Doctoral Thesis

Investigating Language Preferences in Improving Multilingual Swahili Information Retrieval

Joseph Philipo TELEMALA

A Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

in the Department of Computer Science

Faculty of Science

UNIVERSITY OF CAPE TOWN

IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Supervisor:

Prof. Hussein SULEMAN

January 2022

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Joseph Philipo TELEMALA: *Investigating Language Preferences in Improving Multilingual Swahili Information Retrieval,* Thesis presented for the degree of Doctor of Philosophy, © January 2022

SUPERVISOR:
Prof. Hussein SULEMAN

LOCATION:
Cape Town, South Africa

SUBMITTED:
January 2022

DECLARATION

I, Joseph Philipo TELEMALA, do hereby, declare that this thesis is my original work, and that it has never been presented to any University or Institution for any purpose. And that, I know the meaning of plagiarism and declare that all the work in this thesis, save for which is properly acknowledged, is my own.

Cape Town, South Africa, January 2022

Signed by candidate

Joseph Philipo TELEMALA

ACKNOWLEDGEMENTS

First and foremost, I thank the Almighty God for my healthy life and strength throughout the *Safari* of my research, despite the difficult times of the Covid-19 pandemic.

My profound gratitude goes to the Hasso-Plattner Institut (HPI), which enabled me to pursue my PhD studies at the University of Cape Town through the HPI Research School at UCT PhD funding. I deeply appreciate the funding and travel grants.

Prof. Hussein Suleman deserves my heartfelt gratitude not only for his dedication in supervising me, but also for his tireless efforts in making this PhD work a success story. I will never be able to repay him for his invaluable careful guidance, nurturing, encouragement, and mentorship. I appreciate him believing in me and allowing me to work with him despite the fact that we had never met or known each other prior to joining the program.

My family has been so supportive that I cannot believe I have been away for three years and they have remained so strong and happy. My wife, Hollo Kayanda, and our four children, Cephlen, Christy, Charity-Annie, and Chelsie are the most amazing people the Almighty has blessed me with. I am grateful for your unending support. May the Lord God abundantly bless you.

I would like to thank everyone who helped me with my research for their invaluable inputs, insights, and any other assistance that made the entire data collection possible. I would like to express my gratitude to my colleague Odoyo Daud for his contributions to the data collection platform.

I would like to express my heartfelt gratitude to the Digital Libraries Laboratory research group members for their invaluable assistance and ideas sharing. I would like to express my gratitude to Zola Mahlaza, Wanjiru Mburu, Jecton Anyango, and Tezira Wanyana in particular. Also, thanks to the members of the HPI Research School at UCT, Room 300 lab, and ICT4D lab for the great interactions and fun.

Thanks to my friends Alice Bakera, Ansfrid Lekundayo, Stefaan, Mariam Nguvava, Neema Kahabi, Kipembawe, and Elizabeth Moirana. The hiking and long walks kept me sane and physically fit while also assisting me in managing the PhD stresses.

Finally, I would like to express my gratitude to my parents, Philipo Telemala and Cephlen Luge, as well as my siblings, Zabron, Timothy, Hezron, Bernard, Augustino, Pendo, Luge, Allen, Neema, and Mhoja, for their unwavering support in my studies and to my family while I was away.

To my family

my parents Philipo T. Telemala and Cephlen M. Luge,

my wife Hollo S. Kayanda,

my children Cephlen, Christy, Charity-Annie, and Chelsie

for their love *to me.*

ABSTRACT

Multilingual Information Retrieval (MLIR) systems are designed to retrieve information from multiple languages in response to a query posed in another language or in one of the languages in which a user is looking for information. Researchers have proposed a number of approaches for combining the results from individual result lists to produce a single final result list. Some are heuristics, such as round-robin, in which a result is drawn from each result list one at a time until all lists are exhausted, while others are Machine Learning (ML)-based, in which a model is trained using a variety of features from the query and the required documents.

These approaches strive for topical relevance, which is the most important goal in satisfying users' information needs. However, multilingual speakers exhibit a variety of behaviors, some of which are unique to certain individuals based on their historical, cultural, and linguistic backgrounds. Unfortunately, these behaviors are ignored in the current MLIR system design and implementation. Current MLIR systems, in particular, present results that do not take people's language preferences into account when ranking results. Studies have shown that users have different language preferences based on their search topics – *Topic-Language (T-L)* preferences. This study proposes using T-L preferences to improve the relevance of the ranked MLIR results.

To achieve this aim, we used a survey-based study to try to understand the information needs and Web search behaviour of Swahili-speaking Web users in Tanzania. One bold behaviour of such multilingual Web users that emerged is code-switching. Several factors, such as information context and search topic, were identified as reasons for such frequent language switching.

We then created a prototype multilingual search engine with which users interacted in order to quantify how much the language of the query or the selected results is influenced by the search topic. We estimated the relationship between the topic of search and the language of the query and clicked results using the resulting query and click-through logs. The findings revealed that Swahili-speaking Web users have language preferences for certain topics. For example, Kiswahili was significantly preferred as a results language in only 9% of the examined topics, English was preferred in 26% of the topics, and there was no preference for language of results in the remaining 65% of the topics.

Based on these findings, we created the T-L-based algorithm, which re-ranks the results based on T-L associations/preferences. We evaluated our proposed T-L-based algorithm using click-through logs from our prototype guided multilingual search engine. The results show that incorporating language preferences into the ranking model significantly improves the relevance of MLIR results in some specific cases. The strength of the T-L association and the number of relevant results in the preferred language's list were discovered to be driving factors in the performance improvement of the T-L-based algorithm.

This thesis provides evidence that using language preferences can potentially improve the relevance of MLIR results for some topics that are preferentially expressed in specific languages. This is important in communities where information search and access are hampered by a variety of factors and there is a clear lineage in language use, where MLIR's topical relevance alone may not be sufficient.

PUBLICATIONS

Some of the ideas, early versions, data, tables, figures, proof of concept toolkits, and experimental results presented in this thesis have previously appeared in the following publications:

[1] J. P. Telemala and H. Suleman, "Exploring information needs and search behaviour of swahili speakers in tanzania," in *Maturity and Innovation in Digital Libraries*, M. Dobreva, A. Hinze, and M. Žumer, Eds., ser. ICADL 2018. Lecture Notes in Computer Science, vol. 11279, Hamilton, New Zealand: Springer International Publishing, Cham, 2018, pp. 185–190, ISBN: 978-3-030-04257-8. DOI: 10.1007/978-3-030-04257-8_18.

[2] ——, "Exploring topic-language preferences in multilingual swahili information retrieval in tanzania," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 6, Nov. 2021, ISSN: 2375-4699. DOI: 10.1145/3458671.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACM Association for Computing Machinery

AP Average Precision

API Application Programming Interface

BAKITA Baraza la Kiswahili la Taifa

BERT Bidirectional Encoder Representations from Transformers

CLEF Conference and Labs of the Evaluation Forum

CLIR Cross-Lingual Information Retrieval

XLM Cross-lingual Language Model

CM Cascade Model

CR Class Representative

DARPA Defense Advanced Research Projects Agency

DBN Dynamic Bayesian Network

DCG Discounted Cumulative Gain

DT Document Translation

EC European Commission

ERR Expected Reciprocal Rank

FIRE Forum for Information Retrieval Evaluation

FSREC Faculty of Science Research Ethics Committee

GMAP Geometric Mean Average Precision

HCI Human-Computer Interaction

HCS Helsinki Corpus of Swahili

HRM Human Resource(s) Management

IR Information Retrieval

ISO International Organization for Standardization

IT Information Technology

L2R Learning-to-Rank

MAP Mean Average Precision

MBERT Multilingual Bidirectional Encoder Representations from Transformers

ML Machine Learning

MLIR Multilingual Information Retrieval

MRR Mean Reciprocal Rank

MS Microsoft

MT Machine Translation

MUHAS Muhimbili University of Health and Allied Sciences

NDCG Normalized Discounted Cumulative Gain

NER Named Entity Recognition

NLP Natural Language Processing

NMT Neural Machine Translation

NSF National Science Foundation

NTCIR National Institute of Informatics Text Collection for IR

POS Part-of-Speech

QT Query Translation

RR Reciprocal Rank

R-R Round-Robin

SADiLaR South African Centre for Digital Language Resources

SALAMA Swahili Language Manager

SERP Search Engine Results Page

SIGIR Special Interest Group in Information Retrieval

SMT Statistical Machine Translation

SNAL Sokoine National Agricultural Library

SUA Sokoine University of Agriculture

SUZA State University of Zanzibar

SVM Support Vector Machines

TDIL Technology Development for Indian Languages

T-L Topic-Language

TREC Text REtrieval Conference

UCT University of Cape Town

UDOM University of Dodoma

UDSM University of Dar Es Salaam

URL Uniform Resource Locator

US United States

WWW World Wide Web

Part I

BACKGROUND

1 INTRODUCTION

The World Wide Web (WWW) (or the Web) is made up of documents in a variety of languages from all over the world. Through Information Retrieval (IR) technologies such as search engines, Web users can access information from such vast collections. Traditionally, IR systems are designed to retrieve information from document collections in a single language. However, in systems known as Multilingual Information Retrieval (MLIR), they can be extended to support information retrieval from multiple languages. This means that a user of such systems can pose a query in one language and receive responses in multiple languages.

The current MLIR systems share the same IR perspective in that they strive to present the best possible results in terms of the topic of information to satisfy the user's information needs, which is essentially desirable. The behaviour of IR users, on the other hand, may differ from that of MLIR. This is due in part to the fact that IR users may only speak one language, whereas MLIR users speak multiple languages – polyglots. This distinction may result in different behaviour and expectations from these groups of people.

The current design of MLIR systems ignores other factors unique to polyglots, such as cultural, linguistic, and historical backgrounds. This may reduce the relevance of the results to the perspectives of some of these users. To supplement topical relevance, this thesis proposes to investigate a factor called language preferences on improving relevance of the ranked MLIR results. We concentrate on Tanzania's Swahili-speaking Web users because they are polyglots with a clear and unique distinction between English and Swahili language use.

This chapter will cover the top highlights of the research in terms of the context and motivation for carrying out this study in Section 1.1, followed by a brief description of how human behaviour can be incorporated into MLIR design to improve the relevance of the results in Section 1.2. The research problem is stated in Section 1.3, and the research questions are defined in Section 1.4. Section 1.5 highlights the methods and materials used to conduct the research in this thesis. The chapter ends with an organization of the thesis in Section 1.6.

