Exploring Topic-Language Preferences in Multilingual Swahili Information Retrieval in Tanzania

JOSEPH P. TELEMALA, Department of Computer Science, University of Cape Town, South Africa.
HUSSEIN SULEMAN, Department of Computer Science, University of Cape Town, South Africa.

Habitual switching of languages is a common behaviour among polyglots when searching for information on the Web. Studies in information retrieval (IR) and multilingual information retrieval (MLIR) suggest that part of the reason for such regular switching of languages is the topic of search. Unlike survey-based studies, this study uses query and click-through logs. It exploits the querying and results selection behaviour of Swahili MLIR system users to explore how topic of search (query) is associated with language preferences – topic-language preferences. This paper is based on a carefully controlled study using Swahili speaking Web users in Tanzania who interacted with a guided multilingual search engine. From the statistical analysis of queries and click-through logs, it was revealed that language preferences may be associated with the topics of search. The results also suggest that language preferences are not static; they vary along the course of Web search from query to results selection. In most of the topics, users either had significantly no language preference or preferred to query in Kiswahili and changed their preference to either English or no preference for language when selecting/clicking on the results. The findings of this study might provide researchers with more insights in developing better MLIR systems that support certain types of users and in certain scenarios.

CCS Concepts: • **Information systems → Information retrieval**; Specialized information retrieval; Structure and multilingual text search; **Multilingual and cross-lingual retrieval**.

Additional Key Words and Phrases: Language Preferences, Topic-Language, Swahili, MLIR, Guided Multilingual Search.

ACM Reference Format:
Joseph P. Telemala and Hussein Suleman. 2021. Exploring Topic-Language Preferences in Multilingual Swahili Information Retrieval in Tanzania. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 1, 1, Article 1 (January 2021), 30 pages. https://doi.org/10.1145/3458671

1 INTRODUCTION

Multilingual information retrieval (MLIR) is a sub-discipline of information retrieval (IR) dealing with retrieval of information written and/or stored in a language different from the searcher's query language. Fluhr et al. [1999] define MLIR as a system that can process a query and return results/document in essentially any language. Supposedly, MLIR systems enable a Web searcher to get more comprehensive results than a monolingual IR system, taking advantage of a vast amount of information available in other languages, in addition to the language of the query [Rahimi et al. 2015].

Authors' addresses: Joseph P. Telemala, Department of Computer Science, University of Cape Town, Private Bag X3 Rondebosch, Cape Town, Western Cape, 7701, South Africa., jtelemala@cs.uct.ac.za; Hussein Suleman, Department of Computer Science, University of Cape Town, Private Bag X3 Rondebosch, Cape Town, Western Cape, 7701, South Africa., hussein@cs.uct.ac.za.

2375-4699/2021/1-ART1 $15.00
https://doi.org/10.1145/3458671

MLIR systems have been embraced by multiple language speakers (polyglots) especially in digital library settings, for example in works by Alsalmi [2019] and Wu and Chen [2019], as well as general Web search [Ling et al. 2018a]. This may be attributed to the fact that most people in the world are multilingual. As a result, the IR research community has had an increased interest in developing algorithms and systems to better retrieve information from other languages apart from English, or at least, along with English.

The considerable challenge associated with retrieving information from a language different from that of the query is the appropriate query or document translation [Rahimi et al. 2015]. Perfect translation, including context, may mean perfect retrieval (high precision), at least, in the system or technical perspective. However, users in MLIR also want a system that considers their search behaviour in terms of preferences and experiences [Chu and Komlodi 2017; Nzomo et al. 2019]. Studies, mostly in IR, such as those by Aula and Kellar [2009] and Steichen et al. [2014] suggest that one of the common behaviours among polyglots is the frequent switching of languages at different points in the course of Web search. More studies allude that part of the reason for such regular switching of languages in the Web search is the topic of search [Aula and Kellar 2009; Ling et al. 2020; Steichen et al. 2014; Telemala and Suleman 2018]. It is reasonable to assume that such behaviour may replicate when a polyglot Web user interacts with a multilingual search engine.

Thus, focusing on MLIR systems, take an example of a user with an information need in topic T_i, say *tourism*, and two or more language options for use, L_i and L_j, say *Kiswahili* and *English* respectively. Since MLIR supports retrieval of documents in multiple languages, even when this user prefers to query the system with a language L_i, the system retrieves and displays results in both L_i and L_j. The user then clicks on the most relevant result(s), R, oblivious of the language of the query, such that clicked results/URLs, R_x, are from both languages.

It can soundly be assumed that, such interactions from many users of the MLIR and over a particular period of time, may result in patterns (associations). The associations can be between the query (topic) and: i) query language; and ii) language of the results. This paper refers to these patterns as *Topic-Language (T-L) association*, which then leads to a notion of *topic-language preferences* explored and presented in this paper.

This study explores these preferences via exploiting the querying and results selection behaviour of Swahili speaking MLIR system users to understand the association between a topic of search and the preferred language of (i) query and (ii) results. In other words, the objective of this study is to explore T-L association/preferences using query and click-through logs of a MLIR system. The study uses polyglot Web users from Tanzania, an East African multilingual country, where Kiswahili and English are both official languages.

While a small part of this study is based on a questionnaire, the major part is carefully controlled in which participants interacted with a guided multilingual search engine. From both the questionnaire and the users' interaction with the guided search engine, the study endeavoured to answer the following research questions:

RQ1 How do polyglot Swahili speaking Web users rate themselves on the use of English and Kiswahili on Web search?

RQ2 What is the preferred query language among the polyglot Swahili speaking MLIR system users?

RQ3 What is the preferred language of results among the polyglot Swahili speaking MLIR system users?

RQ4 Do the topic-language preferences change at different stages (query to results selection) of MLIR user interaction?

The rest of the paper is organized as follows. Section 2 reviews the related literature. Experimental setup, materials and methods applied in this research are explained in Section 3, followed by the presentation of results in Section 4. Section 5 is about summary and discussions of the results and section 6 Section 6 presents the concluding remarks. Additionally, supplementary materials used in this paper are given in Appendix A.

2 RELATED WORK

Currently, Web search engines present the users with an option to demonstrate or implicitly learn their choice/preference either in terms of the language of the interface or content, layout, themes and system configurations and customization [Chu and Komlodi 2017; Ling et al. 2018b]. Users are reported to switch from one search engine to another due to reasons such as popularity of the search engine or usability of the interface, locality, or search quality, which determines the user satisfaction [Guo et al. 2011]. Language switching – *code-switching*, also called *language alternation* – is a common behaviour of polyglot users when interacting with IR systems [Aula and Kellar 2009].

This study, however, is interested in the language-switching behaviour of polyglots for their search but not of the interfaces of the search engines. Thus, this section presents a review of a number of works that investigated the reasons for such behaviour of polyglots in both classic IR and MLIR settings. The reasons vary from simple ones such as translation purposes to complex ones such as availability of resources. And the methodology for these studies differs from one study to another.

2.1 Reasons for Language Switching and Preferences

In a controlled laboratory experiment, Aula and Kellar [2009] reported that the availability and the quality of information are the major reasons for language switching. The most recent study utilizing the same setting of controlled laboratory experiments, supplemented by interviews, was done by Wang et al. [2018]. Wang et al. divided the notion of code-switching into two categories: situational and metaphorical code-switching. The findings on the reasons for situational code-switching do not differ from the study by Aula and Kellar and other studies discussed below. Factors like language proficiency, information verification, context, and translation purposes were reported.

However, it is worth noting the interesting factors for metaphorical code-switching, because they mostly have to do with only the image and perception the searcher has in mind. Such factors include: perception that results in one language are *accurate and objective*; *sense of belonging* when using a particular language e.g. mother tongue; *credibility and user trust* of the website; and *psychological* reasons.

In addition to the findings by Wang et al. [2018] about sense of belonging for native language, Lowe and Steichen [Lowe and Steichen 2017] observed that any multilingual speaker significantly uses his/her native languages, and that, language preferences depend highly on an individual's characteristics and the type of task they want to achieve. Furthermore, using an online questionnaire, Vassilakaki et al. [2015] observed that users always prioritize their mother language even when there is inadequate/limited information available on the Web. Web users prioritizing their native language may be attributed to struggling in foreign language proficiency, most especially the query formulation problem for non-native speakers as revealed by Nzomo et al. [2016] in a survey on bi/multilingual university students. Even so, a study by Berendt and Kralisch [2009] reported contrary findings in which users presented a tendency of accepting information in English compared to their native languages on the Web.

The same study by Vassilakaki et al. also reported that the purpose of information a user is looking for determines the language to use at a particular moment. Marlow et al. [2008] revealed that language skills have an impact on searching experience in a multilingual search. Using diary interviews, Wang and Komlodi [2018] established several reasons such as: need for translation; availability of resources; language proficiency; and context of the information sought e.g. news and entertainment and social networking.

Focusing on digital multilingual library users, Clough and Eleta [2010] used a questionnaire to examine if two specific factors for language choice – language skills and field of knowledge of the user – are significantly correlated, varying between different fields of knowledge. In a survey of browsing and search behaviour of polyglots in multilingual search engines, Steichen et al. [2014], including the follow-up works [Ling et al. 2018a; Steichen and Freund 2015], revealed that the context of search, such as usage purpose of the information sought and topic domain have a great influence on the choice of language for daily browsing and searching.

A recent study by Ling et al. [2020] used a crowd-sourcing approach to explore the behaviour of users on multilingual news consumption. They revealed that the news topic domain determines the search language.

Table 1. A Summary of factors for code-switching among polyglots

No.	Factor	Studies
1.	User's language skills and proficiency	[Clough and Eleta 2010; Marlow et al. 2008; Steichen et al. 2014; Wang and Komlodi 2018]
2.	Knowledge and profession of the user	[Clough and Eleta 2010; Si et al. 2017; Telemala and Suleman 2018]
3.	Resources availability	[Aula and Kellar 2009; Kralisch and Mandl 2006; Telemala and Suleman 2018; Wang and Komlodi 2018]
4.	Query formulation challenges	[Nzomo et al. 2016; Si et al. 2017]
5.	Search context	[Telemala and Suleman 2018; Vassilakaki et al. 2015; Wang and Komlodi 2018]
6.	Topic domain	[Ling et al. 2020; Lowe and Steichen 2017; Steichen and Freund 2015; Steichen et al. 2014; Telemala and Suleman 2018; Wang et al. 2018]
7.	Task type	[Steichen et al. 2014]
8.	Information quality and accuracy	[Aula and Kellar 2009; Kralisch and Mandl 2006]
9.	Information verification	[Wang et al. 2018]
10.	Translation purposes	[Wang et al. 2018]
11.	Beliefs, credibility and user trust	[Berendt and Kralisch 2009; Lowe and Steichen 2017; Vassilakaki et al. 2015]

2.2 Summary and Synthesis

In summing up, these studies revealed the following reasons/factors for code-switching behaviour in information search shown in Table 1. The most important observation that this study was further interested in, is the code-switching because of the topic domain (topic of search), revealed in several studies such as: Steichen et al. [2014]; Steichen and Freund [2015]; Lowe and Steichen [2017]; Wang et al. [2018]; Ling et al. [2020]; and Telemala and Suleman [2018].

However, survey-based studies, mostly observed in the reviewed works, have issues. Vigo et al. [2019] warn that survey-based studies are unreliable due to self-reporting biases. Furthermore, survey-based studies on human behaviour in Web search: are costly, for large scale data collection; cannot scale for a large geographical region; and are static, as they represent a human behaviour at a specific point in time [Mueller et al. 2017].

