. Yahoo! and AOL followed at third and fourth place. Google also returned more useful results; AOL came next and Yahoo! and Bing followed in that order.

Google returned far more precise results for both the diacrized and undiacrized queries than the other three; Yahoo! and Bing came second and third respectively, while AOL provided the least precise results for the two categories of search queries. The difference between the precision value of any of the four search engines for diacritized and undiacritized queries were found not to be significant. Also, the difference in the precision values of AOL, and Yahoo! when diacritics are applied and when it is not applied is significant, but insignificant for Bing and Google. The data gathered from Yahoo! and Bing are almost the same this is because they are both owned by Microsoft, therefore, the algorithm used might have been identical.

The differences in the performance of some of the search engines when diacritics are applied on search queries and when not applied suggest the search engines do not find relationship between diacritized texts and their undiacritized versions. It is therefore recommended that the search engines should normalize Yoruba search queries. Normalization is beyond removal of diacritics to form the base word, it goes ahead to establish functional relationships between the variants of such base word. Text normalization has been used in several languages and speech to text applications to modify textual representation of a speech sound. Text normalization is the transformation of words into a base form in order to establish relationship between terms from a common class. Normalization is

necessary for retrieval of Yoruba texts because the relationship between variants of a base word that is caused by the application of diacritics and the base word could be established for retrieval. Therefore, with normalization, a relationship will be established between undiacritized words (base words) and their diacritized versions (its variants). The process of grouping words in the a common class referred to as canonicalization. With canonization of Yoruba texts, auto-diacritization of undiacritized Yoruba queries will lead to optimal solution to retrieval problems caused by diacritics.

It is a deterrent that there are no Yoruba language search engines or specialized bilingual search engines for the language. Future work is expected on the application of normalization principles on Yoruba search queries for retrieval. One of the limitations of this research work is that the overlap between the search results from diacritized and undiacritized search queries was not sought, this could be worked upon in the future.

References

Alpkocak, A., & Ceylan, M. (2012). Effects of diacritics on Turkish information retrieval. *Journal of Electrical Engineering & Computer Science, 20*(5), 787-804. doi:10.3906/elk-1010-819

Clarke, L., Craswell, N., & Voorhees, M. (2012). *Overview of the TREC 2012 web track*. Paper presented at the Proceedings of The Twenty-First Text REtrieval Conference (TREC 2012), Gaithersburg, MD. Retrieved from http://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf

comScore.com. (2012). *comScore releases December 2011 U.S. search engine rankings*. Retrieved from http://www.comscore.com/Press_Events/Press_Releases/2012/1/comScore_Releases_December_2011_U.S._Search_Engine_Rankings

Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research, 9*(4). Retrieved from http://InformationR.net/ir/9-4/paper189.html

Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P., & Sheridan, P. (2001). CLIR evaluation at TREC. *Cross-Language Information Retrieval and Evaluation Lecture Notes in Computer Science, 2069*(2001), 7-23. doi: 10.1007/3-540-44645-1_2

Harris, C., & Srinivasan, P. (2012). *Using hybrid methods for relevance assessment in TREC crowd'12*. Paper presented at the Proceedings of The Twenty-First Text REtrieval Conference (TREC 2012), Gaithersburg, MD. Retrieved from http://trec.nist.gov/pubs/trec21/papers/UIowaS.crowd.final.pdf

Internet World Stats. (2010). *Internet world users by language top 10 languages*. Retrieved from http://www.internetworldstats.com/stats7.htm

Kumar, B. T. S., & Pavithra, S. M. (2010). Evaluating the searching capabilities of the search engines and meta search engine: A comparative study. *Annals of Library and Information Studies, 57*, 87-97.

Lewandoski, D. (2012). A framework for evaluating the retrieval effectiveness of search engines. In C. Jouis, I. Biskri, J.-G. Ganascia, & M. Roux (Eds.), *Next generation search engines: Advanced models for information retrieval* (pp. 456-479). Hershey, PA: IGI Global. doi: 10.4018/978-1-4666-0330-1.ch020

Magdy, W., & Jones, G. (2010). *A new metric for patent retrieval evaluation*. Paper presented at the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10), Milton Keynes, United Kingdom.

Pehcevski, J., & Piwowarski, B. (2007). Evaluation metrics. In L. Liu, & M. T. Özsu (Eds.), *Encyclopedia of database systems*. Boston, MA: Springer US.

Radlinski, F., & Craswell, N. (2010). Comparing the sensitivity of information retrieval metrics. In *SIGIR '10: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 667-674). New York, NY: ACM. doi: 10.1145/1835449.1835560

Rampton, J. (2011). *comScore.com: Bing takes no. 2 spot from Yahoo! in December 2011*. Retrieved from http://searchenginewatch.com/article/2137562/comScore-Bing-Takes-No.-2-Spot-From-Yahoo!-in-December-2011

Sakai, T., & Dou, Z. (2013). Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *SIGIR'13: Proceedings of the 36th International ACM SIGIR conference on research and development in information retrieval* (pp. 473-482). New York, NY: ACM. doi: 10.1145/2484028.2484031

Shafi, S. M., & Rather, R. A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology, 2*(2). Retrieved from http://www.webology.org/2005/v2n2/a12.html

Tawileh, W., Mandl, T., & Griesbaum, J. (2011). *Evaluation of five web search engines' in Arabic language*. Paper presented at the LWA in Kassel, Hessen, Germany.