1.1 CONTEXT AND MOTIVATION

1.1.1 *Multilingualism on the Web*

The content on the World Wide Web (WWW) (or Web for short) has been growing at an exponential rate, not only in English (the original language used on the Web), but also in other languages. The top ten languages with the most Wikipedia articles, for example, are shown in Figure 1.1. This rapid increase can be attributed to the multilingual nature of the world's population, i.e., people can speak and/or understand multiple languages – polyglots – and thus, they write and/or share information in those languages. This is evident in social me-

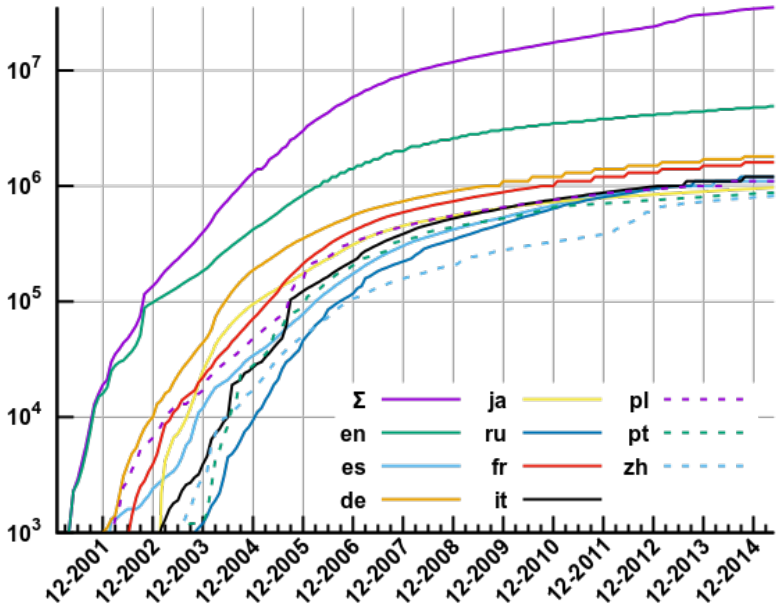


Figure 1.1: Top 10 languages with the most articles on Wikipedia

Where $\sum$ – Total, en – English, es – Spanish, de – German, ja – Japanese, ru – Russian, fr – French, it – Italian, pl – Polish, pt – Portuguese, and zh – Chinese.
Source: Wikipedia

dia platforms such as Facebook and Twitter, where users can post content in any language they are fluent in.

As a result, polyglots may be interested in information in another language, for example, a news reporter who wants to access news that is well covered in another language, or a user who is having difficulty composing a query in one language. The increased content in multiple languages and multiple language speakers (polyglots) poses a challenge to the field of Information Retrieval (IR) research. Because of the multilingual nature of the world and the need to meet users' information needs without language barriers, two related subfields of IR (Cross-Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR)) are of research interest. The goal is to better retrieve information from languages other than English, or at least in addition to English.

CLIR strives, by definition, to retrieve documents from a resource collection whose language differs from the query [158]. For example, a user may query in Arabic and receive results in English. MLIR is concerned with information retrieval in virtually any language [65]. Even if the documents, query, or both are in multiple languages, retrieval is possible. MLIR extends the concept of CLIR, which is basically limited to a pair of documents and requires either the query to be translated to match the language of the documents or the other way around [158].

Multilingual Information Retrieval (MLIR) systems provide users (e.g., Web searchers) with more comprehensive results than monolingual IR systems by utilizing the vast amount of information available in languages other than the language of the query [166]. There is a growing interest in developing MLIR systems

that support low-resource languages such as Kiswahili, as demonstrated by the works of Yarmohammadi, Ma, Hisamoto, *et al.* [226], Zbib, Zhao, Karakos, *et al.* [228] and Zhang, Westerfield, Shim, *et al.* [230]. This thesis focuses on improving results in the context of MLIR.

1.1.2 *Web Search User Behaviour*

The multilingualism of Web users, as well as the availability of multilingual content on the Web, encourages some polyglots to pose queries in their native languages or any language in which they are fluent. As a result, language switching (code-switching) occurs at various points during the course of a Web search. Researchers in Human-Computer Interaction (HCI) and IR are increasingly interested in studying the reasons for code-switching on Web search [197]. Studies have identified several factors, including the availability of documents/resources on the Web [109], [216], the searcher's language proficiency [196], [197], the nature of the task the searcher wishes to complete [10], [70], [120], the topic of the search [10], [120], and query formulation challenges [152], [190].

The majority of the studies from which these factors are derived are based on monolingual IR systems (search engines), which can handle queries in multiple languages. Fewer studies have focused on comprehending language preferences when users query and/or access information from an MLIR. As a result, it is unclear whether the same factors that influence code-switching in monolingual IR also influence language preferences in MLIR.

1.1.3 *MLIR Implementation Challenges*

There are two approaches to developing an MLIR system: centralized architectures and distributed architectures. The centralized MLIR system has a central corpus of all document collections, known as a mixed corpus, and the original query is translated to all supported languages [149]. The distributed approach separates individual document collections and retrieves the results in three steps: i) translating the query to match the languages of the supported document collections; ii) performing document retrieval in a traditional IR style on each collection; and iii) performing result merging, in which the system combines the individual result lists and presents them. This thesis' research is based on a distributed architecture.

Implementing MLIR using a distributed approach presents translation and merging challenges [116], [158]. Rahimi, Shakery, and King [166] assert that, if the translation of the query or documents is perfect, the retrieval model can get relevant results in the independent monolingual result lists. The final difficult problem is determining how to combine (merge/re-rank) the individual result lists to produce the final result list. The merging problem is difficult, in part because each result list to be merged has incomparable document scores, owing to the various statistics used to score the documents [149].

Several approaches to merging MLIR results have been proposed in the existing literature. Some use the relevance scores of each document to the query, such as raw-score [181] and normalized-by-topk [116], while others, such as round-robin

[181], [214], ignore the scores. Other approaches to merging employ ML strategies such as logistic regression [111], handcrafted features [206], and multi-view [209]. Other approaches such as those in the works by Gialampoukidis, Moumtzidou, Tsikrika, *et al.* [72], Liu, Meng, Qiu, *et al.* [122], and Qin and Liu [164], which are specifically designed for federated search, can also be used for results merging. These convergent approaches are concerned with achieving topical relevance. In addition to topical relevance, the current thesis proposes to incorporate user behaviour to improve the relevance of merged MLIR results in terms of the preferred language of search.

1.1.4 *Kiswahili and Swahili-speaking Web Users*

Kiswahili is a Bantu language that descends from the Benue-Congo branch of the Niger-Congo language family [64], [205]. It is primarily spoken in East and some parts of Central Africa, with an estimated population of over 100 million speakers [64], [221]. The characteristics of Swahili speakers differ greatly from country to country, owing to differences in education systems and culture. While Kiswahili is only taught in primary and secondary schools in Kenya and Uganda, it is a medium of instruction in public primary schools and adult education in Tanzania [172]. For high school and tertiary levels of education, the medium of instruction shifts to English [172], [205]. English is also used in district and high courts, international transactions, international diplomacy, and as the language of science and technology [154].

Furthermore, Tanzania uses Kiswahili as a national and official language, while English is an official language [154], [205]. Tanzania, in addition to these languages, has over 110 tribes, each with its own tribal language [148]. This implies that Swahili speakers in Tanzania are polyglots, able to communicate in at least two languages, namely tribal language as the first language for most rural dwellers, Kiswahili as the second language (or first language for most urban dwellers), and English as the third language (or second language for most urban dwellers). Because there are no textual documents in those languages, the availability of tribal languages on the Web is almost non-existent.

Despite the fact that Kiswahili and English have official status in Tanzania, their use in daily business by citizens varies greatly. It is unusual, for example, to see such people conversing in English on the streets or in office corridors, even among highly educated individuals, government officials, university students, and judicial officers. They primarily communicate in Kiswahili. Ngonyani [148] summarizes different domains in which the two languages are used. The author shows that Kiswahili dominates, as a communication language, almost all domains of life such as politics, mass media, local business, worship and literature. English, on the other hand, is used alongside Kiswahili for education and commerce.

When compared to other multilingual communities around the world, there is a clear separation in how the languages are used in Tanzania's Swahili-speaking community, where English is merely a language of records (documents), which means that all official documents must be recorded or written in that language. Other multilingual communities may exhibit different characteristics, such as

equal use of languages in all spheres of life. In Kenya, for example, the Swahili-speaking community can use both Kiswahili and English equally in parliament, whereas this is not the case in Tanzania.

This distinction in how Swahili speakers use the two languages complicates the mechanism by which users interact with information on the Web. This exacerbates the consequences and implications for MLIR design and implementation. Given that polyglots are mostly fluent in multiple languages, the MLIR system design is centered on achieving topical relevance while ignoring other factors. However, given the use of English and Swahili in Tanzania, the information needs for work-related tasks and the language of information they use may differ from those for non-work-related tasks. The unique characteristics of Tanzania's polyglots necessitate a study that not only focuses on achieving topical relevance, but also incorporates human behaviours to improve the relevance of MLIR system results.

1.2 INCORPORATING USER BEHAVIOUR IN MLIR

This thesis proposes to investigate the code-switching factors identified in monolingual IR in an MLIR environment. The thesis is specifically about the *topic of search* factor, which has been identified in several survey-based studies, including by Steichen, Ghorab, O'Connor, *et al.* [196], Wang and Komlodi [216], and Lowe and Steichen [129]. Because users prefer a specific language under certain conditions, such as search topic, we argue that such users may also want their preferences and expectations to be incorporated into the MLIR system. According to Chu and Komlodi [32] and Nzomo, Vaughan, Ajiferuke, *et al.* [153], users prefer to see a system that takes their search behaviour into account.

The behaviour of switching or preferring a specific language for a specific topic of search is referred to as *Topic-Language (T-L) preference* in this thesis. For the purposes of this thesis, we may also use the term *language preferences* to mean the same thing as *Topic-Language (T-L) preference*.

1.2.1 *Example Scenario*

The example scenario below depicts T-L preferences for a typical polyglot Swahili-speaking MLIR user.

> The user has an information need in the *tourism* topic, and two or more language options for use, such as *Kiswahili* and *English*. The user composes a query in Kiswahili. The system retrieves and displays results in the two languages: Kiswahili and English, disregarding the query language – MLIR displays results in all the supported languages regardless of query language. With these results, the user will obviously click on the most relevant result(s), oblivious to the language he/she used to compose the query, such that clicked results/Uniform Resource Locators (URLs) are from both languages.

We can reasonably expect such interactions from a large number of users over a specific time period to produce patterns or associations of interest. The associ-

ations can be between the search topic and the language used to formulate the query, or between the search topic and the language of the clicked results. This thesis refers to these patterns as *Topic-Language (T-L) association*, which leads to the concept of *T-L preferences*, which is used in this thesis. As a result, the current thesis argues that these associations (T-L preferences) may improve the relevance of results in MLIR.