This brings a set of fundamental questions such as: i) can the code-switching due to topic (topic-language preferences) in MLIR be identified using the cheaply and massively available click logs?; and ii) are there other latent behaviours related to topic-language preferences in the MLIR click logs? In that regard, this study uses click-through data from a guided MLIR system to perform an analysis and address these questions. We find out that the topic of search is indeed a reason for language switching, and thus, we suggest the use of topic-language preferences in improving relevance of MLIR results.

3 METHODOLOGY

3.1 Setting up the Corpus of Topics and Queries

This subsection details the preparation of search topics and queries used in this research. The study used Tanzania, an East African country, as a case study, due to her large population of Swahili speakers and at the same time the status of Kiswahili and English as official languages. The multilingual nature of the country and the official status of Kiswahili and English guarantees that Web searchers are likely to use both languages. All the topics and queries prepared had an affiliation to Tanzanian Web searchers or, at least, originated from Tanzania.

Participation in this study was entirely unpaid. To save participants' time, it was necessary to ensure participants do not spend much time thinking of scenarios and queries to search from. There are several other studies in IR and MLIR, such as Ling et al. 2020, Lowe and Steichen 2017 and Yamamoto and Yamamoto 2020, where tasks and topics or queries were prepared before hand. The prepared tasks and topics/queries indicate (simulated) information needs.

Web directories organize websites according to major themes (topics) and sub-themes (sub-topics) of the information the websites contain. For example, a tourism website is organized around a tourism theme. The Web directories, specific to Tanzania's websites, and Google Trends[1] were used to identify diverse topics on the Web. The Web directories with a good coverage of Tanzania's websites used in this study are: Alexa[2]; 123Tanzania[3]; Yalwa[4] and, the deprecated, WWW Virtual Library[5]. Google Trends, on the other hand, offers a high level way to analyse the trending queries on the Web. It categorizes search topics and queries as Web, image, news, Google shopping and YouTube. The interest of this study was on the Web search operations.

To ensure that all the queries were in the geographical region of Tanzania and that the topics are not just mentioned in the Web directories but also are used in the region under study, the configurations of the Google Trends Explore[6] system were as follows: category – Web search; location – Tanzania; and duration – 2004 to 2019. Then, we queried each of the topics obtained

[1] https://trends.google.com/trends/

[2] https://www.alexa.com/topsites/category/Regional/Africa/Tanzania

[3] www.123tanzania.com

[4] https://www.yalwa.co.tz/

[5] http://vlib.org/

[6] https://trends.google.com/trends/explore

from the Web directories against Google Trends Explore to identify the related queries in each of the topics.

From the Google Trends Explore system, topics and their respective queries were exported as comma-separated values (CSV) files, then combined into a single file. This was followed by removing all single-word queries due to their (deemed) ambiguity, [Jansen and Spink 2006; Sanderson 2008], even when translated. For example, an English query *apple*, may imply a *fruit* or a *technology company*, but it may only translate to *tufaa* in Kiswahili. This translation only accounts for the fruit meaning, leaving out important and relevant results related to the information about the technology company.

Only topics with at least two queries were retained for further processing. Topics with only one query were removed; regarding such topics as less important to the community under study. Furthermore, queries under one topic but with slight variations in either the spelling or pre- and post-fixes or related information needs were merged. This produced a total of 1184 queries on 123 different topics.

Owing to the reason that Bing Web Search API[7] was used for retrieving results from the Web, we mainly used Bing Microsoft Translator[8] to translate all queries from one language to another i.e. if a query was in Kiswahili, it was translated to English or vice-versa. In cases with translation problems such as term ambiguity and lexical-semantic issues, Google Translate[9] was used for verification and/or as an alternative.

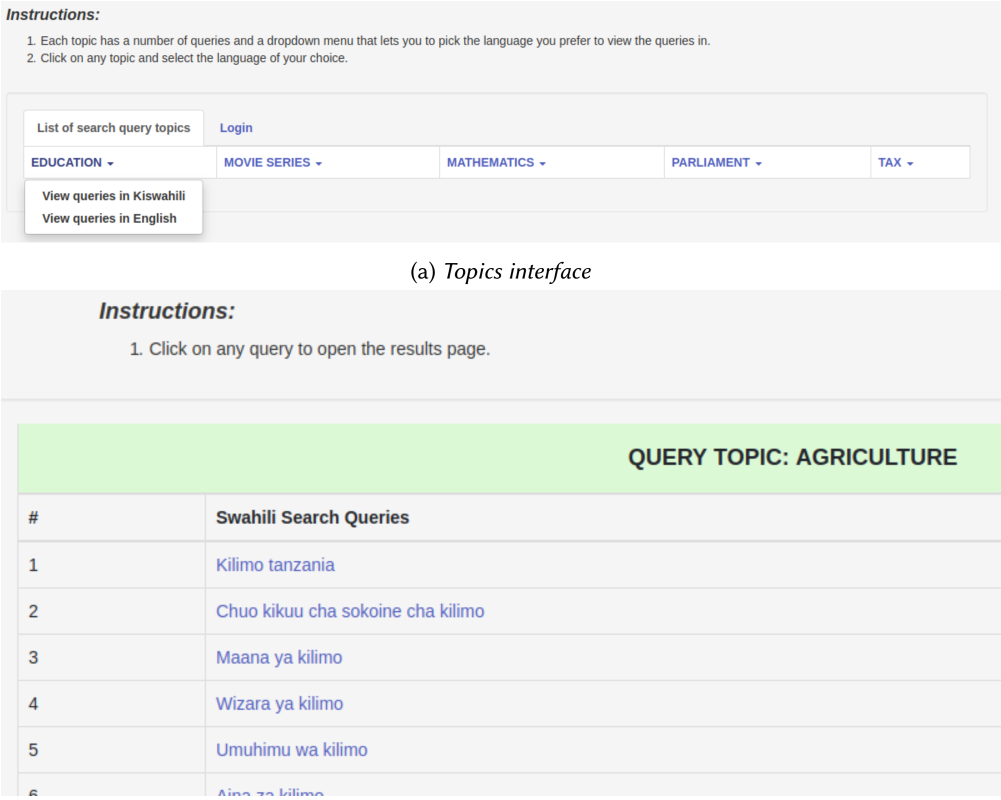
3.2 Data Collection Platform

The platform for collecting data had 3 main sections: *demographics*; *topic and queries system* and the *search engine interface*. Before accessing the demographics page, each participant had to sign a consent form at the index page of the platform[10]. The system then redirected a user to a page that collected a few personal demographics details, namely: sex, age group, education level, and occupation. On the same page, participants were also required to rate their use of English and Kiswahili when searching for information on the Web, on a scale of 5 such that: 1 - never; 2 - rarely; 3 - sometimes; 4 - often; and 5 - always use the language for Web search.

The topic and queries system section had two pages: topics and queries. A user was presented with five randomly generated topics, from among the 123 topics mentioned above (see Figure 1a for an excerpt of sample topics). A random display of five topics per user session ensured that users were not fed up or confused with all the 123 topics available, and at the same time ensuring equal chances of selection for each topic from the users. The user was the instructed to click on a topic of his/her interest from among the displayed topics, then asked to choose, from a drop-down list attached at each of the displayed topics, the language for viewing queries. Only two language options were provided: English and Kiswahili. Note that, for the purpose and settings of this study, the language chosen is referred to as a *query language*.

After choosing the preferred query language, the queries page opens up, displaying only queries in the selected topic and in the selected language of the user. To avoid the bias of users only choosing short and/or easy to type queries, each query had an embedded clickable link to the search engine interface. Users were instructed to click on a query of their choice to open the results page (search engine interface). Refer to Figure 1b for an example of displayed Swahili queries in the Agriculture topic.

[7] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/
[8] https://www.bing.com/translator
[9] https://translate.google.com/
[10] http://simba.cs.uct.ac.za/~joseph/



Instructions:

1. Each topic has a number of queries and a dropdown menu that lets you to pick the language you prefer to view the queries in.
2. Click on any topic and select the language of your choice.

(a) *Topics interface*

Instructions:

1. Click on any query to open the results page.

QUERY TOPIC: AGRICULTURE

#	Swahili Search Queries
1	Kilimo tanzania
2	Chuo kikuu cha sokoine cha kilimo
3	Maana ya kilimo
4	Wizara ya kilimo
5	Umuhimu wa kilimo
6	Aina za kilimo

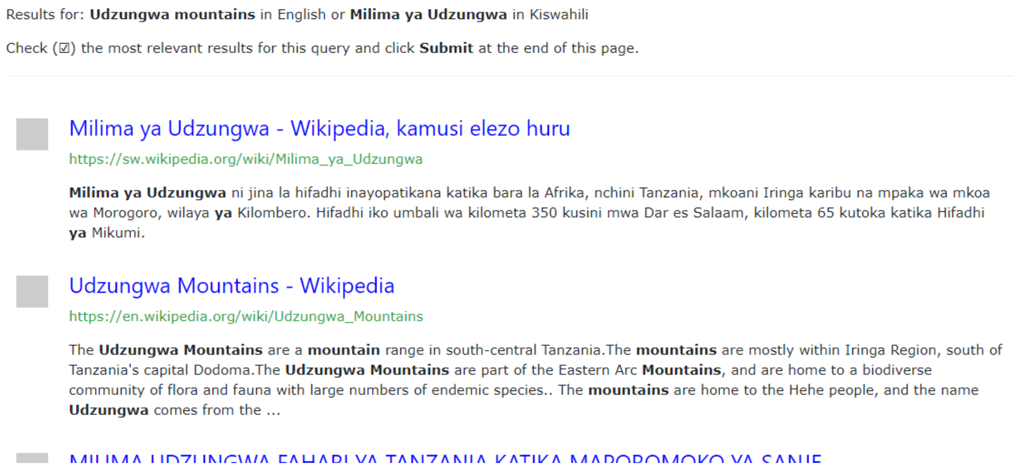
(b) *Queries interface*

Fig. 1. Topic and Queries interfaces of the guided multilingual search system.

The last part of the platform was the search engine results page, which used Microsoft Bing Web Search API at the back-end to retrieve (multilingual) results. Search results were presented in an interleaved round-robin style in both English and Kiswahili regardless of the query language the user specified when querying. See Figure 2 for an excerpt of the displayed results. To counter the effect of the position bias of results in one language from always appearing as the first, languages of the results appearing at the first position were randomly alternated for each session.

Even though findings about the multilingual display/interface preferences reported by Ling et al. [2018a] suggest that the panel style was mostly preferred, one may argue that this style may introduce bias due to layout. For instance, many Web users may be inclined to the results on the left panel (first results to display) and give less attention to the ones on the right, treating the ones on the right as extras/additional. Additionally, the participants are used to monolingual style of results presentation, thus, the look and feel of a monolingual search engine is appealing for new MLIR users such as the ones used in this study. This way, the participants can make up their minds based on the language of the results rather than being influenced by the results presentation layout.

From Figure 2, it can be observed that the search bar was hidden from the users to restrict query modification and/or users creating their own queries. The clickable links on the results titles were also disabled in such a way that users could not navigate to one result or open several tabs for the most relevant results. Instead, users were supposed to inspect the results based on the snippets and



Results for: **Udzungwa mountains** in English or **Milima ya Udzungwa** in Kiswahili

Check (☑) the most relevant results for this query and click **Submit** at the end of this page.

□ Milima ya Udzungwa - Wikipedia, kamusi elezo huru
https://sw.wikipedia.org/wiki/Milima_ya_Udzungwa

Milima ya Udzungwa ni jina la hifadhi inayopatikana katika bara la Afrika, nchini Tanzania, mkoani Iringa karibu na mpaka wa mkoa wa Morogoro, wilaya **ya** Kilombero. Hifadhi iko umbali wa kilometa 350 kusini mwa Dar es Salaam, kilometa 65 kutoka katika Hifadhi **ya** Mikumi.