(Received: 2013/4/25; Accepted: 2013/8/13)

Appendix

Table 1. Keywords

Search Keywords from Google Insight		Informational Queries	
		Generated without Diacritics	Generated With Diactics
1 Yoruba video	1	fiimu Yoruba	Fîmù Yorùbá
	2	fiimu agbelewo Yoruba	Fîmù àgbéléwò Yorùbá
2 What is Yoruba	3	Eya Yoruba	ẹ̀ yà Yorùbá
	4	Onka Yoruba	oǹkà Yorùbá
3 Yoruba actress	5	Osere Yoruba	òṣèré Yorùbá
	6	Fidio Yoruba	fídíò Yorùbá
4 Yoruba language	7	Ede Yoruba	Èdè Yorùbá
	8	Orisa ogun	Òrìshà ògún
	9	Ikini ni ile Yoruba	Ìkíni ní ilẹ̀ Yorùbá
5 Yoruba history	10	Itan isedale Yoruba	Ìtàn ìṣẹ̀dáleẹ̀ Yorùbá
	11	Aare ona kakanfo	Àárẹ ọ̀nà kakanfò
6 Yoruba bible	12	Bibeli Yoruba	Bíbélí Yorùbá
	13	Olodumare	Olódùmarè
	14	Esin kristieni	Ẹ̀sìn krìstíẹ́nì
7 Yoruba culture	15	Oriki	Oríkì
	16	Isin orisa	Ìsìn Òrìṣà
	17	Asa Yoruba	Àṣà Yorùbá
8 Yoruba names	18	Isomoloruko	Ìsọmọlórúkọ
	19	Awon oruko Yoruba	Àwọn orúkọ Yorùbá
9 Yoruba dictionary	20	Iwe atumo ede Yoruba	Ìwé atúmọ́èdè Yorùbá
	21	Aayan ogbufo Yoruba	Aáyan ògbufọ̀ Yorùbá
10 Nigeria	22	Itan ominira naijiria	Ìtàn Òmìnìra nàìjíríà
	23	Eto Ijoba naijiria	Èto ìjọba nàìjíríà
	24	Ile igbimo asofin	Ilé ìgbìmọ̀ aṣòfin nàìjíríà
11 Love	25	Ife omonikeji	Ìfẹ́ ọmọnìkejì
	26	Ife ninu igbeyawo	Ìfẹ́ nínú ìgbeyàwó
12 News Nigeria	27	Iroyin Ere idaraya	Ìròyìn Eré Ìdárayá
	28	Ijoba aare Jonathan	Ìjọba àárẹ́ Jonathan
13 Pictures	29	Ise Aworan yiya	Iṣẹ́ àwòrán yíyà
	30	Aworan lori intaneeti	Àwòrán lórí Íntánẹ́ẹ́tì

Note. Search keywords collected from http://www.Google.com/insights/search/ on 14/08/2012

圖書資訊學刊　第12卷第1期　（103.6）　頁1-19
doi: 10.6182/jlis.2014.12(1).001

變音符號對搜尋引擎檢索約魯巴語文獻表現之成效

Effects of Diacritics on Web Search Engines' Performance for Retrieval of Yoruba Documents

Toluwase Victor Asubiaro[1]

摘　要

本研究目的在於了解使用變音符號與否，是否影響搜尋引擎（AOL、Bing、Google、Yahoo!）搜尋約魯巴語文獻之成效。本研究自Google search logs整理奈及利亞最常使用的關鍵字，制訂30題約魯巴語問項，包含使用變音符號與未使用變音符號兩類，做為研究之關鍵字彙。研究結果顯示，未使用變音符號之關鍵字彙在所有搜尋引擎中皆獲得較多結果；在準確率（precision values）上，是否使用變音符號，則在AOL和Yahoo!相比時出現顯著差異。本研究結果指出，是否使用變音符號，確實影響搜尋引擎檢索約魯巴語文獻之成效。本研究建議，搜尋引擎有必要針對約魯巴語之問項與索引，預先進行正規化。

關鍵字：資訊檢索、資訊檢索評估、變音符號、搜尋引擎、約魯巴語

[1] 奈及利亞伊巴丹大學醫學院E. Latunde Odeku醫學圖書館
E. Latunde Odeku Medical Library, College of Medicine, University College Hospital Campus, University of Ibadan, Ibadan, Nigeria
E-Mail: toluwaase@yahoo.ca

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：Asubiaro, T. V. (2014). Effects of diacritics on web search engines' performance for retrieval of Yoruba documents. *Journal of Library and Information Studies, 12*(1), 1-19. doi: 10.6182/jlis.2014.12(1).001
以Chicago格式引用本文：Toluwase Victor Asubiaro. "Effects of diacritics on web search engines' performance for retrieval of Yoruba documents." *Journal of Library and Information Studies* 12 no.1 (2014): 1-19. doi: 10.6182/jlis.2014.12(1).001