Consider the case of multilingual Web users with known language preferences for search topics. Assume 50 results are presented to a user in an interleaved format, with one result in language A and the next result in language B. This thesis proposes that the results in a preferred language for that topic of search be prioritized and placed at the top of the list based on the users' language preferences. The thesis attempts to answer the question, "What is the ideal set of results?" that should be presented to the user for a query in a specific topic.

We believe that incorporating MLIR user behaviour, specifically T-L preferences, into the ranking process will improve the relevance of ranked MLIR results, at least in terms other than topical relevance, such as preferences and experiences. And that the improvement may benefit certain types of MLIR users, primarily Swahili-speaking MLIR users, in certain scenarios. Meeting the expectations and preferences of MLIR users from various backgrounds is a success for MLIR. Users who speak low-resource languages, such as Kiswahili, for example, may benefit from relevant results in their preferred language even if they struggle with language fluency and query formulation.

1.3 RESEARCH PROBLEM

To the best of our knowledge, current MLIR systems do not take language preferences (T-L preferences) into account when ranking search results, particularly during the merging stage. As a result, the system may hide potentially relevant results further down the list, and users either miss them or expend extra effort to find them. Current MLIR systems heavily rely on topical relevance as a metric of success in meeting the information needs of users.

According to Ghorab, Zhou, Steichen, *et al.* [71] and Steichen and Freund [195], it is not always the topical relevance that gives users of a system satisfaction; meeting their expectations and preferences is another important aspect. We argue that users are more likely to be satisfied when they first see top-ranked results in the language they want or prefer, rather than scrolling to find results in their preferred language. Thus, we propose incorporating language preferences into the MLIR results ranking to supplement the topical relevance.

1.4 RESEARCH QUESTIONS

We propose a study that makes use of an MLIR system that supports two languages, Kiswahili and English. The thesis seeks to determine how multilingual Swahili-speaking Web users' language preferences can improve MLIR, keeping in mind that English has more resources on the Web than Kiswahili. The scientific goals of this thesis are to estimate T-L preferences in MLIR systems and to investigate the effect of T-L preferences on improving the relevance of ranked results

in MLIR. The thesis aims to provide an answer to the following primary research question.

> *How can the association between topic and language (T-L association) in Multilingual Information Retrieval (MLIR) be estimated, and how can it improve relevance of ranked results in a multilingual Swahili Information Retrieval (IR)?*

To answer this primary research question, we first attempt to understand the information needs and search behaviour of potential MLIR users in current Web use; second, estimate the T-L preferences/associations as a result of user interaction with an MLIR system; and finally, use the T-L preferences in the ranking of MLIR results and evaluate their impact on improving relevance of the results. To accomplish these goals, the primary research question was divided into three research questions.

RQ1 *What are the information needs and search behaviours of Tanzanian polyglot Swahili-speaking Web users?*

This research question aimed to understand the problem domain, which is essential for proposing and/or developing an informed solution. We targeted the population of Swahili-speaking Web searchers in the United Republic of Tanzania. We conducted a survey with Tanzanian librarians/information experts and Swahili specialists. We sampled public and university libraries, Kiswahili departments from various universities, and the Tanzanian Kiswahili Council.

The interviews with key informants from these institutions aimed to gain a better understanding of the problem domain by looking into the search behaviour of Swahili-speaking Web users as well as the needs and uses of Swahili information in various sectors. This survey of Swahili-speaking Web users assisted in understanding the key focus areas for our study by eliciting experiences, perceptions, and opinions from those under investigation.

This research question is addressed in Chapter 4, while the supporting and supplementary materials used to conduct the survey study are presented in Appendix A.

RQ2 *What are the topic-language preferences of the polyglot Swahili-speaking users of the multilingual Swahili Information Retrieval (IR) system?*

After identifying one interesting behaviour among Swahili-speaking Web users – switching the language of search based on the topic of search – we devised an experiment to capture such behaviour implicitly. Thus, to address this research question, we created a prototype multilingual search engine (MLIR system) that supports Kiswahili and English, powered by the Microsoft (MS) Bing Web search Application Programming Interface (API). Tanzanian Swahili-speaking Web users interacted with the system using a prepared set of queries.

We attempted to deduce the association between topic of search and query or results language based on data (query and click-through logs) generated from user interaction with the MLIR system. We analyzed the query and click-through logs from this carefully controlled MLIR system using preference-based statistics to reveal the T-L preferences. Determining these associations/preferences led to

a better understanding of the latent behaviour of multilingual Swahili IR system users.

The question is addressed in Chapter 5, and the supplementary materials are provided in Appendix B.

RQ3 *How can topic-language preferences improve the relevance of ranked results in a multilingual Swahili Information Retrieval (IR) system?*

This question aims to incorporate the identified T-L preferences into the MLIR system's ranking and assess how much they improve the relevance of the ranked MLIR results. We developed a T-L-based algorithm, which included T-L preferences in the ranking. The click-through logs were primarily used in this evaluation. We then evaluated its performance using standard IR evaluation metrics such as Average Precision (AP), Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG).

The proposed T-L-based algorithm in Chapter 6 and its evaluation in Chapter 7 provide answers to this question.

1.5 RESEARCH METHODOLOGY

This study employs *empirical bottom-up* methods [7], which are a collection of empirical studies that use a bottom-up approach to knowledge development. Bottom-up approaches, according to Amigó, Fang, Mizzaro, *et al.* [7], use test cases derived from real-world scenarios and pre-existing IR models. Using this combination of empirical studies and bottom-up approaches in IR allows for empirical validation of systems, where users are mostly subjected to the system in a controlled laboratory setting. Massive log analysis is also included in the empirical bottom-up methods [7].

This thesis' research employed a combination of survey, controlled experiment, and query and click-through logs analysis. To develop and address the primary research question, we went through the following steps.

- *A comprehensive literature review* to understand the state-of-the-art solutions in MLIR and multilingual Swahili IR in particular.
- *A survey-based study* to understand the problem domain in terms of the search behaviour and information needs of Swahili-speaking Web users.
- *Creating a multilingual Swahili IR prototype system* for users to interact with and generate query and click-through logs.
- *A statistical analysis* of query and click-through logs to estimate T-L preferences.
- *An evaluation* to assess the efficacy of the T-L-based approach to improving the relevance of multilingual Swahili IR.

1.5.1 *Research Design*

This thesis employed two primary research designs: survey and controlled experiment. Prior to administering a data collection tool such as an interview or a questionnaire, survey studies typically involve an in-depth sampling of participants from the targeted population [95]. Controlled experiments frequently employ manipulated variables, with subjects divided into control and experimental groups [95]. We used all of the subjects, but limited their ability to change the variables. The study in Chapter 5 describes a prototypical multilingual Swahili search engine, with users querying it using prepared queries and topics.

1.5.2 *Data Collection*

1.5.2.1 *Survey Study*

Survey-based studies involving human subjects typically include a list of prepared questions designed to elicit specific information from respondents such as opinions, thoughts, and insights [95]. A survey's goal could be to provide explanation, description, and/or exploration, and it could use interviews or questionnaires to accomplish this. Interview research was used in our cross-sectional survey presented in Chapter 4 to explore and understand the information needs and search behaviour of Swahili-speaking Web users. We spoke with information science experts as well as Kiswahili specialists.

1.5.2.2 *Query and Click-logs*

Query and click-through logs can effectively capture traces of human behaviour on the Web, where users demonstrate their actual behaviour rather than recalled behaviour [5]. The query and click-through logs were analyzed in our study in Chapter 5 to estimate the latent T-L preferences in topics and super-topics. The click-through logs are also used in Chapter 7 to evaluate the proposed T-L-based approach for merging multilingual Swahili IR.

1.6 THESIS ORGANIZATION

This thesis is mainly divided into five parts.

Part I establishes the context for the argument presented in this thesis.

- Chapter 1 provides an overview of the background and motivation for conducting this research. It also includes the research problem, research questions, and a summary of the methods and materials used to answer the research questions.
- Chapter 2 presents the state-of-the-art in MLIR, with a focus on the detailed background of MLIR, such as history and architecture of MLIR, and evaluation measures in the general IR field, which also apply to MLIR. The second section of the chapter includes a review of the literature on information needs as well as the results of merging strategies in MLIR.

- Chapter 3 restricts the review of literature to the case of Kiswahili. It presents several cutting-edge solutions to Swahili IR, Natural Language Processing (NLP) and Machine Translation (MT).

Part II covers Swahili Web users' perspectives on information search behaviour and further investigates how the topic of search is associated with the language of both query and results

- Chapter 4 explores the information needs and search behaviour of Swahili-speaking Web users.
- Chapter 5 investigates the relationship between topic of search and the preferred language of query or results (T-L associations).

Part III focuses on the evaluation of the proposed approach that uses Topic-Language (T-L) association.

- Chapter 6 introduces the T-L-based algorithm, which is an approach for merging MLIR results.
- Chapter 7 presents an evaluation of the proposed T-L-based algorithm.

Part IV covers the thesis conclusions and suggestions for further research on the multilingual Swahili IR.

- Chapter 8 summarizes the major findings of the study, describes the study's main contributions, and outlines the direction for future research on multilingual Swahili IR.

Part V contains all the additional materials and resources used in carrying out this research.

- Appendix A contains the materials for a study presented in Chapter 4.
- Appendix B presents materials for a study presented in Chapter 5.

2

MULTILINGUAL INFORMATION RETRIEVAL

This chapter provides background information as well as a review of the literature on MLIR. On the background aspect, it covers the fundamental concepts of Information Retrieval (IR), Cross-Lingual Information Retrieval (CLIR), and Multilingual Information Retrieval (MLIR). The history and architectures for developing MLIR systems are then highlighted. It also highlights the evaluation measures that are commonly used in the evaluation of IR systems.

The literature review part of this chapter is concerned with the related research works presented in this thesis. We begin by looking at studies that investigated information needs and search behavior without regard for whether or not the focus was on Web users. We then focus on research on information needs and search behavior in the Swahili-speaking region. The literature review section then delves into Web users' search behavior, with a focus on their code-switching habits. Because there were few works on multilingual search engines, we examined a number of works that used monolingual search engine system users. The final part of the literature review section looks at methods for merging results in MLIR. That is, we examine various works that proposed methods for combining results from various sources to produce a single results list.

The literature review in this chapter forms the first part of the thesis literature review; the specific works on Swahili IR, Natural Language Processing (NLP), Machine Translation (MT), and CLIR/MLIR will be covered in the second part of literature review in Chapter 3.

This chapter is structured as follows. Section 2.1 provides some key terminology related to IR in general and MLIR in particular. Section 2.2 presents the historical perspective of MLIR development. The chapter then presents the architectural frameworks for developing MLIR system in Section 2.3. The IR system evaluation metrics are in Section 2.4. The rest of the sections are concerned with related literature, specifically focusing on information needs and search behaviour in Section 2.5, code-switching in Web search in Section 2.6, and the cutting-edge MLIR results merging approaches in Section 2.7. The chapter ends with a summary in Section 2.8.