□ Udzungwa Mountains - Wikipedia
https://en.wikipedia.org/wiki/Udzungwa_Mountains

The **Udzungwa Mountains** are a **mountain** range in south-central Tanzania.The **mountains** are mostly within Iringa Region, south of Tanzania's capital Dodoma.The **Udzungwa Mountains** are part of the Eastern Arc **Mountains**, and are home to a biodiverse community of flora and fauna with large numbers of endemic species.. The **mountains** are home to the Hehe people, and the name **Udzungwa** comes from the ...

□ MILIMA UDZUNGWA FAHARI YA TANZANIA KATIKA MAPOROMOKO YA SANJE

Fig. 2. Search results layout.

make decisions as to whether a result is relevant or not. It is common to derive or infer relevance judgement using snippets in click logs data. Most studies in click models [Chuklin et al. 2015; Grotov et al. 2015] and learning to rank [Joachims et al. 2017a,b; Liu 2011] rely on relevance judgements via clicks as a result of snippets examination.

On the left of each result, there was a checkbox to let users indicate/check (✓) the most relevant results. This saved a great deal of users' time from navigating through all the relevant results and at the same time allowing users to complete as many search sessions as possible. When a user had indicated the most relevant result(s), he/she had to click on the submit button at the end of the results page. The chosen results (URLs) and search query were stored in a text file on a server. This data represented the click-through logs, used in the analysis of this paper. In addition to this data, all responses from Bing Web Search API call were dumped to another (JSON) file on the server.

The last page of the platform had an acknowledgement message to thank users for participating in the research. The page also had a request message to users who had time and wish to repeat the search process, with different randomly generated topics. With these simple procedures, participants were able to follow flawlessly and some participants were able to perform several iterations of search on their own and complete the exercise in less than 10 minutes.

3.3 Participant Recruitment

Participant recruitment targeted the general public of Web users in Tanzania, but it was supplemented by university students, specifically from the Sokoine University of Agriculture[11], Tanzania. The recruitment of the general public was via: i) social media such as WhatsApp messenger groups and individuals in a snowball style, LinkedIn[12], Instagram[13] and JamiiForums[14]; and ii) mailing lists of organizations and universities in Tanzania, with the help of friends and colleagues working there.

[11] https://sua.ac.tz/
[12] https://www.linkedin.com/
[13] https://www.instagram.com/
[14] https://www.jamiiforums.com/ – The (local) Tanzania's leading social networking forum/website

After obtaining research clearance from the Sokoine University of Agriculture, student participants were recruited with the help of lecturers and class representatives (CRs). Lecturers communicated the invitation message to the CRs using WhatsApp messenger. The CRs used WhatsApp class groups, a common means of communication among students to deliver the message. Student participants interacted with our data collection platform using university computer laboratories while non-student participants interacted with our data collection platform using their own gadgets such as smartphones and computers. The invitation message, written in both English and Kiswahili, had a link (URL) to the data collection platform. The system only granted permission to the study after signing the informed consent form at the homepage of the platform.

3.4 Data set Description

The data collection process spanned three month from 06 November 2019 to 5 February 2020. As explained in the data collection platform in Section 3.2 above, both queries and click-through data were logged. In addition, user ratings on the use of English and Kiswahili on the Web search as well as demographics information were collected.

The first (query) data set contained 2387 query records from the user interactions with the 123 topics mentioned in Section 3.1 above. Each query record had the query and topic it belonged to. The query records per topic of search (also called counts or frequency) for each language were aggregated. The distribution of query records per topic of search in Kiswahili was as follows: minimum – 0; maximum – 28; and average – 9 queries per topic. For English, the distribution of queries per topic of search was: minimum – 0; maximum – 23; and average – 7 queries per topic.

For brevity and demonstration reasons, related topics were aggregated into large groups of topics, called *super-topics*. For example, topics such as *Computer*, *Hardware*, *Internet*, *Phones*, *Software*, *Telecommunications* and *Television* were all grouped into the *IT and Electronics* super-topic. The grouping resulted in 19 super-topics for the original 123 topics, as shown in the Supplemental Materials Table A1.

The second (click-through) data set contains 3157 click-through records (or clicked URLs). Every click-through record has a minimum of 0 to a maximum of 10 relevant clicked results (URLs). The records that had neither Kiswahili nor English clicked URLs were deleted i.e. records where users did not find relevant results from both languages and did not click on any of the results. Each record has three columns: language identifier, query, and a list of clicked URLs. Each query was associated with its corresponding topic by looking up the original topic and query corpus. Finally, all the queries that belonged to one topic for both English and Kiswahili were put together, and further grouped by super-topics as in the query (first) data set case.

3.5 Data Analysis

The analysis of demographic information as well as the user ratings on the use of Kiswahili and English in searching for information on the Web, was mainly via descriptive statistics. While MS Excel was used to perform the descriptive statistical analyses and the exploratory analysis, mainly hypothesis testing (Mann-Whitney test) was done using an online test tool[15]. The analysis of the query and the click-through logs from the user interaction with the guided multilingual system is explained in the following subsections.

3.5.1 Estimating Query Language Preferences.

Recall that the guided multilingual system was designed in such a way that a user is presented with a list of five topics and a drop-down menu for choosing the language for viewing the queries. The language chosen is treated as a preferred query language over the other language. Since there

[15] https://www.statskingdom.com/170median_mann_whitney.html

were only two language options, the user was forced to choose from the two, commonly called a forced-choice paired preference test [Lawless and Heymann 2013; Meilgaard et al. 2006]. Meilgaard et al. [2006] define a preference test as one in which a respondent is forced to choose one item over another or others. Using this test to answer research question 2 (**RQ2**), the objectives were:

(1) to determine the overall preferred query language;
(2) to determine the preferred query language in super-topics; and
(3) to determine the preferred query language in topics.

The tests were one-tailed tests i.e. one language is preferred than the other. In order not to conclude falsely that a preference exists, it was important to adjust differently the sensitivity values α, β and P_{max} to address each of the above objectives. Note that α (or α-risk) is the probability of concluding that there is a preference, while actually there is not (Type I error) and β (or β-risk) is the probability of concluding that there is no preference, while in fact there is (Type II error). Meilgaard et al. [2006] define P_{max} as *"the departure from equal intensity (i.e., a 50:50 split of opinion among respondents) that represents a meaningful difference to the researcher"*. For example, for a 90% confidence level of detecting a 70:30 split in preferences, then $P_{max} = 70\%$ and $\beta = 0.10$. The rule of thumb states that: if $P_{max} < 55\%$, $55\% \leq P_{max} \leq 65\%$ and $P_{max} > 65\%$, then there is a small, medium and large departure from equal intensity respectively.

The values of α can be calculated depending on the obtained number of responses n and the minimum number of common responses x using the formula:

$$\alpha = 1 - BINOMDIST(x - 1, n, P_0, 1) \quad (1)$$

The values of β are calculated depending on the minimum number of common responses x and the P_{max}, using the formula:

$$\beta = BINOMDIST(x - 1, n, P_{max}, 1) \quad (2)$$

The P_{max} is based on the probability of common guess (P_0) and the proportion of distinguishers (P_d) such that:

$$P_{max} = P_d + P_0(1 - P_d) \quad (3)$$

Setting P_0 and P_d at 0.5 each yields a $P_{max} = 75\%$. This value is large enough to ensure a clear split in user preferences between English and Kiswahili. Particularly, some topics had higher values of β than the maximum desired (i.e. 0.20), thus were omitted to avoid large Type II errors. This was partly caused by a small number of respondents/responses in those topics. Thus, only 47 out of 123 topics qualified for analysis.

The minimum number of common responses x can be calculated as follows:

$$x = (n/2) + z\sqrt{n/4} \quad (4)$$

where n is the total number of responses in a super-topic or topic and $z = 1.645$ for a one-tailed test with $n \leq 30$. Different values of z were obtained from Table 17.3 in [Meilgaard et al. 2006] for $n > 30$. If the observed number of common responses c is greater or equal to x, then the conclusion is that there is a preference for a language with common responses and no preference otherwise. The formulas in Equations 1 – 4 are adopted from [Meilgaard et al. 2006].

3.5.2 *Estimating Preferences for Language of Results.*

Treating each URL as an independent choice/response from a user, the previously discussed method for estimating query language preferences is adopted to estimate the preferred language of results. Then, we attempted to answer the third research question (**RQ3**) using the following objectives.

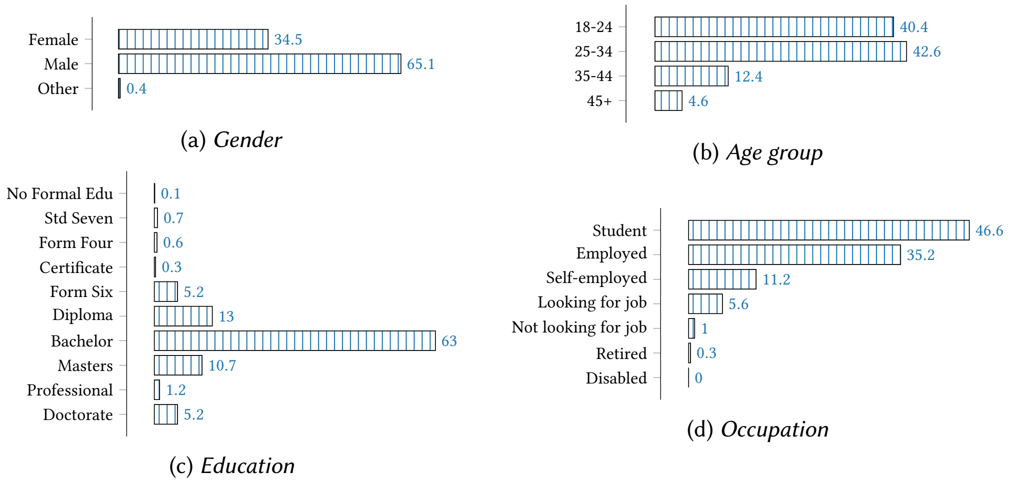
(1) To determine the generic preferred language of results.
(2) To determine the preferred language of results in super-topics.
(3) To determine the preferred language of results in topics.

Topics with higher values of β than the desired i.e. $\beta > 0.20$ were omitted in the analysis to avoid committing large Type II errors. 66 out of 123 topics qualified for analysis. Note that about 1% error was allowed to accommodate some of the topics whose β-values were closer to the desired value.

4 RESULTS

4.1 Demographic Information

The experiment involved 676 participants: 65.1% male, 34.5% female, and 0.4% who preferred not to disclose their gender (Figure 3a). The largest proportion of participants were young and middle aged, between 18-24 (40.4%), 25-34 (42.6%) and 35-44 (12.4%). Only 4.6% of participants were aged above 45 years, as shown in Figure 3b.



(a) *Gender*

(b) *Age group*

(c) *Education*

(d) *Occupation*

Fig. 3. Participants' demographics (**N=676**).

In Figure 3c, Standard (Std) Seven represents the highest level of primary school education, while Form Four and Form Six represents ordinary and advanced level of secondary (high) school education respectively.

Most of the participants had (or were at) tertiary level of education: 63.0% with bachelors degree, 13.0% with diploma, 10.7% with masters degree, 5.2% with doctorate degree, and 1.2% with professional degree e.g. Veterinary Doctor. 0.3% had a Certificate. 0.1% of participants had no formal education, 0.7% had primary school education and 5.8% had high school education (0.6% form four and 5.2% form six), as can be seen in Figure 3c.