A: BACKGROUND

2.1 DEFINITIONS OF KEY CONCEPTS

There is a long history of organizing and managing information sources to allow for easy access and retrieval of information. Libraries have existed for generations to facilitate information access and search. In the early 1950s, the invention of computers resulted in the introduction of computer-based systems to enable the organization, storage, management, dissemination, and retrieval of information – Information Retrieval (IR) systems [11], [174]. IR came into play for two primary reasons: one, the traditional cataloging strategy, which was commonly used in libraries, could not keep up with the growing number of documents; and two, the unprecedented type of information/data that emerged with the invention of computers, which was different from the textual data typically found in traditional libraries [178].

Libraries and IR systems share the goal of locating information based on the information needs of the user. An IR system finds information in unstructured and/or semi-structured resources, such as text (e.g., documents and Web pages), videos, audios, and images. Unlike structured data in databases, dealing with unstructured data is difficult. IR system users have varying levels of information needs, intentions, and preferences.

The complexity of the information source and the information needs of the user adds another challenge to determining the success of an IR system in delivering useful information to the user. The measures vary, with some focusing on the system, such as speed (of both indexing and searching), and others on the user, such as satisfying their information needs (or answering their queries). Some metrics, such as system speed, ability to express and resolve complex queries, and so on, can be measured (quantified).

However, the fundamental challenge with IR is measuring user satisfaction, which is difficult to quantify. For example, there is no point in developing a fast IR system that retrieves useless information. Thus, IR research employs the concept of *relevance* to quantify the success of an IR system.

The concepts introduced in this section are related in such a way that the IR, CLIR and MLIR systems are built around the need to satisfy a user's *information need*.

2.1.1 *Information Needs*

Information need is a cognitive need of humans [222]. It is perhaps similar to a human's basic needs (food, shelter, and clothing); the only difference is that this is cognitive in the sense that it assists him/her in reasoning, remembering, and deciding (to make informed decision). Information needs can range from a desire to improve/expand or even correct knowledge [107] to a simple perception of a lack of information in one's mind that causes him/her to develop a desire for it [169], [222].

There are numerous sources and channels through which one can obtain information to meet their needs. These might be: traditional oral sources, such

as directly asking or listening to a friend, colleague, elder, and expert; printed sources such as books, newspapers, brochures, and magazines; and electronic sources such as the Web (Internet) – social media, blogs, websites – radio, and television [3].

The invention of the Internet (and the Web in particular) has dramatically changed the way humans who have access to it obtain information. Because of the Internet's widespread penetration and access, humans now rely primarily on electronic sources organized in the World Wide Web (WWW) or Web for short [184]. Clarke, Belden, Koopman, *et al.*'s [35] study on physicians, for example, found that there is an increase in the use of the Web as their primary source of information.

The availability of information in electronic form, particularly on the Web, has resulted in a slightly altered definition of information need. In the Web environment, Schultz [184] defines an information need as the amount of information required by a Web user to fulfill his/her search intent. This thesis views an information need in the Web environment as encompassing a broad range of aspects, such as a knowledge gap that needs to be filled with information or uncertainty that requires resolution from Web sources.

However, obtaining information from the Web is not a simple task. For example, as of June 2021, the Google search engine had indexed over *50 billion pages*[1]. Unlike oral sources and relational databases, there is no formal information structure on the Web (unstructured). There must be a way to easily and efficiently obtain information from such massive collections of information. As a result, IR systems are required.

In IR, it is common practice for users to express their information needs as a *query*, which is a simple string of text (or multimedia, as in image search), on some form of IR system interface. This adds to the already difficult problem of determining relevance. That is, how does the IR system translate the expressed query in order to derive and retrieve useful information based on the user's latent information need? This is due to the fact that relevance is related (or evaluated relative to) an information need rather than a specific query. For example,

Information need: *I'm looking for information on the opportunities and challenges of establishing a tourist agency in poor countries.*

Query: *tourist agency*

2.1.2 *Search Query*

A query is an information need that is presented to the IR system. Different IR systems support various forms of expressing information needs, most notably text, voice/speech, video and image [31].

1 https://www.worldwidewebsize.com/

2.1.3 *Information Retrieval (IR)*

The classical definition of IR was proposed by Salton [174]; it states that:

> "*Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.*"

Baeza-Yates and Ribeiro-Neto [11] define IR as a process that encompasses

> "*...representation, storage, organization of, and access to information.*"

In terms of the access to information or searching component, according to Baeza-Yates and Ribeiro-Neto's definition, IR matches the requested information/resources from specific (typically unstructured) sources to satisfy a user's *information need*. Plain-text in documents, documents themselves, metadata, images, videos, sound, and text databases may all be involved as source collections. As a result, before being displayed to the user, the information is properly stored, organized, and represented.

2.1.4 *Information Retrieval Models*

A model in IR is used to predict the relevant documents for the user's query [80]. The IR models can be classified based on their modeling approaches, which include Boolean models, vector space models, and probabilistic approaches (i.e., probabilistic retrieval models, Bayesian Network models, and language models). For reference, Hiemstra [80] provides a thorough examination of these IR models.

2.1.5 *Applications of Information Retrieval*

An IR system is a software system that handles the IR process by comparing a set of documents to the submitted query(ies) to generate a list of ranked results/-documents [44]. The system manages how the retrieved information/resources are represented, stored, organized, and accessed by the end user [11]. Search engines, information filtering and recommendation systems, digital libraries, media search, and domain specific systems, such as geographic IR (GIR), vertical search, and legal IR [113], are examples of common applications (systems) where IR techniques are used.

In classical IR, all the retrieved resources must be in one language [158]. However, the digital space has a lot of information in different languages. Thus, there are sub-fields of IR to address such cases as either Cross-Lingual Information Retrieval (CLIR) or Multilingual Information Retrieval (MLIR). According to [158], all retrieved resources in classical IR must be in one language. However, the digital space contains a wealth of information in a variety of languages. As a result, there are sub-fields of IR to address such cases – Cross-Lingual Information Retrieval (CLIR) or Multilingual Information Retrieval (MLIR).

2.1.6 *Cross-lingual Information Retrieval (CLIR)*

Cross-Lingual Information Retrieval (CLIR) is a sub-field of IR that deals with retrieving results from a collection of resources in a language other than the query [158]. Querying from a collection with documents in a language other than the query implies that translation, merging, and presentation are required, [158]. For example, suppose a user is not competent enough to compose an English query or for some reasons expresses her query in Kiswahili and wishes to see results from a collection of English documents. The CLIR system translates the Swahili query to English, then performs the standard IR processes in the English document collection and returns English results.

The most common translation techniques used in CLIR are: *Document Translation (DT)* – documents are translated into the language of the query; and *Query Translation (QT)* – the query is translated into the languages of the documents [149], [158]. However, DT has a number of limitations, including: storage requirements, particularly when the number of supported languages is large; maintaining an up-to-date collection of translated documents; and time consumption in creating the translated documents [166].

QT avoids most of these drawbacks while incurring little overhead when compared to translating the entire resource collection, making it a preferred option [158]. The disadvantage of this approach, though, is that it produces poor translations of ambiguous (typically short) queries and is slightly slower due to on-the-fly translation [158].

2.1.7 *Multilingual Information Retrieval (MLIR)*

Multilingual Information Retrieval (MLIR) is another sub-field of IR that processes resources in multiple languages, i.e., handles resource retrieval when either documents, queries, or both are in multiple languages [149], [158]. Fluhr, Frederking, Oard, *et al.* [65] define MLIR as a system that can process a query and return results/documents in essentially any language. MLIR also employs CLIR techniques to overcome language barriers from either the query or the document collection [149].

Unlike CLIR, the final MLIR results are typically displayed in the resource collections' original languages [158]. Thus, the query is translated to match the languages of resource collections. For example, if the MLIR system supports three languages (including that of the query), the query must be translated to two languages apart from that of the query. This thesis focuses on MLIR, defined as a user querying in one language and receiving results in multiple languages supported by the system, in this case English and Kiswahili.

2.2 BRIEF HISTORY OF MULTILINGUAL INFORMATION RETRIEVAL

2.2.1 *Motivation*

There were a few websites, mostly written entirely in English, when Tim Berners-Lee invented the WWW (or Web) in 1989. As a result, early search systems such

as AltaVista and Yahoo! implemented systems that met the needs of the English-speaking community [158]. However, it has never been the same for the past 30 years; the Web has been used all over the world, supporting documents in almost every written and spoken language. Table 2.1 summarizes the global Internet user population by continent.

The number of users by language has also increased dramatically, most notably in languages other than English, such as Arabic, which has increased by 9,348% in the last 20 years, as shown in Table 2.2. As a result, Web content in languages other than English is increasing, as evidenced by Wikipedia documents [220].

Table 2.1: Estimates of the world Internet usage and population statistics as of December 2020.

Region	**Internet Users (%)**	**Growth 2000-'21**	**Internet World (%)**
Africa	590M (43.0%)	12,975%	11.7%
Asia	2,707M (62.6%)	2,268%	53.6%
Europe	728M (87.1%)	592%	14.4%
Latin America/Caribbean	477M (72.4%)	2,544%	9.4%
Middle East	188M (70.8%)	5,627%	3.7%
North America	333M (89.9%)	208%	6.6%
Oceania/Australia	29M (67.4%)	284%	0.6%
Total	**5,053,911,722 (64.2%)**	**1,300%**	**100.0%**

Source: Internet World Stats [193].

The increased number of non-English documents on the Web, as well as the ease with which cultures can interact, are challenging the position of traditional IR [65], [149]. Search engines are no longer just for English speakers or users who are proficient in understanding and expressing their queries to the system in English. In order to meet such demands, the two IR sub-fields of CLIR and MLIR gained traction and popularity in the mid-1990s [29], [152]. However, the first use of CLIR appeared in the early 1970s in Gerard Salton's early works, using a multilingual thesaurus [175].

2.2.2 *Efforts to Facilitate and Enable the Development of CLIR and MLIR*

Because of the importance of CLIR and MLIR to the world's large population, there have been systemic initiatives and efforts to support the growth of these two sub-fields, such as:

1. *Introduction of standards and technologies*. Several standards have been established to support computer processing of text in languages other than English. The International Organization for Standardization (ISO) introduced

Table 2.2: The top 10 Web languages as of March 31, 2020 (number of Internet users by language)

Language	**Population (2020 est.)**	**Internet Users**	**Growth (2000-2020)**	**Internet Users (%)**
English	1,531M	1,186M	742.9%	25.9%
Chinese	1,477M	888M	2,650.4%	19.4%
Spanish	517M	364M	1,511.0%	7.9%
Arabic	448M	237M	9,348.0%	5.2%
Portuguese	291M	172M	2,167.0%	3.7%
Indon./Malay.	306M	198M	3,356.0%	4.3%
French	432M	152M	1,164.6%	3.3%
Japanese	126M	119M	152.0%	2.6%
Russian	146M	116M	3,653.4%	2.5%
German	99M	93M	236.2%	2.0%
Top 10 Lang.	**5,274M**	**3,525M**	**1,188.2%**	**76.9%**
Other Lang.	2,523M	1,061M	1,114.1%	23.1%
Total	**7,797M**	**4,586M**	**1,170.3%**	**100.0%**

Source: Internet World Stats [194].

the ISO 5964:1985 standard in 1985 to support multilingual thesauri [89]. It was later revised to allow for interoperability with other vocabularies, [90]. Another development was the formation of the Unicode Consortium, which resulted in the publication of Unicode Version 1.0 in 1991[2] [158]. Furthermore, through the ISO-639, the ISO developed a scheme for classification of all human languages and their related dialects.

2. *Availability of publicly funded research*. Several projects have been implemented to allow access to data in languages other than English. In the United States, for example, National Science Foundation (NSF) [3] and the Defense Advanced Research Projects Agency (DARPA)[4] have funded cross- and multilanguage research. The European Commission (EC) funds research to promote and support the use of tools and information sharing in European languages. In India, the Technology Development for Indian Languages (TDIL)[5] supports tools that enable the sharing of information and resources in Indian languages. In South Africa, the South African Centre for Digital

2 https://unicode.org/history/
3 https://www.nsf.gov/
4 https://www.darpa.mil/
5 https://tdil.meity.gov.in/

Language Resources (SADiLaR)[6] supports all the official languages of South Africa in terms of research and development.

3. *Introduction of multilingual and cross-lingual tracks in the conferences and evaluation forums and campaigns*. In 1996 the Association for Computing Machinery (ACM)-Special Interest Group in Information Retrieval (SIGIR)[7] hosted the first CLIR workshop [158], which stimulated evaluation campaigns such as Text REtrieval Conference (TREC)[8], which introduced multilingual and cross-lingual IR tracks. The TREC multilingual and cross-lingual track prompted the creation of other evaluation campaigns, such as the Conference and Labs of the Evaluation Forum (CLEF)[9] for European languages, Forum for Information Retrieval Evaluation (FIRE)[10] for Indian languages, and the National Institute of Informatics Text Collection for IR (NTCIR) for Asian languages.

Despite these efforts, the success of CLIR and MLIR remains elusive, particularly for low-resource languages. Part of the reason is a lack of public funding, a lack of evaluation campaigns, and the difficulties in achieving perfect translation, which is required to achieve MLIR and CLIR [149], [166].

Because it is difficult to achieve good topical relevance in MLIR for low-resource languages, perhaps searchers' behaviour can be used in results ranking [32], [153]. This study addresses the challenge in part by utilizing language preferences, which are a common search behaviour among polyglot Web searchers, as seen in Section 2.5.

2.3 ARCHITECTURE OF MULTILINGUAL INFORMATION RETRIEVAL

Approaches for developing MLIR systems can broadly be categorized into two: centralized architecture and distributed architecture [116].

2.3.1 *Centralized MLIR Architecture*

This method combines all of the supported document collections into a single centralized collection known as a mixed document corpus (a pool of documents in different languages), where the original query is translated into all of the supported languages [149]. For a centralized MLIR framework, see Figure 2.1.

The centralized approach proposed by Nie and Jin [149] combines all of the translated queries, including the original query, to form a single large query with a language tag associated with each term (stem). Each document in the mixed document corpus, on the other hand, is tagged with its original language. Because the mixed document corpus has a single central index, the retrieval process is identical to that of a traditional monolingual IR.

6 https://sadilar.org/index.php/en/
7 http://sigir.org/
8 https://trec.nist.gov/
9 http://www.clef-initiative.eu/
10 http://fire.irsi.res.in/fire/2020/home

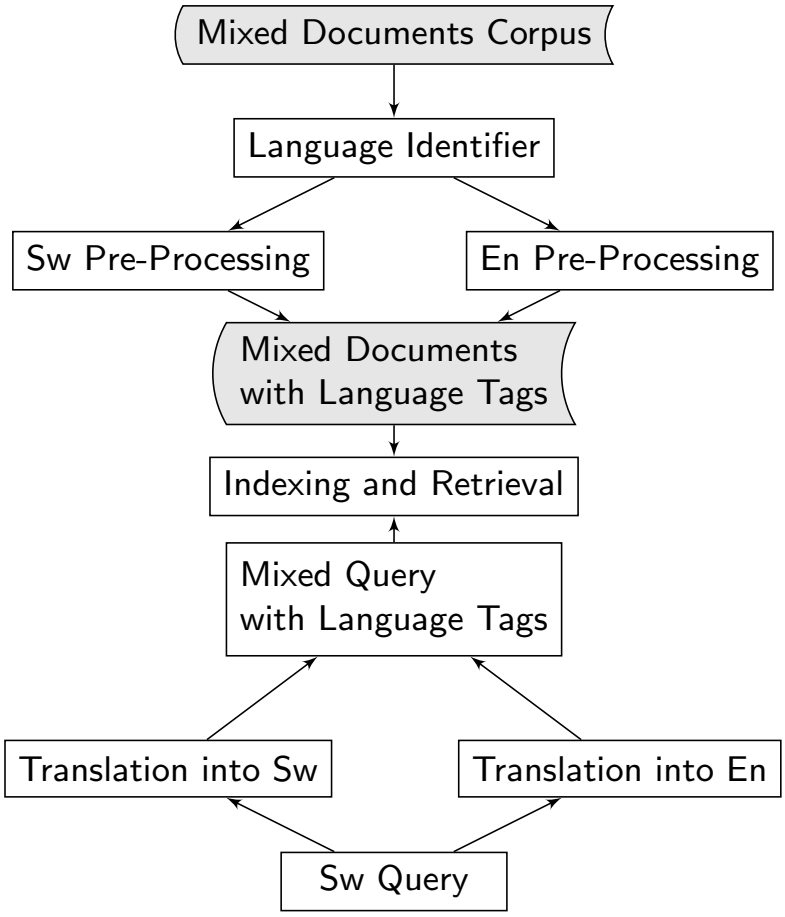


Figure 2.1: A centralized MLIR framework that supports English and Kiswahili.

This centralized MLIR architecture is adapted from Nie and Jin [149] to illustrate a case of two supported languages – English (En) and Kiswahili (Sw).

The centralized index has several advantages, the most important of which is the avoidance of different IR models, which would otherwise be required for each document collection; as a result, term weights are comparable [116], [149]. The use of the same index in this architecture avoids an important and essentially difficult problem in MLIR – result merging.

It should be noted, however, that the centralized architecture suffers from index term over-weighting [116]. This is primarily due to the *tf-idf* scheme, in which the query term is determined by the number of documents in which it appears; thus, combining all of the document collections across the supported languages increases the number of documents while individual terms may not. As a result, the `idf` of a term increases, over-weighting the indexed terms from a high resource language.

Consider the $\approx 60,000$ Swahili Wikipedia article collection combined with the $\approx 6,000,000$ English article collection [220]. The new total number of documents (N) increases by about 100 times in the combined collection and disproportionately increases the weights of the English index terms, which are more likely to have a large `df`.

2.3.2 *Distributed MLIR Architecture*

The distributed MLIR architecture is primarily composed of three steps: i) Query Translation (QT); ii) document retrieval; and iii) result merging [208]. After translating the query and running it against the separate document collections, which essentially have separate indices for each collection for each language, several results are produced, as shown in a simplified distributed MLIR framework in Figure 2.2. Several result lists, however, come at a cost: the challenge of determining how to best combine (merge) these result lists into the most relevant single list.

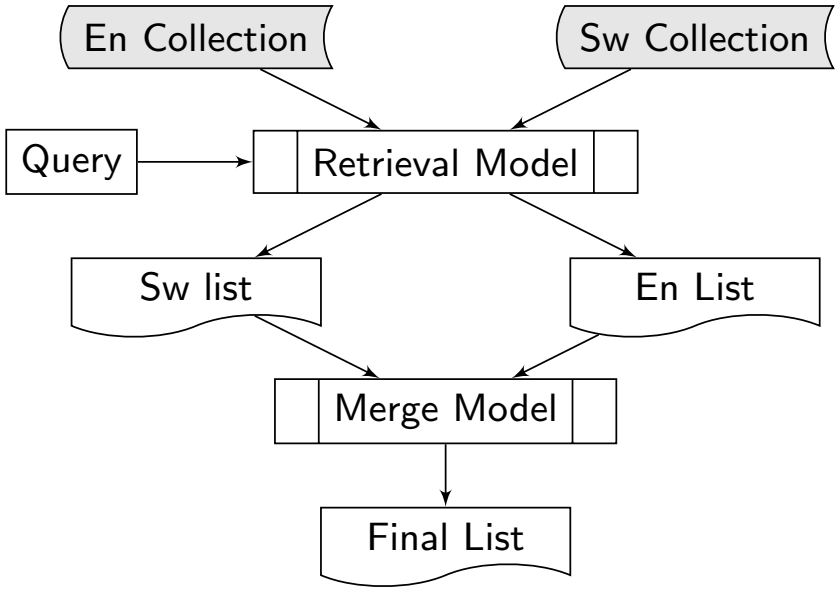


Figure 2.2: A distributed MLIR framework that supports English and Kiswahili.
This distributed MLIR architecture is adapted from Tsai, Wang, and Chen [208] to demonstrate a case of two supported languages – English (En) and Kiswahili (Sw).

The appropriate query or document translation is a significant challenge associated with this architecture [166]. If the translation is flawless, the *Retrieval Model* can produce relevant results in the independent monolingual result lists. The only remaining issue is determining the best *Merge Model* to generate the final multilingual result list. The merging problem is difficult because each result list to be merged has incomparable document scores, owing to the various statistics used to score the documents [149].

This architecture is the focus of this thesis. However, the processes and approaches used to generate the *Retrieval Model* in Figure 2.2 are beyond the scope of this thesis because the focus is on the final merged multilingual result list rather than how to generate those individual result lists.

The pre-trained deep neural network language models such as Multilingual Bidirectional Encoder Representations from Transformers (MBERT) [58] and Cross-lingual Language Model (XLM) [23] helps with a variety of tasks in multilingual IR and NLP. Studies that have used it have shown a significant improvement over the state-of-the-art. However, for the purpose of centralized MLIR, this enhancement is limited to the "retrieval model," not the merge model. Furthermore, fine-

tuning such models for low-resource languages remains a challenge, so other approaches other than language models should be considered.