In terms of occupation (Figure 3d), 46.6% of participants were students, 35.2% were employed, and 11.2% were self-employed or entrepreneurs or businesspersons. Meanwhile, 5.6% of participants were not employed but seeking jobs, and 1.0% were not employed and not looking for jobs. 0.3% of participants were retired and there was no participant who was disabled to the extent of not able to work.

4.2 The Use of English and Kiswahili on Web Search

The participants' self-assessment of the use of English and Kiswahili in Web search indicate that the largest number of our participants rated themselves to "always" (41.9%), "often" (28.3%) and "sometimes" (25.1%) use English in their daily Web search. The remaining participants "rarely" (2.7%) or "never" (2.1%) use English to search for information on the Web (Figure 4). On the other hand, 39.9% of participants rated themselves as they "sometimes" do use Kiswahili in their Web search. A substantial number of participants “always” and “often” use Kiswahili (24.7% and 8.9% respectively). 19.8% “rarely” and 6.7% “never” use Kiswahili in their Web search.

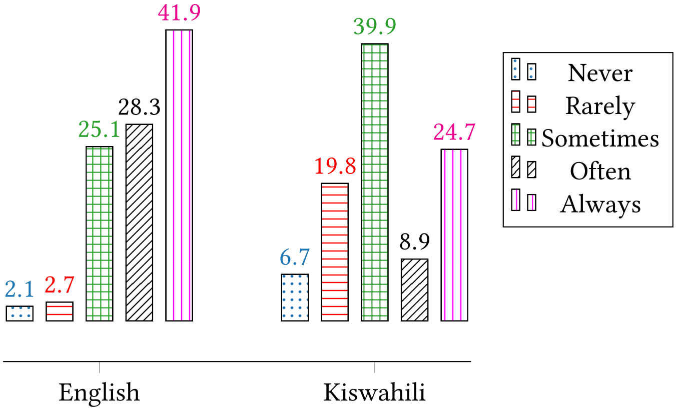


Fig. 4. Participants ratings on their language use on the Web search **(N=676)**.

The descriptive statistics in Table 2 indicate that participants’ rating of Kiswahili and English use in Web search had a mode of 3 and 5 respectively. These modes imply that participants “sometimes" use Kiswahili while “always" using English to search for information on the Web.

Table 2. Ratings on language use on the Web search **(N=676)**.

	Median	Mode	Min.	Max.
English	4	5	1	5
Kiswahili	3	3	1	5

Using a Mann-Whitney U test to compare the ratings of English and Kiswahili using $\alpha = 0.05$, the null hypothesis H_0 is rejected. This means, the difference in the medians of English and Swahili use is big enough to be statistically significant (U=142098.5, p=0.0000). The results suggest that English is significantly preferred to Kiswahili as a language for Web search.

4.3 User Interaction with the Topic and Queries System

The interaction of users with the topic and queries system is summarized in Figure 5. While some searched only in a single topic, presumably once, others searched multiple times in different topics, either sticking to one language across all topics they searched from or switching between English and Kiswahili as topics changed. 23.6% and 17.7% of users searched in a single topic using solely Kiswahili and English respectively. Users who searched in multiple topics were divided into two types: i) those using one language irrespective of topic, 18.8% and 12.9% for Kiswahili and English respectively; and ii) those using both languages across topics. The latter can be divided into three groups: i) those using both languages equally (5.4%) e.g. 2 topics in English and 2 topics in Kiswahili; ii) those using more topics in English than Kiswahili (11.4%) e.g. 5 topics in English and 2 topics in Kiswahili; and iii) those using more topics in Kiswahili than English (10.1%) e.g. 7 topics in Kiswahili and 3 topics in English.

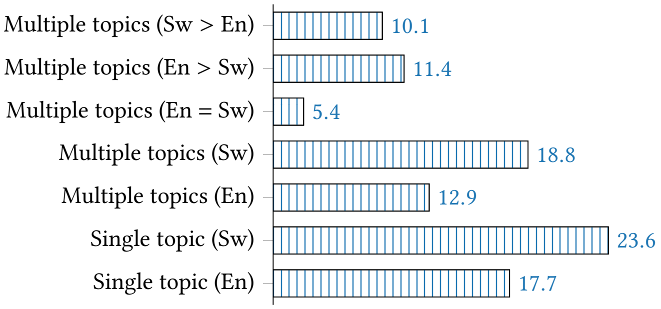


Fig. 5. User behaviour when interacting with the topics and queries.

The bar chart in Figure 6 represents the aggregated counts of responses per super-topic. The responses represent the number of of queries in a particular language. The figure shows that bars in each super-topic are not equal in length, indicating that there were variations in which English and Kiswahili were used in different super-topics. In most super-topics, it can be seen that Swahili bars are taller than English bars. Out of these visual observations, three sub-questions may be formulated, which are addressed in the next sub-sections.

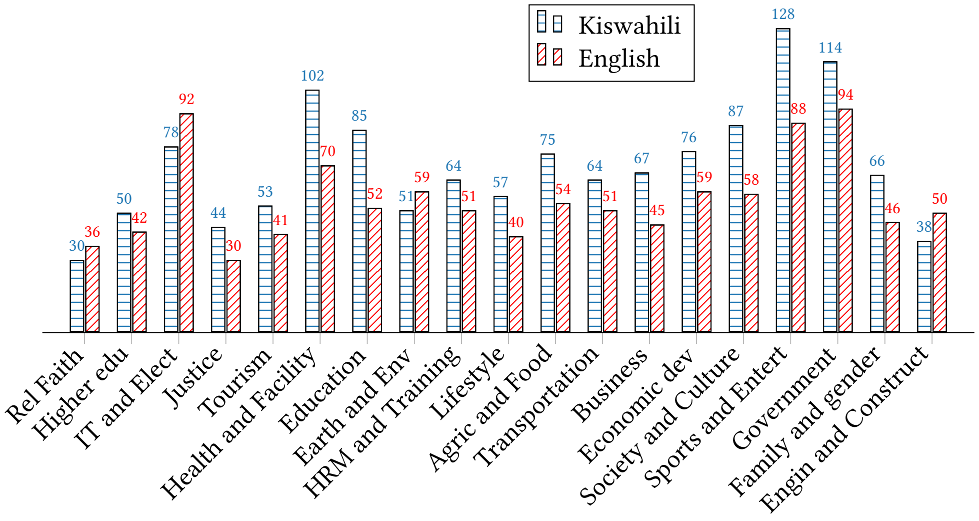


Fig. 6. Frequency/counts of query language vs super-topic of search.

4.3.1 *What is the generic preferred query language?*

There were a total of 2387 responses from all the topics, in which 1329 and 1058 responses were in favour of Kiswahili and English respectively. Using Equation 4, a minimum of 1250 common responses were needed in order to conclude that there is a preference for one of the languages. Since Kiswahili had 1329 responses compared to 1058 of English, it can be concluded that there was a significant preference for Kiswahili as a generic query language, at the calculated (using Equation 1 − 3) sensitivity values of $\alpha = 0.01$, $\beta = 0.00$ and $P_{max} = 75\%$.

4.3.2 *What is the preferred query language in super-topics?*

The Supplemental Materials Table A2 details the tests for query language preferences in all the 19 super-topics, tested at different sensitivity values using Equation 1 – 3 such that: $0.04 \le \alpha \le 0.08$; $0.0000 \le \beta \le 0.0033$; and $P_{max} = 0.75$. There was a statistically significant preference for Kiswahili as a query language in 9 out of the 19 super-topics (47%, Figure 9a). The super-topics are: Justice; Health and Facility; Education; Lifestyle; Agriculture and Food; Business; Society and Culture; Sports and Entertainment; and Family and Gender. There was no preference (ties) for language in the remaining 10 super-topics (53%, Figure 9a). These topics are: Religious faith; High education; IT and Electronics; Tourism; Earth and Environment; HRM and Training; Transportation; Economic development; Governance; and Engineering and Construction.

Contrary to the visual observation in the bar chart (Figure 6), where it was observed that English bars were marginally taller than Swahili bars in 4 super-topics, it can be seen that English is not a significantly preferred query language in any of the super-topics. Swahili bars that are taller than the English counterparts in 15 super-topics indicate that Kiswahili is significantly preferred as a query language in only 9 super-topics. Visual differences in the sizes of the bars in 10 super-topics are not statistically significant to conclude preferences for any language and therefore users equally used Kiswahili and English as a query language.

4.3.3 *What is the preferred query language in topics?*

Refer to Supplemental Materials Table A3 for the details of the tests for 47 topics, which passed the requirement of a minimum number of responses and the sensitivity values. These values were calculated using Equation 1 – 3 such that: $0.03 \le \alpha \le 0.09$; $0.02 \le \beta \le 0.21$; and $P_{max} = 0.75$. It is observed that there was a significant preference for Kiswahili as a query language in 16 topics (34%, Figure 9b). Some of the topics include: Law; National park; HIV/AIDS; Waste management; and Agriculture. On the other hand, English was significantly preferred as a query language in only 4 topics (9%, Figure 9b). Such topics include: Religion; Computer; Heart; and Water. There was significantly no preference for language in 27 topics (57%, Figure 9b), including: Phones; Environment; Training; and Fashion.

4.4 User Interaction with the Results Page

A total of 20 results, per search query, were displayed in the search engine results page (SERP), with each language having equal share of the number of results i.e. 10 each. Every query had a minimum of 0 to a maximum of 10 relevant clicked results (URLs) in both English and Kiswahili. A descriptive statistical summary in Table 3, from 809 sessions, indicates that the mean number of URLs clicked per query is 2.14 and 1.77 for English and Kiswahili respectively. Both English and Kiswahili had modes and medians of 1 URL per query respectively. The modes of 1 imply that most queries had at least 1 clicked relevant answer for both English and Kiswahili.

Table 3. Query-URLs descriptive statistics (**N=809**).

	Mean	Median	Mode	Std Dev.	Variance	Min.	Max.
English	2.14	1	1	2.08	4.33	0	9
Kiswahili	1.77	1	1	2.07	4.29	0	9

User interaction with the SERP can be categorised as follows: i) users who clicked on results in only one language – English (35.6%) and Kiswahili (24.1%); and ii) users who clicked on results in both languages. In this group, some users clicked on equal number of results from both languages

(10%), some clicked on more results in English (15.9%) and others clicked on more results in Kiswahili (14.3%) (Figure 7).

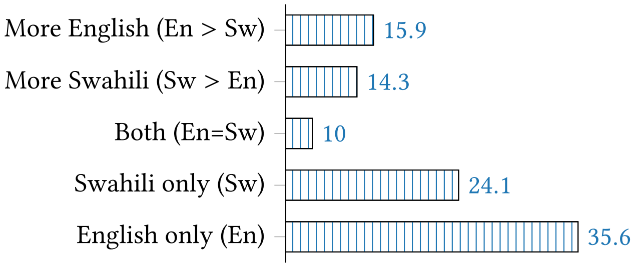


Fig. 7. User behaviour in interacting with the topics and queries.

As described in Sub-section 3.4, every query was mapped to its respective topic and related topics were grouped into 19 super-topics. The frequencies of clicked URLs/results per super-topic are plotted in Figure 8. Visually, the bars are varying in length within super-topics, which consequently implies the same in topics, leading to three sub-questions addressed in the next subsections.