2.4 INFORMATION RETRIEVAL EVALUATION MEASURES

Evaluation measures quantify the system's effectiveness and ability to *successfully* deliver satisfying information to the user based on his/her information needs. However, as Clough and Sanderson [37] put it, the question of success is difficult and multifaceted:

> *... But what does it mean to be successful? It might refer to whether an information retrieval system retrieves relevant (compared with non-relevant) documents; how quickly results are returned; how well the system supports users' interactions; whether users are satisfied with the results; how easily users can use the system; whether the system helps users carry out their tasks and fulfil their information needs; whether the system impacts on the wider environment; how reliable the system is etc.*

An IR and MLIR evaluation process traditionally consists of three components [132]:

1. a corpus of information needs expressed as queries;
2. a corpus of documents from which information needs must be met;
3. and a set of explicit relevance judgements for each query-document pair, indicating *relevance* or *non-relevance*.

There are two common types of evaluation metrics in IR. The first are those that measure performance based on binary relevance judgements, such as Precision, Recall, Average Precision (AP), Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP), and Mean Reciprocal Rank (MRR). The second type is those that assess performance based on multiple levels of relevance (or graded relevance), such as Discounted Cumulative Gain (DCG), and Normalized Discounted Cumulative Gain (NDCG), and Expected Reciprocal Rank (ERR).

2.4.1 *Precision-Based Measures*

2.4.1.1 *Precision and Recall*

The Precision P is the fraction of retrieved documents that are relevant [229], which is calculated:

$$P = \frac{\text{relevant retrieved documents}}{\text{retrieved documents}} \tag{2.1}$$

Recall R, on the other hand, denotes the system's ability to locate relevant documents [25], and is calculated as

$$R = \frac{\text{relevant documents retrieved}}{\text{relevant documents}} \tag{2.2}$$

The nature of most IR tasks such as Web search requires a system that can retrieve a few relevant results at the top of the list, i.e., one that aims for a high precision rate. Recall, on the other hand, insists on retrieving more relevant documents from the collection, regardless of where they appear in the final results list. As a result, it is not commonly used as a measure of performance in IR tasks such as Web search, because a system may have a perfect recall but has retrieved a large number of (all) documents in a collection, for example, 1000. In general, the measure used is determined primarily by the IR task at hand. For example, if a user is looking for a patent or conducting a systematic review and needs all of the relevant documents, the recall measure becomes relevant.

Precision, according to the definition in Equation 2.1, takes into account all retrieved documents. However, in most IR tasks, it can be truncated to only a specific rank of retrieved documents k, which is known as precision at k (P@k). *Precision@k* represents the proportion of relevant top k documents [40]. For instance, if r relevant documents are found at rank k, then

$$P@k = \frac{r}{k} \tag{2.3}$$

Consider a retrieval task for a single query in which the system has retrieved n documents, as shown in Figure 2.3.

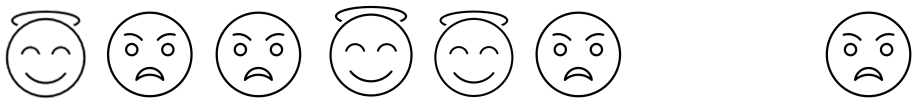


Figure 2.3: An example of a retrieval task in which some of the results retrieved are relevant.

We get the following results when we use Equation 2.3:

$$P@1 = 1/1 = 1 \tag{2.4a}$$
$$P@2 = 1/2 = 0.5 \tag{2.4b}$$
$$P@3 = 1/3 = 0.33 \tag{2.4c}$$
$$P@4 = 2/4 = 0.5 \tag{2.4d}$$
$$P@5 = 3/5 = 0.6 \tag{2.4e}$$
$$P@n = 3/n \tag{2.4f}$$

2.4.1.2 *Average Precision and Mean Average Precision*

The Average Precision (AP) and Mean Average Precision (MAP) are derived from the Precision measure, which is commonly used in IR to evaluate the average performance of the system in which the list is ranked [150]. To obtain the AP, the mean of the precision at each relevant rank position is computed, [229]. In other words, consider each relevant document's precision at a rank position, R, i.e., P@R. When calculating the AP@k, only a subset of the top-ranked documents is considered [40]. It is critical to treat the AP with caution in order to avoid

assuming that it is simply the arithmetic mean of the relevant rank positions, as it carries more significance in terms of recall. That is, the recall base is the denominator of AP. When relevant documents are not retrieved, the recall base division penalizes IR systems that do not retrieve all relevant documents.

Consider our example in Figure 2.3 once more. Consider only precision at ranks 1 (Equation 2.4a), 4 (Equation 2.4d) and 5 (Equation 2.4e) to get the AP@5.

$$AP@5 = \frac{1}{3}(1 + 0.5 + 0.6) = 0.7 \tag{2.5}$$

An AP score represents the performance for a single query. An arithmetic average of APs for multiple queries n or retrieval models yields the Mean Average Precision (MAP) [14]. Formally,

$$MAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{2.6}$$

2.4.1.3 *Reciprocal Rank and Mean Reciprocal Rank*

The Reciprocal Rank (RR) measure computes the reciprocal of the position at which the first relevant document R was found, i.e., $RR = R^{-1}$ [40]. If a relevant document is found at position 1 in the ranked list, the Reciprocal Rank (RR) is 1; if it is found at position 2, the RR is 0.5, and so on. It is especially useful when a user only wants to retrieve one of the most relevant documents (known item), such as in question answering retrieval applications [40], [215]. While AP considers the whole ranking, RR accounts just for the first relevant document.

Referring back to our previous example in Figure 2.3, because the relevant document is ranked first, the RR is 1. The (arithmetic) mean of multiple queries yields a metric known as the Mean Reciprocal Rank (MRR). The MRR is equivalent to the MAP when each query has exactly one relevant document.

2.4.1.4 *R-Precision*

R-Precision is the ratio of relevant documents r retrieved until the position that equals the number of relevant documents for a query in the collection R i.e., $R - Precision = \frac{r}{R}$ [9], [40]. Assume a collection contains 150 documents, 25 of which are relevant to query Q ($R = 25$), and the system retrieves 8 relevant documents ($r = 8$). The precision@25 is then R-Precision, which is $\frac{8}{25} = 0.32$.

R-Precision is another measure that is strongly related to AP [9]. However, unlike other precision-based measures, it requires a thorough understanding of the relevant set of documents (R) for a specific query [40]. The R-Precision algorithm addresses the Precision@k problem, in which the total number of relevant documents in the collection affects the final score. For example, a perfect system with only 5 relevant documents that retrieves all 5 documents results in a $P@15 = \frac{5}{15} = 0.33$, whereas the $R - Precision = \frac{5}{5} = 1$. This implies that the R-Precision is appropriate for evaluating the overall system's effectiveness. It does not, however, place as much emphasis on fine-grained ranking quality as other precision-based measures such as the P@k [40].

To calculate the Average R-Precision (ARP), take the arithmetic average of the R-Precision values from multiple topics/questions.

2.4.2 *Gain-Based Measures*

2.4.2.1 *DCG and NDCG*

Precision-based measures, such as those discussed above, are best suited when there are binary judgments, such as relevant and irrelevant. When there are multiple degree or ordered relevance judgements, a measure called Discounted Cumulative Gain (DCG) becomes appropriate [94], [150].

Each relevance level is assigned a gain value; for example, 3 can represent highly relevant (hr), 2 can represent relevant (r), 1 can represent marginally relevant (mr), and 0 can represent completely irrelevant (ci). The gain vector is then created from the list of documents that were retrieved. For example, if 5 documents are retrieved with gain values of r, r, hr, ci, mr, the gain vector $G = 2, 2, 3, 0, 1 >$ (*CASE I*). The cumulative gain (CG) is expressed mathematically as:

$$CG_{@k} = \sum_{i=1}^{k} G[i] \tag{2.7}$$

In our case, $CG = 2+2+3+0+1 = 8$. Assume another set with $G = 3, 2, 1, 2, 0$ (*CASE II*); the $CG = 3+2+1+2+0 = 8$. Both sets are equally good in terms of CG, with the same CG score. However, the second set is superior to the first because it places the most relevant result at the top of the list.

When the position of the document is used in conjunction with the score of a document, the problem can be solved. Documents ranked lower in the list are penalized using a discount function, assuming that users are typically uninterested in documents ranked lower in the list [22], [94].

The Discounted Cumulative Gain (DCG) is

$$DCG_{@k} = \sum_{i=1}^{k} \frac{G[i]}{\log_2(i+1)} \tag{2.8}$$

The log base is usually set to 2, but it can also take other values e.g., 10, depending on the task. While the DCG for first set is,

$$\begin{aligned} DCG &= \frac{2}{\log_2(1+1)} + \frac{2}{\log_2(2+1)} + \frac{3}{\log_2(3+1)} + \frac{0}{\log_2(4+1)} + \frac{1}{\log_2(5+1)} \\ &= \frac{2}{1} + \frac{2}{1.585} + \frac{3}{2} + \frac{0}{2.322} + \frac{1}{2.585} \\ &= 2 + 1.262 + 1.5 + 0 + 0.387 \\ &= 5.15 \end{aligned}$$

the second set produces,

$$\begin{aligned} DCG &= \frac{3}{\log_2(1+1)} + \frac{2}{\log_2(2+1)} + \frac{1}{\log_2(3+1)} + \frac{2}{\log_2(4+1)} + \frac{0}{\log_2(5+1)} \\ &= \frac{3}{1} + \frac{2}{1.585} + \frac{1}{2} + \frac{2}{2.322} + \frac{0}{2.585} \\ &= 3 + 1.262 + 0.5 + 0.86 + 0 \\ &= 5.62 \end{aligned}$$

To calculate the Normalized Discounted Cumulative Gain (NDCG), first normalize the DCG vector against the "ideal" DCG vector (IDCG). The NDCG establishes appropriate lower and upper bounds [0,1]. Formally,

$$NDCG_{@k} = \frac{DCG_k}{IDCG_k} \tag{2.9}$$

Consider the two examples above; the IDCG is calculated by assuming that all relevant documents are ideally at the top of the document list, i.e., $G = 3, 2, 2, 1, 0 >$. Thus,

$$\begin{aligned} IDCG &= \frac{3}{\log_2(1+1)} + \frac{2}{\log_2(2+1)} + \frac{2}{\log_2(3+1)} + \frac{1}{\log_2(4+1)} + \\ &\frac{0}{\log_2(5+1)} \\ &= \frac{3}{1} + \frac{2}{1.585} + \frac{2}{2} + \frac{1}{2.322} + \frac{0}{2.585} \\ &= 3 + 1.262 + 1 + 0.431 + 0 \\ &= 5.69 \end{aligned}$$

As a result, the NDCG for our first case is:

$$NDCG = \frac{5.15}{5.69} = 0.91$$

And now for the second case:

$$NDCG = \frac{5.62}{5.69} = 0.99$$

Using the NDCG measure highlights the distinction between the two systems (cases). Apart from graded relevance, NDCG (and any other graded measures) can also be used for binary relevance, for example, in experiments involving click-through data.

2.4.2.2 *Expected Reciprocal Rank*

The Expected Reciprocal Rank (ERR) [27] is another measure with a discounting spirit. It discounts the number of documents displayed below the highly relevant documents by defining the expected reciprocal length of time that the user will take to find a relevant document [27], [150]. Chapelle, Metlzer, Zhang, *et al.* [27] contends that NDCG's assumptions are static, i.e., the gain is the same for a document in a rank position, and the discount affects the document in the rank above it equally. The ERR metric is based on the Cascade Model (CM) [27]. The metric is appropriate for situations where a user is only interested in a few highly relevant documents, such as navigational search [173].

The ERR is calculated as follows:

$$\sum_{k=1}^{n} \frac{1}{k} P(\text{stop at position } k) \tag{2.10}$$

where n represents the total number of documents in the ranked list. The likelihood that a user will stop at position k is determined by the likelihood R_i that a document will satisfy the user. As a result,

$$P(\text{stop at position } k) = \prod_{i=1}^{r-1}(1 - R_i)R_k \tag{2.11}$$

Assume g_k is the *k-th* document's grade; we want to map from relevance grades to probability of relevance, based on Chapelle, Metlzer, Zhang, *et al.* [27]

$$P(\text{satisfied by doc } k) = \frac{2^{g_k} - 1}{2^{g_{max}}} \tag{2.12}$$

where g_{max} is the highest grade/point on the scale used; in our CASE I example, this is 3. Refer to Table 2.3 for details on an ERR calculation of CASE I.

Table 2.3: Example – calculating the Expected Reciprocal Rank (ERR).

k	k^{-1}	g_k	P(satis. by doc k)	P(stop at doc k)
1	$\frac{1}{1}$	2	$\frac{2^2-1}{2^3} = 3/8$	$3/8$
2	$\frac{1}{2}$	2	$\frac{2^2-1}{2^3} = 3/8$	$3/8 * (1 - 3/8)$
3	$\frac{1}{3}$	3	$\frac{2^3-1}{2^3} = 7/8$	$7/8 * (1 - 3/8) * (1 - 3/8)$
4	$\frac{1}{4}$	0	$\frac{2^0-1}{2^3} = 0/8$	$0 * (1 - 3/8) * (1 - 3/8) * (1 - 7/8)$
5	$\frac{1}{5}$	1	$\frac{2^1-1}{2^3} = 1/8$	$1/8 * (1 - 3/8) * (1 - 3/8) * (1 - 7/8) * (1 - 0/8)$

k represents a document's position/rank, g_k represents the assumed grade of the k^{th} document, P(satis. by doc k) represents the probability of a user being satisfied by a document at k, and P(stop at doc k) represents the probability of the user stopping at k after being satisfied by the document.

The final ERR score can now be easily calculated with Equation 2.10 as follows:

$$\begin{aligned} ERR = {} & \frac{1}{1} * 3/8 + \frac{1}{2} * 3/8 * (1 - 3/8) + \frac{1}{3} * 7/8 * (1 - 3/8) * (1 - 3/8) + \\ & \frac{1}{4} * 0 * (1 - 3/8) * (1 - 3/8) * (1 - 7/8) + \\ & \frac{1}{5} * 1/8 * (1 - 3/8) * (1 - 3/8) * (1 - 7/8) * (1 - 0/8) \\ & = 0.375 + 0.117 + 0.114 + 0 + 0.001 \\ & = 0.61 \end{aligned}$$

B: Literature Review

This section contains a review of works that are related to the works presented in this thesis. We begin by looking at studies that investigated information needs and search behaviour without regard for whether or not the focus was on Web users. We then focus on research on information needs and search behavior in the

Swahili-speaking region. The second part of this section delves into Web users' search behaviour, with a focus on their code-switching habits. Because there were few works on multilingual search engines, we examined a number of works that used monolingual search engine system users. The final part of this section looks at methods for merging results in MLIR. That is, we examine various works that proposed methods for combining results from various sources to produce a single results list.

2.5 INFORMATION NEEDS AND SEARCH BEHAVIOUR OF USERS

2.5.1 *Information Needs and Search Behaviour*

Researchers in information science, IR, and information systems, as well as information/resource providers/designers, work hard to understand what users of systems and/or services want, and how they behave when using such systems/services [120], [196]. Researchers and designers make an effort to understand users before designing a system/service, as well as to assess/evaluate it once in use and to deliver more specific information rather than general information.

Thus, studies on information needs and seeking behaviour vary greatly depending on a variety of factors, including the area/specialization under study, such as agriculture or health care, as well as geographic region/location, including rural vs. urban dwellers, as discussed below.

The diary study by Church and Smyth [34] on the information needs and seeking behaviour of mobile users discovered that information needs are primarily location- and time-dependent. They change with geographical location and time, implying that they can be spatial-temporal. Another study on mobile users is that of Aliannejadi, Harvey, Crestani, *et al.* [6], which shows that the context and situation determine mobile users' information needs and seeking behaviour. The study by Kassab and Yuan [104] had similar findings on the information needs and seeking behaviour of mobile and smartphone users.

According to Clarke, Belden, Koopman, *et al.* [35], studies on the information needs and search behaviour of health workers in primary care services show that physicians and nurses are interested in information about diagnoses, drugs, treatments, and/or therapies. Furthermore, the authors report that primary case service workers consider their coworkers to be their primary source of information.

In terms of the Internet (Web) as a source of information, the review by Younger [227] revealed that the information needs of doctors and nurses on the Web are the same, namely patient care (diagnosis, drugs, and treatments) and professional development. The authors also report a lack of interest in and awareness of the importance of libraries as a source of information. Demergazzi, Pastore, Bassani, *et al.* [56] investigated neurologists' information needs and search behaviour. They discovered that neurologists treating patients with multiple sclerosis and migraine are most interested in therapy management, drugs, and diagnostic strategies and procedures information by analyzing query logs.

Islam and Zabed Ahmed [92] examined studies on the information needs and search behaviour of rural residents in both poor and rich countries. Their analysis

revealed that, despite socioeconomic differences between developed and developing countries, information needs are similar. They are interested in information about agriculture, health, education, religion, occupation and income generation, self-governance, current events, and recreation. Furthermore, rural dwellers in developed countries are interested in environmental, legal, and civil rights information. Additionally, rural dwellers in developed countries are also interested in information about the environment, and legal and civil rights. Phiri, Chipeta, and Chawinga [160] discovered that small-holder farmers primarily search for animal husbandry information, and that the primary source of information was colleagues' personal experiences.

Another area where several studies on users' information needs and search behaviour exist is music and entertainment. According to Lee and Downie [112], music searchers require contextual information (metadata) in addition to the bibliographic information that is normally associated with the music. Their search behaviour is influenced by public knowledge or other people's opinions, such as reviews, recommendations, and ratings. Considering the search behaviour influenced by recommendations, Cheng and Shen [30] proposed a location-aware music recommendation. A study conducted by Deng, Zhao, Fu, *et al.* [57] investigated the behaviour of females and males who ask questions about music on a social Q&A site. They discovered that male and female behaviour is distinct. Females' questions were indirect, promoting discussions about the music they seek, whereas males provided enough contextual information expecting a ready reference.

The list of different sectors/areas with different information needs and search behaviour continues, despite the fact that there is a massive body of research in this area. We summarize the discussed works in Table 2.4, where the information needs and search behaviour are primarily occupation and/or area oriented.

2.5.2 *Information Needs and Search Behaviour in the Swahili-speaking Region*

Work on information needs and search behaviour in the Swahili-speaking region (in East Africa) is limited, with most studies focusing on specific sectors/domains such as agriculture [16], [62], [130], informal sectors [91], health [156], and rural societies [141]. Furthermore, these works are concerned with identifying the various types and sources of information. However, to the best of our knowledge, there have been a few/no studies on Swahili-speaking Web users. The information language was either ignored or assumed to be an obvious variable in the studies.

2.6 CODE-SWITCHING BEHAVIOUR OF POLYGLOTS ON THE WEB

The previous section's discussion of information needs and search behaviour included several sources of information such as word of mouth from a colleague, experiences, libraries, television and radio programs, and the Web (e.g., social media, blogs, and websites). In this section, we will concentrate on the Web as the current primary source of information for many people around the world, particularly those who can read or understand multiple languages – polyglots.

Table 2.4: A summary of research on information needs and search behaviour.

No.	Study	Area	Major Findings
1	Church and Smyth [34]; Aliannejadi, Harvey, Crestani, *et al.* [6]; and Kassab and Yuan [104].	Mobile phone.	Mobile phone users search behaviour is driven by the situation, and context e.g., location and time.
2	Clarke, Belden, Koopman, *et al.* [35]; Younger [227]; and Demergazzi, Pastore, Bassani, *et al.* [56].	Health and Life sciences.	Physicians mainly look for information on diagnosis, drugs, treatment, therapy, procedures and professional development.
3	Islam and Zabed Ahmed [92]; and Phiri, Chipeta, and Chawinga [160].	Rural Dwellers.	Rural dwellers mainly search for agriculture, health, education, religion, occupation and income generation, self-governance, current affairs, recreation, environment, and legal and civil rights.
4	Lee and Downie [112]; Cheng and Shen [30]; and Deng, Zhao, Fu, *et al.* [57].	Music.	Music searchers' behaviours rely on public knowledge or opinions of others, such as reviews, recommendations and ratings.

Language switching – *code-switching*, also known as *language alternation* – is a common behaviour of polyglot users when interacting with IR systems [10]. We wanted to understand the reasons for their code-switching behaviour, thus, this section contains a review of several works that investigated the reasons for such behaviour of polyglots in both classic IR and MLIR settings.

Most Web search engines allow users to display their choices/preferences in terms of interface language, content, layout, themes, and system configurations and customization [32], [119]. According to studies by Chu and Komlodi [32] and Ling, Steichen, and Choulos [119], users switch from one search engine to another based on the popularity of the search engine, usability of the interface, locality, and search quality. These elements may influence user satisfaction with the IR system [78]. The thesis, on the other hand, is interested in the language-switching behaviour of polyglots for their search, rather than the search engine interfaces.

We look at research on the reasons for code-switching. The reasons range from straightforward, such as translation requirements, to complex, such as resource availability. Furthermore, the methodology used in these studies to determine such reasons differs from one another.