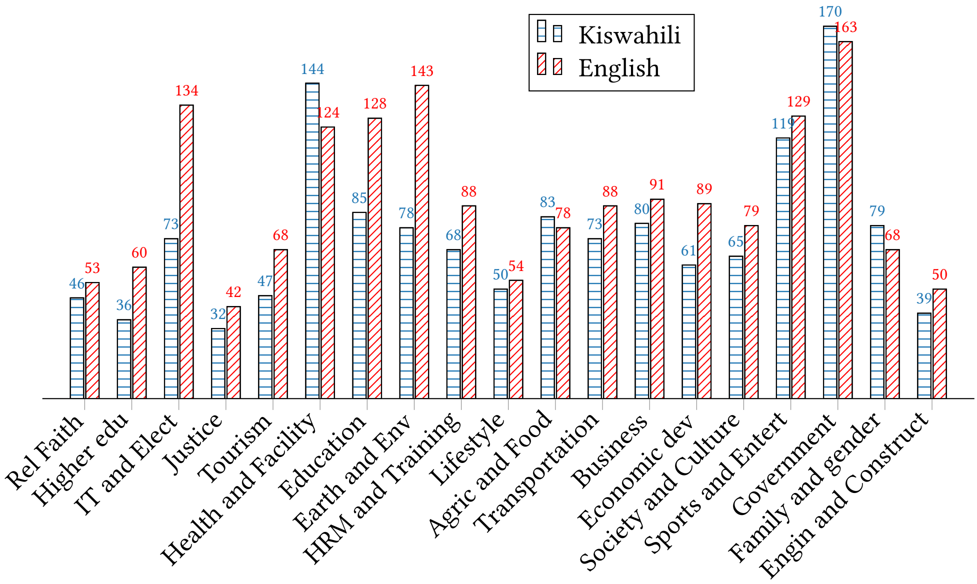


Fig. 8. Total URLs/results clicked in each language vs super-topic of search results.

4.4.1 *What is the overall preferred language of results?*

A total of 3157 URLs/results were clicked, out of which 1729 were in English and 1428 were in Kiswahili. Using Equation 4, a minimum of 1644 common clicked results (responses) were needed in order to conclude that there is a preference of one language over the other. Since English had 1729 clicked results compared to 1428 in Kiswahili, then, there is a generic significant preference for English as a language of results, given the calculated (using Equation 1 – 3) sensitivity values of $\alpha = 0.01$, $\beta = 0.00$ and $P_{max} = 75\%$.

4.4.2 *What is the preferred language of results in super-topics?*

Refer to Supplemental Materials Table A4 for the details of the tests for preference in each super-topic at different sensitivity values calculated using Equation 1 – 3 such that: $0.04 \leq \alpha \leq 0.06$; $0.0000 \leq \beta \leq 0.0024$; and $P_{max} = 0.75$. The results indicate that there was significantly no preference for language in 12 (63%, Figure 10a) super-topics. English was significantly preferred as a language of results in the remaining 7 (37%, Figure 10a) super-topics. Kiswahili was not significantly preferred in any of the super-topics, contrary to the visual observation in Figure 8, where Kiswahili was marginally preferred to English as a query language in 4 super-topics.

4.4.3 *What is the preferred language of results in topics?*

Table A5 in the Supplemental Materials details the tests for preference for the language of results in 66 topics that passed the criteria for sensitivity values. The values were calculated using Equation 1 – 3 such that: $0.03 \leq \alpha \leq 0.09$; $0.00 \leq \beta \leq 0.21$; and $P_{max} = 0.75$. The results show that there was significantly no preference for language of results in 46 (70%, Figure 10b) topics. Such topics include: Phones; Court; National park; Business; and Constitution. There was a significant preference for English as a language of results in 15 (23%, Figure 10b) topics, such as: University admission; Scholarship; Law; Energy; and Chemistry. The results also show that Kiswahili was significantly preferred as a language of results in 5 (7%, Figure 10b) topics, which are: HIV/AIDS; Livestock; News; Music; and Election.

5 DISCUSSION

Developing effective MLIR systems requires understanding the user behaviour and preferences. However, due to the complexity of user behaviour and preferences, survey-based studies may not be able to effectively capture them. Instead, query and click-through logs have become an invaluable source of implicit information on user behaviour and preferences. As a result, this controlled study explored user preferences, particularly language preferences and how they are associated with topic of search and results. A small part of this study used a questionnaire to understand the way multilingual Web users evaluate themselves on how they use English and Kiswahili in searching for information on the Web. The major part of this study allowed participants to interact with a guided multilingual search engine, in which the query and click-through logs are analysed.

5.1 RQ1: How do polyglot Swahili speaking Web users rate themselves on the use of English and Kiswahili on Web search?

It was observed that users had mixed feelings about their use of English and Kiswahili in their Web search. Most users indicated that they "always" and "sometimes" use English and Kiswahili respectively. There may be several reasons as to why English was rated as always being used in Web search. There was no follow up question to explore the reasons, as that was out of scope of this study. However, the massive amount of information available in English compared to a low-resourced language (Kiswahili) on the Internet may be part of the reason. For example, as of May 25, 2020, Wikipedia had 6,085,840 English articles compared to only 59,033 Kiswahili articles[16]. On the other hand, Tanzania uses English as a medium of instruction from Secondary schools to University level, including all the vocational colleges. This implies that most Swahili speaking Web users are trained in English and are likely to look for professional information in English and non-professional in Kiswahili as revealed in [Telemala and Suleman 2018].

[16] https://meta.wikimedia.org/wiki/List_of_Wikipedias

5.2 RQ2: What is the preferred query language among the polyglot Swahili speaking MLIR system users?

Before answering the questions about query language preferences amongst users, it was necessary to look at the interaction behaviour of the multilingual search system users, using query logs. It was observed that, apart from users who searched only once in one topic, users who searched in multiple topics were divided into two groups: i) those searching in multiple topics using a single language, where it was found that more participants used Kiswahili than English as a query language; and ii) those searching in multiple topics using both English and Kiswahili, either with varying numbers of topics in the two languages – e.g 3 topics in English and 5 topics in Kiswahili – or with equal numbers of topics in both languages.

The results for language preferences suggest that the smaller the topic granularity, the clearer the language preference. In other words, though statistically significant, it may be misleading to generalize that Kiswahili is an overall preferred query language (refer to Section 4.3.1), while deep down in the super-topics and topics, no preference (ties) or English had a fair share of preference (refer to Sections 4.3.2 and 4.3.3). It was observed that: there was significantly no preference for language in 53% of the super-topics – both English and Kiswahili had significantly equal chance of being used as query language; Kiswahili was found to be a significantly preferred query language in the remaining 47% of the super-topics; and English was not significantly preferred. At the topic level, there was significant no preference in 57% of the topics, and preference for Kiswahili and English in 34% and 9% of the topics respectively.

The results largely suggest, from the user interaction with the topic and query system (query logs), that users had either no preference for language or preferred Kiswahili as a query language (Figure 9). Kiswahili was a significantly preferred query language almost equally to the proportion of "no preference" in super-topics and in about one third of the topics. English, on the other hand, was significantly preferred at a very small scale, particularly in topics. The term "language use in the Web search" as used in the questionnaire part of the study was broad in the sense that it may mean the language used to query the Web (query language) or the language of results from search engines or even the language used to surf the Web. Thus, to some degree, these findings may correlate with the "sometimes" use of Kiswahili found in user opinion in **RQ1** (Section 5.1), as the query language preference is biased towards either Kiswahili or no preference as opposed to English.

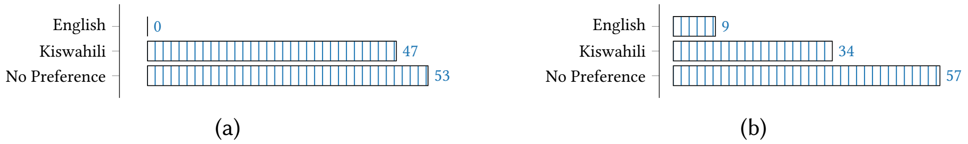


Fig. 9. Query language preferences in: (a) 19 super-topics; and (b) 47 topics

5.3 RQ3: What is the preferred language of results among the polyglot Swahili speaking MLIR system users?

At a higher generic level, English was found to be a significantly preferred language of results, but as the level of granularity decreased to super-topics and topics, it was observed no preference scored the largest margin and at the same time Kiswahili was not significantly preferred in any of the super-topics. There was significantly no preference for any language in 63% of the super-topics and the remaining 37% of the super-topics were significantly preferred in English as a language of results. On the other hand, there was significantly no preference for language of results in 70% of the topics and 23% and 7% of the topics were significantly preferred in English and Kiswahili respectively.

Contrary to the behaviour of the same users during querying, Figure 10 indicates that only a handful of topics were significantly preferred in Kiswahili. English results seem to be significantly preferred in more than one third of the super-topics and about a quarter of the topics. The largest proportion of preferences were "no preference" (ties), where both Kiswahili and English had equal chance of being used as language of results.

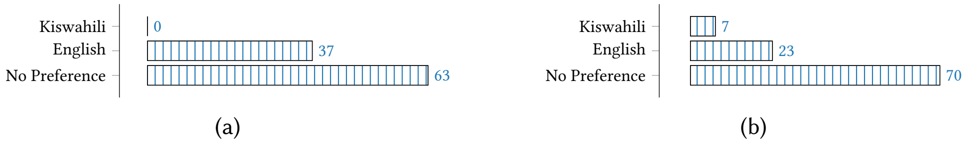


Fig. 10. Preferences of language of results for: (a) 19 super-topics; and (b) 66 topics.

The results for research questions **RQ2** and **RQ3**, discussed in Sections 5.2 and 5.3, can lead us to important inferences that form the contribution of this work towards building better multilingual information retrieval that caters for every group of users and scenarios. The inferences are explained in the next subsection.

5.4 RQ4: Do the topic-language preferences change at different stages (query to results selection) of MLIR user interaction?

Referring to Figures 9 and 10, it can be noted that there were changes in language preferences in the course of searching at the stage of querying the system (query language preference) and the results selection (language of results preference). For demonstration purposes, Table 4 shows the percentage of super-topics and topics whose language preferences changed. Note that only 36 topics, which had a perfect match or alignment, are presented in this paper, i.e., topics that existed in both the analysis of query language preferences (Section 4.3.3) and language of results preferences (Section 4.4.3).

In the super-topics, it can be seen that preferred language changed from Kiswahili to English in only the Education super-topic (5%). The preferred language changed from Kiswahili to no preference in 8 (42%) super-topics, which are: Justice; Health and Facility; Lifestyle; Agriculture and Food; Business; Society and Culture; Sports and Entertainment; and Family and Gender. Further, the preferred language changed from no preference to English in 6 (32%) super-topics, which are: Higher education; IT and Electronics; Tourism; Earth and Environment; HRM and Training; and Economic development. No preference did not change in 4 (21%) super-topics, which are: Religious faith; Transportation; Government; and Engineering and Construction.

In the topics, it is observed that there was a change in language preference from Kiswahili to English for only the Law topic (3%). The preferred language changed from Kiswahili to no preference in 9 (25%) topics. These topics are: University; National park; Clinic; Education; Waste management; Agriculture; Development; Industry; and Award. The change in preference for language from English to no preference occurred for one (3%) topic: Religion. The same number, 3%, happened from no preference to Kiswahili for the Election topic. The preference change from no preference to English occurred in 6 (17%) topics, which are: University admission; Computer hardware; Environment; Training; Food; and Social. There was no change in language preference of Kiswahili for HIV/AIDS and Music topics (6%). No change in preference for English was observed in 3 (8%) topics, which are: Computer; Heart; and Water. No preference did not change in 13 (36%) topics, which are:

Table 4. Changes in language preferences in both super-topics and topic from query to results languages.

	Percent Changing	
Type of Change	**Super-topics (N=19)**	**Topics (N=36)**
sw → en	5	3
sw → np	42	25
sw → sw	0	6
en → sw	0	0
en → np	0	3
en → en	0	8
np → sw	0	3
np → en	32	17
np → np	21	36

Where sw = Kiswahili, en = English and np = no preference.

Christianity; Software; Phones; Mathematics; School; Human resource; Fashion; Movies series; Culture; Public service; Government; Family; and Child.

The justification for these changes (or no changes) in language preferences from query language to the language of results are not in the of scope of this work. However, some may be associated with (un)availability of online documents [Aula and Kellar 2009; Telemala and Suleman 2018; Wang and Komlodi 2018] in Kiswahili or due to the fact that users prefer to search in languages they are familiar with (in this case Kiswahili) [Lowe and Steichen 2017; Wang and Komlodi 2018]. It is interesting to see that despite the scarcity of online Swahili documents, users were able to get the information they wanted to satisfy their information needs. This is demonstrated by the fact that the proportion of "no preference" where results in Kiswahili and English could potentially be equally used, was large (more than two third of the topics).

Findings for which users significantly preferred one language over the other in some topics (or super-topics) may imply that a topic can be associated to a language used – *Topic-Language association/preferences.* For topics such as HIV/AIDS and Music, it was shown that, users significantly preferred querying and consuming (selecting) results in Kiswahili. On the other hand, topics such as Computer and Heart were significantly preferred in English as both a query language and language of results. Even when users changed their preference from say Kiswahili to English for topics such as Law, it may be said to be circumstantial i.e. availability (quantity) of documents [Aula and Kellar 2009; Domingues and Lopes 2019; Telemala and Suleman 2018; Wang and Komlodi 2018] and quality of documents [Aula and Kellar 2009; Domingues and Lopes 2019; Lopes and Ribeiro 2013].

There may be several other reasons such as prior knowledge about whether sufficient number of documents in a certain language, time constraints, and whether one wants high recall or precision, etc. Nevertheless, the investigation about these factors and others influencing language preferences were not in the scope of this paper. Our intention was only to demonstrate, using query and click-through logs, that topic of search may be associated with (or varies with) language of either query or results – *Topic-Language Preferences.*

5.5 Limitations of the Study

One limitation of this study, which may affect making a proper generalization to the overall Swahili speaking Web users community in Tanzania, is the type of participants. As it was presented in Figure 3 that the largest proportion of the participants were between the age of 18 to 34, mostly students in their bachelor's degree and some employed individuals. For practical application, it would be beneficial to have a balanced representation of the Swahili speaking Web users community in terms of education, age and occupation.

Another limitation is the number of participants. Most of research that involves query logs, for example in the learning to rank studies, involve a very large number of users and/or dataset e.g. the Yahoo! Weboscope dataset [Chapelle and Chang 2011]. Unlike such studies, which mostly use commercial search engines query logs and essentially restricted to monolingual IR datasets, the experimental MLIR search engine used in this study did not have access to such a large audience or a dataset with MLIR query logs. There are also studies that involve click-through log data but use limited numbers of users, for example, Joachims et al. [2017a].

Furthermore, only the results snippets were used for (perceived) relevance judgement. Liu [2011] argue that snippets must be treated with care when inferring relevance despite that there is a strong correlation between perceived relevance and absolute relevance. There is a large body of works on user modeling (click models) that rely on snippets to make relevance judgements e.g. [Craswell et al. 2008; Dupret and Piwowarski 2008; Guo et al. 2009b,a]. Other user modeling approaches e.g. Dynamic Bayesian Model (DBN) [Chapelle and Zhang 2009] believe that snippets are not enough to infer relevance of a result, suggesting users need to visit a given page.

One would argue that it is important to let users of the system develop their own queries as it helps to reveal the actual information needs and relevance judgements from real users. That is, with the current experimental design, users were forced to make relevance judgements on queries that they did not create. Even so, the reasons for opting for this carefully controlled experimental design are two-fold: i) avoiding data skewness problem; and ii) avoiding machine translation problem.

Using a small group of users, it was possible to get feedback on a wide range of topics as opposed to asking the very same small group of users to search for whatever topics they wanted. The dataset obtained would not have the same (normal) distribution and thus, skewed. One possible way to avoid skewed data is getting data from a very large number of users. However, a very large number of users for log-based studies is only possible when using commercial search engines. Thus, by giving extra guidance for the users, with a small number of users, it is possible to get a dataset that can be assumed to be under the normal distribution.

(Machine) translation (MT) is paramount in achieving MLIR [Peters et al. 2012]. The challenge with letting users develop their own queries is the overhead in producing/obtaining perfect translation, especially for the under-resourced language like Kiswahili. MT for low resource languages is far from being perfect, partly due to inadequate parallel corpora [Karakanta et al. 2018]. Letting users develop queries on-the-fly would, therefore, reduce robustness of the system due to translation errors; thus, affecting retrieval results in one of the languages, and consequently imbalance the language preferences investigated in this study.

6 CONCLUSIONS AND FUTURE WORK

Inspired by the code-switching behaviour among polyglot Web users when interacting with an information retrieval system, specifically the search engine, this study explored the topic-language preferences in MLIR. The objective is to aid the development of novel MLIR solutions based on user behavior and preferences. To the best of our knowledge, this paper presented the first study that utilized multilingual query and click-through logs to explore the topic-language preferences and show that language preferences may change in the course of a MLIR search. More specifically, the work forms part of the efforts to support "fair" multilingual information retrieval in low-resourced languages such as Kiswahili, yet spoken by a large number of users. The study used the case of polyglot Swahili speaking Web users in Tanzania.

The study involved a small part based on a questionnaire and the main part was based on a controlled multilingual search engine, from which we exploited the query and click-through logs to derive the topic-language preferences. Results from the questionnaire study indicated that, on average, users often use English and sometimes use Kiswahili in their daily Web search. Generally, from the query logs, it can be concluded that either Kiswahili is a significantly preferred query language or there were no preference for language at all. Whereas, from the click-through logs, it can be concluded that either English was a significantly preferred/used language of results or there were no language preference at all.

Focusing only at a topic level, it is evident that both query language and language of results can be associated with the topic of search – *Topic-Language association*. That is, the preferences for query and results languages differ from one topic to another. It is important to note that, language preferences were not observed in all topics; for example, there were language preference in only 30% of the topics (7% for Kiswahili and 23% for English), while there were no preferences for language in the remaining 70%. It is further observed that users can change or stick with their preferred language in the course of MLIR Web search i.e. topic-language associations may change depending on the stage of search (querying stage and results selection stage).

The results of this study open up avenues for MLIR research towards developing better systems by shedding light on topic-language associations and preferences and can be used as a basis for building a better information retrieval system to support users in certain scenarios. The scenarios may include: i) populating the SERP (by re-ranking) with more (or top) results in the preferred language as opposed to equally interleaving results from both languages; ii) support at the results level for users who indicated to prefer querying in Kiswahili, despite the low number of documents on the Web.

To this end, our plan is to build statistical models that learn the query language preferences as well as the language of results preferences and devise algorithms for re-ranking the results according to query language preferences. It is important to note that this study was mostly dominated by student participants, and it might not well generalize to other groups. A further study using more diverse groups of users and/or more data may give more insights on the topic-language preferences and language preference changes. For more perspectives into the user topic-language preferences, another study might also collect/decode a lot more information (than just clicks) such as, asking users for graded relevance judgements, and reasons for their choices/judgements.

ACKNOWLEDGMENTS

This work was financially supported by the Hasso-Plattner Institute for Digital Engineering.

REFERENCES

Hany M Alsalmi. 2019. Information-seeking in multilingual digital libraries: Comparative case studies of five university students. *Library Hi Tech* ahead-of-print, ahead-of-print (2019), 23–32.

Anne Aula and Melanie Kellar. 2009. Multilingual search strategies. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*. ACM, Boston, MA, USA, 3865 – 3870.

Bettina Berendt and Anett Kralisch. 2009. A user-centric approach to identifying best deployment strategies for language tools: the impact of content and access language on Web user behaviour and attitudes. *Information Retrieval* 12, 3 (2009), 380.

Olivier Chapelle and Yi Chang. 2011. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*. JMLR: Workshop and Conference Proceedings, Sunnyvale, CA, 1–24.

Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain (WWW '09)*. Association for Computing Machinery, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/1526709.1526711

Peng Chu and Anita Komlodi. 2017. TranSearch: A Multilingual Search User Interface Accommodating User Interaction and Preference. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2466–2472. DOI: http://dx.doi.org/10.1145/3027063.3053262

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services* 7, 3 (2015), 1–115.

Paul Clough and Irene Eleta. 2010. Investigating Language Skills and Field of Knowledge on Multilingual Information Access in Digital Libraries. *International Journal of Digital Library Systems* 1, 1 (2010), 89–103.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto, California, USA (WSDM '08)*. Association for Computing Machinery, New York, NY, USA, 87–94. DOI: http://dx.doi.org/10.1145/1341531.1341545

Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and Comparing Portuguese and English Wikipedia Medicine-Related Articles. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1203–1207. DOI: http://dx.doi.org/10.1145/3308560.3316758

Georges E. Dupret and Benjamin Piwowarski. 2008. A User Browsing Model to Predict Search Engine Click Data from Past Observations.. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, Singapore (SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 331–338. DOI: http://dx.doi.org/10.1145/1390334.1390392

Christian Fluhr, Robert E Frederking, Doug Oard, Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. 1999. Multilingual (or Cross-lingual) Information Retrieval. *Proceedings of the Multilingual Information Management: Current Levels and Future Abilities* . (1999), 10–13.

Artem Grotov, Aleksandr Chuklin, Ilya Markov, Luka Stout, Finde Xumara, and Maarten de Rijke. 2015. A Comparative Study of Click Models for Web Search. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Josanne Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Capellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 78–90.

Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael Taylor, Yi-Min Wang, and Christos Faloutsos. 2009b. Click Chain Model in Web Search. In *Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain (WWW '09)*. Association for Computing Machinery, New York, NY, USA, 11–20. DOI: http://dx.doi.org/10.1145/1526709.1526712

Fan Guo, Chao Liu, and Yi Min Wang. 2009a. Efficient Multiple-Click Models in Web Search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining. Barcelona, Spain (WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 124–131. DOI: http://dx.doi.org/10.1145/1498759.1498818

Qi Guo, Ryen W. White, Yunqiao Zhang, Blake Anderson, and Susan T. Dumais. 2011. Why Searchers Switch: Understanding and Predicting Engine Switching Rationales. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China (SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 335–344. DOI: http://dx.doi.org/10.1145/2009916.2009964

Bernard J Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information processing & management* 42, 1 (2006), 248–263.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017a. Accurately Interpreting Clickthrough Data as Implicit Feedback. *SIGIR Forum* 51, 1 (Aug. 2017), 4–11. DOI: http://dx.doi.org/10.1145/3130332.3130334

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017b. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 781–789. DOI: http://dx.doi.org/10.1145/3018661.3018699

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation* 32, 1 (2018), 167–189.

Anett Kralisch and Thomas Mandl. 2006. Barriers to information access across languages on the internet: Network and language effects. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 3. IEEE, Kauia, HI, USA, 54b–54b.

Harry T Lawless and Hildegarde Heymann. 2013. *Sensory evaluation of food: principles and practices.* Springer Science & Business Media, Berlin, Germany.

Chenjun Ling, Ben Steichen, and Alexander G. Choulos. 2018a. A Comparative User Study of Interactive Multilingual Search Interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval - CHIIR '18*. ACM, New Brunswick, NJ, USA, 211–220.

Chenjun Ling, Ben Steichen, and Alexander G Choulos. 2018b. A Comparative User Study of Interactive Multilingual Search

Interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval.* ACM, New York, NY, 211–220.

Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual News–An Investigation of Consumption, Querying, and Search Result Selection Behaviors. *International Journal of Human–Computer Interaction* 36, 6 (2020), 516–535. DOI: http://dx.doi.org/10.1080/10447318.2019.1662636

Tie-Yan Liu. 2011. *Learning to rank for information retrieval.* Springer Science & Business Media, Berlin, Germany.

Carla Teixeira Lopes and Cristina Ribeiro. 2013. Measuring the value of health query translation: An analysis by user language proficiency. *Journal of the American Society for Information Science and Technology* 64, 5 (2013), 951–963.

Ryan Lowe and Ben Steichen. 2017. Multilingual Search User Behaviors – Exploring Multilingual Querying and Result Selection Through Crowdsourcing. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17.* ACM, Bratislava, Slovakia, 303–307.

Jennifer Marlow, Paul Clough, Juan Cigarrán Recuero, and Javier Artiles. 2008. Exploring the Effects of Language Skills on Multilingual Web Search. In *Advances in Information Retrieval*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 126–137.

Morten C Meilgaard, B Thomas Carr, and Gail Vance Civille. 2006. *Sensory evaluation techniques.* CRC press, Boca Raton, Florida.

Willi Mueller, Thiago H Silva, Jussara M Almeida, and Antonio AF Loureiro. 2017. Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science* 6, 1 (2017), 5.

Peggy Nzomo, Isola Ajiferuke, Liwen Vaughan, and Pamela McKenzie. 2016. Multilingual Information Retrieval & Use: Perceptions and Practices Amongst Bi/Multilingual Academic Users. *Journal of Academic Librarianship* 42, 5 (2016), 495–502.

Peggy Nzomo, Liwen Vaughan, Isola Ajiferuke, and Pam McKenzie. 2019. Multilingual Information Access (MLIA) Tools on Google and WorldCat: Bi/Multilingual University Students' Experience and Perceptions. *Journal of Library Administration* 59, 8 (2019), 831–853.

Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual Information Retrieval - From Research to Practice.* Springer Science & Business Media, Berlin/Heidelberg, Germany.

Razieh Rahimi, Azadeh Shakery, and Irwin King. 2015. Multilingual information retrieval in the language modeling framework. *Information Retrieval Journal* 18, 3 (2015), 246–281.

Mark Sanderson. 2008. Ambiguous queries: test collections need more sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, New York, NY, 499–506.

Li Si, Qiuyu Pan, and Xiaozhe Zhuang. 2017. An empirical analysis of user behaviour on multilingual information retrieval. *The Electronic Library* 35, 3 (2017), 410–426.

Ben Steichen and Luanne Freund. 2015. Supporting the Modern Polyglot. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15.* ACM, Seol, Korea, 3483–3492.

Ben Steichen, M Rami Ghorab, Alexander O'Connor, Séamus Lawless, and Vincent Wade. 2014. Towards personalized multilingual information access-exploring the browsing and search behavior of multilingual users. In *International Conference on User Modeling, Adaptation, and Personalization.* Springer, Berlin, Germany, 435–446.

Joseph P. Telemala and Hussein Suleman. 2018. Exploring Information Needs and Search Behaviour of Swahili Speakers in Tanzania. In *Maturity and Innovation in Digital Libraries*, Milena Dobreva, Annika Hinze, and Maja Žumer (Eds.). Springer International Publishing, Cham, 185–190.

Evgenia Vassilakaki, Emmanouel Garoufallou, Frances Johnson, and R J Hartley. 2015. An Exploration of Users' Needs for Multilingual Information Retrieval and Access. In *Metadata and Semantics Research*, Gaitanou P Garoufallou E., Hartley R. (Ed.). Vol. 544. Springer, Cham, 249–258.

Markel Vigo, Nicolas Matentzoglu, Caroline Jay, and Robert Stevens. 2019. Comparing ontology authoring workflows with Protégé: In the laboratory, in the tutorial and in the 'wild'. *Journal of Web Semantics* 57 (2019), 100473.

Jieyu Wang and Anita Komlodi. 2018. Switching Languages in Online Searching : A Qualitative Study of Web Users ' Code -Switching Search Behaviors *. In *Proceedings of the CHIIR'18 March 11-15, 2018.* ACM, New Brunswick, NJ, USA, 201–208.

Jieyu Wang, Anita Komlodi, and Omar Ka. 2018. Understanding multilingual web users' code-switching behaviors in online searching. *Proceedings of the Association for Information Science and Technology* 55, 1 (2018), 534–543.

Anping Wu and Jiangping Chen. 2019. Sustaining multilinguality: Case studies of two American multilingual digital libraries. *iConference 2019 Proceedings* ., . (2019), 1–5.

Yusuke Yamamoto and Takehiro Yamamoto. 2020. Personalization Finder: A Search Interface for Identifying and Self-Controlling Web Search Personalization. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20).* Association for Computing Machinery, New York, NY, USA, 37–46. DOI: http://dx.doi.org/10.1145/3383583.3398519

A SUPPLEMENTAL MATERIALS

A.1 Grouping of Topics

Table A1. Grouping of related topics into super-topics

SN.	Super-topic	Topics
1	Religious Faith	Religion, Islam, and Christianity.
2	Higher education	University, University Admission, and Scholarship.
3	IT and electronics	Television, Computer, Computer Hardware, Computer Software, Telecommunications, Internet, and Phones.
4	Justice	Law, Judiciary, and Court.
5	Tourism	Tourism, National park, Restaurant, Resort, Lodging, Hotel, Guide, and Taxi.
6	Health and facility	Health, Medical, Hospital, Clinic, Hiv/aids, Cancer, Heart, and Safety.
7	Education	Science (subject), Chemistry, Mathematics, Education, School, and Books.
8	Earth and Environment	Earth, Environment, Water, Weather, Survey, Waste management, and Energy.
9	HRM and Training	Human resource, Training, Management, and Conference.
10	Lifestyle	Fashion, Clothing, Hairstyle, Beauty, Massage, and Shopping.
11	Agriculture and Food	Agriculture, Farming, Animal, Livestock, and Food.
12	Transportation	Airport, Flight, Transport, Freight transport, Cargo, Railway, Ferry, Car, Motorcycle, and Traffic sign.
13	Business	Accounting, Bookkeeping, Banking, Insurance, Marketing, Sales, Business, Import, and Tax.
14	Economic development	Budget, Planning, Development, Aid, Economy, Industry, and Finance.
15	Sports and Entertainment	chat, Entertainment, Film, Movie series, Movie theater, Game, Sports, Award, Photograph, and Party.
16	Society and culture	Wedding, Culture, Society, News, Social, and Issues.
17	Governance	Public service, Government, President, Constitution, Parliament, Ministry, Election, Embassy, and Security.
18	Family and Gender	Family, Child, Female, and Feminism.
19	Engineering and Construction	Building, Design, Furniture, Engineering, and Electricity.

A.2 Estimating Query Language Preferences in Super-topics and Topics

Table A2. Testing query language preferences in super-topics.

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
1	Religious Faith	30	36	66	41	0.5	0.5	0.75	0.05	0.00	1.00	NP
2	Higher education	50	42	92	55	0.5	0.5	0.75	0.06	0.00	1.00	NP
3	IT and electronics	78	92	170	97	0.5	0.5	0.75	0.05	0.00	1.00	NP
4	Justice	44	30	74	44	0.5	0.5	0.75	0.07	0.00	1.00	Sw
5	Tourism	53	41	94	56	0.5	0.5	0.75	0.06	0.00	1.00	NP
6	Health and facility	102	70	172	98	0.5	0.5	0.75	0.05	0.00	1.00	Sw
7	Education	85	52	137	79	0.5	0.5	0.75	0.04	0.00	1.00	Sw
8	Earth and Environ.	51	59	110	65	0.5	0.5	0.75	0.05	0.00	1.00	NP
9	HRM and Training	64	51	115	67	0.5	0.5	0.75	0.05	0.00	1.00	NP
10	Lifestyle	57	40	97	57	0.5	0.5	0.75	0.08	0.00	1.00	Sw
11	Agric. and Food	75	54	129	75	0.5	0.5	0.75	0.06	0.00	1.00	Sw
12	Transportation	64	51	115	67	0.5	0.5	0.75	0.05	0.00	1.00	NP
13	Business	67	45	112	66	0.5	0.5	0.75	0.05	0.00	1.00	Sw
14	Economic dev.	76	59	135	78	0.5	0.5	0.75	0.04	0.00	1.00	NP
15	Society and culture	87	58	145	83	0.5	0.5	0.75	0.05	0.00	1.00	Sw
16	Sports and Entert.	128	88	216	121	0.5	0.5	0.75	0.04	0.00	1.00	Sw
17	Governance	114	94	208	117	0.5	0.5	0.75	0.06	0.00	1.00	NP
18	Family and Gender	66	46	112	66	0.5	0.5	0.75	0.05	0.00	1.00	Sw
19	Engin. and Const.	38	50	88	53	0.5	0.5	0.75	0.05	0.00	1.00	NP

Where n = total number of responses, x = minimum number of common responses given by $x = (n/2)+z\sqrt{n/4}$, $z = 1.64$, P_0 = probability of common guess, P_d = proportion of distinguishers, $P_{max} = P_d + P_0(1 - P_d)$ = probability of common response @ P_d, $\alpha = 1 - BINOMDIST(x-1, n, P_0, 1)$ = Type I error, $\beta = BINOMDIST(x-1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

Table A3. Testing query language preferences in topics.

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
1	Christianity	15	8	23	17	0.5	0.5	0.75	0.05	0.20	0.80	Sw
2	Religion	14	24	38	24	0.5	0.5	0.75	0.07	0.03	0.97	En

... Continued on next page

Table A3 – continued from previous page

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
3	University	35	13	48	31	0.5	0.5	0.75	0.06	0.02	0.98	Sw
4	University adm.	7	13	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
5	Computer	13	24	37	24	0.5	0.5	0.75	0.05	0.06	0.94	En
6	Computer H/W	21	20	41	27	0.5	0.5	0.75	0.06	0.03	0.97	NP
7	Internet	9	12	21	14	0.5	0.5	0.75	0.09	0.13	0.87	NP
8	Phones	16	11	27	19	0.5	0.5	0.75	0.06	0.11	0.89	NP
9	Software	10	13	23	17	0.5	0.5	0.75	0.05	0.20	0.80	NP
10	Law	21	7	28	19	0.5	0.5	0.75	0.04	0.14	0.86	Sw
11	National Park	22	8	30	21	0.5	0.5	0.75	0.05	0.11	0.89	Sw
12	Heart	13	22	35	22	0.5	0.5	0.75	0.09	0.04	0.96	En
13	HIV/Aids	20	6	26	18	0.5	0.5	0.75	0.04	0.18	0.82	Sw
14	Clinic	23	7	30	21	0.5	0.5	0.75	0.05	0.11	0.89	Sw
15	Mathematics	9	16	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
16	Education	30	13	43	28	0.5	0.5	0.75	0.06	0.03	0.97	Sw
17	School	21	11	32	22	0.5	0.5	0.75	0.06	0.08	0.92	Sw
18	Environment	12	13	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
19	Waste Mgnt	19	10	29	19	0.5	0.5	0.75	0.07	0.09	0.91	Sw
20	Water	9	17	26	17	0.5	0.5	0.75	0.08	0.09	0.91	En
21	Human resrc.	12	14	26	18	0.5	0.5	0.75	0.04	0.18	0.82	NP
22	Management	24	8	32	22	0.5	0.5	0.75	0.06	0.08	0.92	Sw
23	Conference	8	12	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
24	Training	20	17	37	24	0.5	0.5	0.75	0.05	0.06	0.94	NP
25	Fashion	17	13	30	21	0.5	0.5	0.75	0.05	0.11	0.89	NP
26	Agriculture	30	8	38	25	0.5	0.5	0.75	0.04	0.07	0.93	Sw
27	Animals	16	16	32	22	0.5	0.5	0.75	0.06	0.08	0.92	NP
28	Food	15	12	27	19	0.5	0.5	0.75	0.06	0.11	0.89	NP
29	Development	22	11	33	22	0.5	0.5	0.75	0.04	0.10	0.90	Sw
30	Industry	19	10	29	19	0.5	0.5	0.75	0.07	0.09	0.91	Sw
31	Society	23	11	34	23	0.5	0.5	0.75	0.06	0.06	0.94	Sw
32	Culture	15	21	36	24	0.5	0.5	0.75	0.07	0.05	0.95	NP

... Continued on next page

Table A3 – continued from previous page

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
33	Social	20	16	36	24	0.5	0.5	0.75	0.07	0.05	0.95	NP
34	Award	23	14	37	23	0.5	0.5	0.75	0.09	0.03	0.97	Sw
35	Movie series	12	16	28	20	0.5	0.5	0.75	0.04	0.14	0.86	NP
36	Music	17	6	23	17	0.5	0.5	0.75	0.05	0.20	0.80	Sw
37	Game	12	8	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
38	Photographs	31	12	43	28	0.5	0.5	0.75	0.06	0.03	0.97	Sw
39	Election	18	11	29	20	0.5	0.5	0.75	0.03	0.17	0.83	NP
40	Government	16	12	28	20	0.5	0.5	0.75	0.04	0.14	0.86	NP
41	Ministry	11	9	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
42	Public Service	12	19	31	21	0.5	0.5	0.75	0.04	0.13	0.87	NP
43	Child	18	17	35	23	0.5	0.5	0.75	0.04	0.08	0.92	NP
44	Family	15	17	32	22	0.5	0.5	0.75	0.06	0.08	0.92	NP
45	Female	30	10	40	26	0.5	0.5	0.75	0.04	0.06	0.94	Sw
46	Design	10	15	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
47	Engineering	12	11	23	17	0.5	0.5	0.75	0.05	0.20	0.80	NP

Where n = total number of responses, x = minimum number of common responses given by $x = (n/2)+z\sqrt{n/4}$, $z = 1.64$ for $n \leq 30$, P_0 = probability of common guess, P_d = proportion of distinguishers, $P_{max} = P_d + P_0(1 - P_d)$ = probability of common response @ P_d, $\alpha = 1 - BINOMDIST(x - 1, n, P_0, 1)$ = Type I error, $\beta = BINOMDIST(x - 1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

A.3 Estimating Preferences for Language of Results in Super-topics and Topics

Table A4. Testing preferences for language of results in super-topics.

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
1	Religious Faith	53	46	99	59	0.5	0.5	0.75	0.5	0.0001	1.0	NP
2	Higher education	60	36	96	57	0.5	0.5	0.75	0.4	0.0003	1.0	En
3	IT and electronics	134	73	207	116	0.5	0.5	0.75	0.5	0.0000	1.0	En
4	Justice	42	32	74	45	0.5	0.5	0.75	0.4	0.0024	1.0	NP
5	Tourism	68	47	115	67	0.5	0.5	0.75	0.5	0.0000	1.0	En
6	Health and facility	124	144	268	148	0.5	0.5	0.75	0.5	0.0000	1.0	NP

... Continued on next page

Table A4 – continued from previous page

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
7	Education	128	85	213	119	0.5	0.5	0.75	0.5	0.0000	1.0	En
8	Earth and Environ.	143	78	221	124	0.5	0.5	0.75	0.5	0.0000	1.0	En
9	HRM and Training	88	68	156	88	0.5	0.5	0.75	0.6	0.0000	1.0	En
10	Lifestyle	54	50	104	61	0.5	0.5	0.75	0.5	0.0001	1.0	Sw
11	Agric. and Food	78	83	161	92	0.5	0.5	0.75	0.06	0.0000	1.0	NP
12	Transportation	88	73	161	92	0.5	0.5	0.75	0.06	0.0000	1.0	NP
13	Business	91	80	171	97	0.5	0.5	0.75	0.05	0.0000	1.0	NP
14	Economic dev.	89	61	150	86	0.5	0.5	0.75	0.04	0.0000	1.0	En
15	Society and culture	79	65	144	83	0.5	0.5	0.75	0.06	0.0000	1.0	NP
16	Sports and Entert.	129	119	248	138	0.5	0.5	0.75	0.06	0.0000	1.0	NP
17	Governance	163	170	333	182	0.5	0.5	0.75	0.05	0.0000	1.0	NP
18	Family and Gender	68	79	147	84	0.5	0.5	0.75	0.05	0.0000	1.0	NP
19	Engin. and Const.	50	39	89	53	0.5	0.5	0.75	0.04	0.0004	1.0	NP

Where n = total number of responses, x = minimum number of common responses given by $x = (n/2)+z\sqrt{n/4}$, $z = 1.64$, P_0 = probability of common guess, P_d = proportion of distinguishers, $P_{max} = P_d + P_0(1 - P_d)$ = probability of common response @ P_d, $\alpha = 1-BINOMDIST(x-1, n, P_0, 1)$ = Type I error, $\beta = BINOMDIST(x-1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

Table A5. Testing preferences for language of results in topics.

SN.	Topic	En	Sw	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
1	Christianity	20	19	39	26	0.5	0.5	0.75	0.05	0.04	0.96	NP
2	Religion	28	26	54	34	0.5	0.5	0.75	0.02	0.02	0.98	NP
3	Scholarship	19	8	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
4	University	16	18	34	23	0.5	0.5	0.75	0.06	0.06	0.94	NP
5	Univ. admiss.	25	10	35	23	0.5	0.5	0.75	0.04	0.08	0.92	En
6	Computer	20	7	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
7	Software	18	10	28	20	0.5	0.5	0.75	0.04	0.14	0.86	NP
8	Hardware	48	8	56	35	0.5	0.5	0.75	0.04	0.01	0.99	En
9	Phones	25	23	48	31	0.5	0.5	0.75	0.06	0.02	0.98	NP
10	Law	26	11	37	24	0.5	0.5	0.75	0.05	0.06	0.94	En

... Continued on next page

Table A5 – continued from previous page

SN.	Topic	En	Sw	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
11	Court	13	18	31	21	0.5	0.5	0.75	0.04	0.13	0.87	NP
12	Tourism	21	11	32	22	0.5	0.5	0.75	0.06	0.08	0.92	NP
13	National park	16	13	29	20	0.5	0.5	0.75	0.03	0.17	0.83	NP
14	Health	15	11	26	18	0.5	0.5	0.75	0.04	0.18	0.82	NP
15	Medical	10	15	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
16	Hospital	18	13	31	21	0.5	0.5	0.75	0.04	0.13	0.87	NP
17	Clinic	23	31	54	34	0.5	0.5	0.75	0.04	0.02	0.98	NP
18	HIV/Aids	16	31	47	30	0.5	0.5	0.75	0.04	0.03	0.97	Sw
19	Cancer	16	23	39	26	0.5	0.5	0.75	0.05	0.04	0.96	NP
20	Heart	22	6	28	20	0.5	0.5	0.75	0.04	0.14	0.86	En
21	Chemistry	24	3	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
22	Mathematics	12	11	23	17	0.5	0.5	0.75	0.05	0.20	0.80	NP
23	Education	64	49	113	66	0.5	0.5	0.75	0.04	0.00	1.00	NP
24	School	17	14	31	21	0.5	0.5	0.75	0.04	0.13	0.87	NP
25	Environment	44	16	60	37	0.5	0.5	0.75	0.05	0.01	0.99	En
26	Water	44	26	70	43	0.5	0.5	0.75	0.06	0.00	1.00	En
27	Weather	14	12	26	18	0.5	0.5	0.75	0.04	0.18	0.82	NP
28	Waste Mgnt.	13	10	23	17	0.5	0.5	0.75	0.05	0.20	0.80	NP
29	Energy	20	9	29	20	0.5	0.5	0.75	0.03	0.17	0.83	En
30	Human resrc.	42	44	86	52	0.5	0.5	0.75	0.05	0.00	1.00	NP
31	Training	28	16	44	28	0.5	0.5	0.75	0.05	0.03	0.97	En
32	Fashion	37	26	63	39	0.5	0.5	0.75	0.04	0.01	0.99	NP
33	Agriculture	28	25	53	33	0.5	0.5	0.75	0.05	0.01	0.99	NP
34	Farming	13	12	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
35	Livestock	13	24	37	24	0.5	0.5	0.75	0.09	0.03	0.97	Sw
36	Food	23	7	30	21	0.5	0.5	0.75	0.05	0.11	0.89	En
37	Airport	20	20	40	26	0.5	0.5	0.75	0.04	0.05	0.95	NP
38	Flight	22	14	36	24	0.5	0.5	0.75	0.07	0.05	0.95	NP
39	Transport	20	7	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
40	Railway	6	14	20	15	0.5	0.5	0.75	0.06	0.21	0.79	Sw

... Continued on next page

Table A5 – continued from previous page

SN.	Topic	En	Sw	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
41	Accounting	22	13	35	22	0.5	0.5	0.75	0.09	0.04	0.96	En
42	Banking	7	13	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
43	Business	11	12	23	17	0.5	0.5	0.75	0.05	0.20	0.80	NP
44	Import	13	12	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
45	Tax	13	13	26	18	0.5	0.5	0.75	0.04	0.18	0.82	NP
46	Budget	15	10	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
47	Development	32	34	66	41	0.5	0.5	0.75	0.05	0.00	1.00	NP
48	Industry	18	9	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
49	Chat	14	11	25	18	0.5	0.5	0.75	0.05	0.15	0.85	NP
50	Movie series	34	26	60	37	0.5	0.5	0.75	0.05	0.01	0.99	NP
51	Movie theater	25	22	47	30	0.5	0.5	0.75	0.04	0.03	0.97	NP
52	Music	8	25	33	22	0.5	0.5	0.75	0.04	0.10	0.90	Sw
53	Award	10	10	20	15	0.5	0.5	0.75	0.06	0.21	0.79	NP
54	Culture	29	23	52	33	0.5	0.5	0.75	0.06	0.01	0.99	NP
55	News	13	27	40	26	0.5	0.5	0.75	0.04	0.05	0.95	Sw
56	Social	21	9	30	21	0.5	0.5	0.75	0.05	0.11	0.89	En
57	Public Service	21	27	48	31	0.5	0.5	0.75	0.06	0.02	0.98	NP
58	Government	18	9	27	19	0.5	0.5	0.75	0.06	0.11	0.89	En
59	President	22	21	43	28	0.5	0.5	0.75	0.06	0.03	0.97	NP
60	Constitution	15	12	27	19	0.5	0.5	0.75	0.06	0.11	0.89	NP
61	Parliament	29	36	65	40	0.5	0.5	0.75	0.04	0.01	0.99	NP
62	Election	17	30	47	30	0.5	0.5	0.75	0.04	0.03	0.97	Sw
63	Security	32	25	57	36	0.5	0.5	0.75	0.06	0.01	0.99	NP
64	Family	34	45	79	48	0.5	0.5	0.75	0.06	0.00	1.00	NP
65	Child	27	21	48	31	0.5	0.5	0.75	0.06	0.02	0.98	NP
66	Building	13	15	28	19	0.5	0.5	0.75	0.04	0.14	0.86	NP

Where x = minimum number of common responses given by $x = (n/2) + z\sqrt{n/4}$, $z = 1.64$ for $n \leq 30$, P_0 = probability of common guess, P_d = proportion of distinguishers, $P_{max} = P_d + P_0(1 - P_d)$ = probability of common response @ P_d, $\alpha = 1 - BINOMDIST(x-1, n, P_0, 1)$ = Type I error, $\beta = BINOMDIST(x-1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.