Language switching is primarily motivated by the availability and quality of information, according to a controlled laboratory experiment study conducted by Aula and Kellar [10]. The most recent study, conducted by Wang, Komlodi, and Ka [217], used the same setting of controlled laboratory experiments supplemented by interviews. They classified code-switching into two types: situational

code-switching and metaphorical code-switching. The findings on the reasons for situational code-switching are consistent with the findings of Aula and Kellar and other studies discussed below. Language proficiency, information verification, context, and translation purposes were all mentioned in these studies.

However, the interesting factors for metaphorical code-switching are worth noting because they mostly have to do with the image and perception the searcher has in mind. Information *accuracy and objectivity* perception in one language; *sense of belonging* when using a specific language, such as mother tongue; *credibility and user trust* of the website; and *psychological* reasons are examples of such factors.

Lowe and Steichen [129] 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, in addition to the findings by Wang, Komlodi, and Ka [217] about sense of belonging for native language. Using an online questionnaire, Vassilakaki, Garoufallou, Johnson, *et al.* [210] discovered that even when there is insufficient/limited information on the Web, users always prioritize their mother language. Users who are having difficulty with their foreign language proficiency, particularly the query formulation problem for non-native speakers, prioritize their native language Nzomo, Ajiferuke, Vaughan, *et al.* [152].

Other studies, such as Berendt and Kralisch [15], Petrelli, Levin, Beaulieu, *et al.* [159], and Rieh and Rieh [170], reported contradicting findings in which users demonstrated a preference for accepting information in English over their native languages as long as it is appropriate for completing the task.

The study by Vassilakaki, Garoufallou, Johnson, *et al.* [210] also reported that the purpose of the information a user is looking for determines the language to use at a given time. Marlow, Clough, Recuero, *et al.* [133] discovered that language skills influence the searching experience in a multilingual search. Wang and Komlodi [216] established several reasons using diary interviews, including the need for translation, the availability of resources, language proficiency, and the context of the information sought, such as news and entertainment and social networking.

Clough and Eleta [36] used a questionnaire to examine if two specific factors for language choice – language skills and the user's field of knowledge – significantly correlate, varying between different fields of knowledge. In a survey of polyglots' browsing and search behaviour in multilingual search engines, Steichen, Ghorab, O'Connor, *et al.* [196], including the follow-up works (Steichen and Freund [195], Ling, Steichen, and Choulos [119], and Steichen and Lowe [197]), revealed that the context of search, such as usage purpose of the information sought and topic domain, and language fluency have a significant influence on the choice of language for daily browsing and searching.

A recent study by Ling, Steichen, and Figueira [120] used a crowd-sourcing approach to investigate user behaviour regarding multilingual news consumption. They discovered that the search language is determined by the news topic domain.

In conclusion, these studies revealed the factors influencing code-switching behaviour in information search, as stated in Table 2.5. An intriguing and significant observation is the code-switching caused by the topic domain (search topic),

which has been revealed in several studies, such as Steichen, Ghorab, O'Connor, *et al.* [196], Steichen and Freund [195], Lowe and Steichen [129], Wang, Komlodi, and Ka [217], Ling, Steichen, and Figueira [120], and Steichen and Lowe [197].

Table 2.5: A Summary of code-switching factors in polyglots.

Study	Methodology	Setting	Factor
[36]	Questionnaire	IR	Language skills and proficiency
[196]	Survey		
[216]	Interviews		
[197]	Crowdsourcing	MLIR	
[36]	Questionnaire	IR	Knowledge and Profession
[202]	Interviews		
[190]	Controlled lab	MLIR	
[109]	Log mining	IR	Resources availability
[10]	Controlled lab		
[202]	Interviews		
[216]	Interviews		
[152]	Survey	IR	Query formulation challenges
[190]	Controlled lab	MLIR	
[210]	Questionnaire	IR	Search context
[216]	Interviews		
[196]	Survey	IR	Topic domain
[129]	Survey-based		
[217]	Controlled lab and Interview		
[195]	Controlled lab	MLIR	
[197]	Crowdsourcing		
[120]	Crowd-sourcing		

... Continued on next page

Table 2.5 – continued from previous page

Study	Methodology	Setting	Factor
[217]	Controlled lab and Interview	IR	Translation purposes Information verification
[196]	Survey	IR	Task type
[170]			
[159]	Controlled lab	CLIR	
[197]	Crowdsourcing	MLIR	
[10]	Controlled lab	IR	Information quality and accuracy
[109]	Log mining		
[210]	Questionnaire	IR	Beliefs, credibility and user trust
[129]	Survey-based		

To the best of our knowledge, only a few studies have examined query and click-through logs to identify, implicitly, the factors influencing code switching. Query and click-through logs are now the cheapest source of large amounts of data and may avoid the problems associated with survey-based data. Survey-based human behaviour studies are expensive for large-scale data collection; they cannot scale for a large geographical region; and they are static, as they represent human behaviour at a specific point in time [143]. Furthermore, Vigo, Matentzoglu, Jay, *et al.* [212] warn that survey-based studies are unreliable due to self-reporting biases.

2.7 RESULTS MERGING IN MULTILINGUAL INFORMATION RETRIEVAL

This section describes the most common approaches to the MLIR merging task, which is required to achieve final results in a distributed architecture. While some are based on heuristics (*traditional merging* approaches), others are based on Machine Learning (ML). Before delving into the merging approaches for MLIR, let us first define a few terms.

The term *Ranking*, as used in IR, CLIR, and MLIR, refers to the arrangement of objects in a specific order, typically in the order of relevance, importance, or preference to the user's query [11]. To achieve the required/likely relevance [113], features of the objects and/or user preferences are used.

Before being presented to a user, *re-ranking* should extract and re-order a sample of the original ordered results [123]. The top N results are retrieved and then run through a model to produce fewer results (top n), where $N >> n$.

Merging (or aggregation) is the process of combining the results of two or more ranked lists into a single list [81]. At this stage, results re-ranking can also be performed to reduce the total number of results presented to the user. Each of the returned result lists has a score or rank position [81], [116]. Some traditional approaches ignore this score when matching to produce the merged list, while others map these scores to a comparable level.

2.7.1 *Traditional Merging and/or Re-ranking Approaches*

2.7.1.1 *Round-Robin Merging*

The Round-Robin (R-R) merging approach [181], [214] assumes that the intermediary result lists have similar ranking approaches, that each list has roughly the same number of relevant documents, and that the distribution of relevant documents across the lists is similar [116]. Thus, the merged list is obtained by interleaving a single result from each of the intermediate result lists until all of the intermediate result lists are exhausted. However, assuming that a low- and high-resource language pair, such as Kiswahili and English, have the same number of relevant documents is incorrect. Lin and Chen [116] reports that the number of relevant documents varies depending on the document collection.

2.7.1.2 *Raw-Score Merging*

The raw-score merging method [181] assumes comparable similarity scores across individual result lists [81]. Aside from the merged result lists, each document in the list must have a similarity score [157]. Based on the raw similarity scores of the documents involved, the system can generate the final merged list. However, this assumption is flawed because MLIR's intermediary result lists are generated in different languages, which means that the similarity score may not be comparable and, as a result, inaccurate [24].

2.7.1.3 *Normalized Score Merging*

The normalized score merging approaches address the assumption in raw-score merging, which assumes comparable scores across the result lists, as well as the problem of skewed distribution of relevant documents in the R-R approach. Prior to merging, the normalized score approaches normalize the scores in each individual list. Because the algorithms use scores, each document in the result list must have a relevance score in order to produce the final scores of each document to be merged [158].

The *normalized-by-top1* [162], [179] approach divides each document's score value by the highest score in the list. The model organizes the final result list based on the new normalized scores, which range from 0 to 1. The main issue with this approach is when there is a large difference in score between the top and second ranked documents on the list. The model normally penalizes the second document, reducing its score; the same is true for the scores of all other documents in their positions. Even if their original scores were low, the model

may rank these documents lower in the combined list if the second list did not have significant differences in scores.

The *normalized-by-min-max* merging method differs slightly from the previous method in that it uses minimum and maximum document scores. The *CombRSVnorm*, proposed by Savoy [180], is one approach that employs this style. Each document in the final result list has been assigned a normalized score of:

$$S_{\overline{D}} = \frac{S_D - S_{D_{min}}}{S_{D_{max}} - S_{D_{min}}} \tag{2.13}$$

Where $S_{D_{min}}$ is the minimum or a score at some cut-off point for the document to be included in the final result list, $S_{D_{max}}$ is the maximum score, S_D is the document score, and $S_{\overline{D}}$ is the normalized score.

The *normalized-by-top*k merging method [116] uses the same normalization method as the normalized-by-top1 method described above. Normalization, on the other hand, is based on a cut-off point determined by a certain number of documents with the highest scores. Each document's score is divided by the average of the scores from the topk documents. This approach, like the normalized-by-top1 or normalization based on max and min score, combines and sorts the adjusted scores to produce the final merged results. In the row-score merging approach [157], the *normalization* approaches address the problem of incomparable scores.

2.7.1.4 *Weighted-Score Merging*

Normalization approaches produce better results, but they perform poorly when the corpora from which the result lists are drawn have different statistics, resulting in incomparable scores [157]. Consider the term "water" in a corpus of computer science documents; it will rank high if it appears in a query but will rank lower in a corpus of water-related documents. The weighted-score merging approach entails assigning a score to each document in the collection based on its relevance and the corpus to which it belongs [157], [168]. As a result, high-scoring relevant documents from a low-scoring corpus rank lower than relevant high-scoring documents from a high-scoring corpus. The final score of each document is calculated as follows, according to Paltoglou, Salampasis, and Satratzemi [157]:

$$S_{\overline{C}} = \frac{S_C - S_{C_{min}}}{S_{C_{max}} - S_{C_{min}}} \tag{2.14}$$

$$S_{\overline{D}} = \frac{S_D - S_{D_{min}}}{S_{D_{max}} - S_{D_{min}}} \tag{2.15}$$

$$\overline{\overline{D}} = \frac{S_{\overline{D}} + 0.4 * S_{\overline{D}} * S_{\overline{C}}}{1.4} \tag{2.16}$$

Where $S_{\overline{C}}$ is the normalized corpus score, S_C is the corpus score, and $S_{C_{min}}$ and $S_{C_{max}}$ are the minimum and maximum scores assigned to any corpus, respectively. $S_{\overline{D}}$ is the normalized document score, S_D is the respective document

score, and $S_{D_{min}}$ and $S_{D_{max}}$ are the minimum and maximum document scores in a corpus, respectively. Finally, the final score of a document is $\overline{\overline{D}}$. Thus, Equations 2.14 and 2.15 normalize the corpus and document scores, respectively, whereas Equation 2.16 computes the final relevance score for each document, normalized between 0 and 1 by dividing by 1.4 [157].

2.7.1.5 *Sub-Collection Based Merging*

The approaches based on sub-collection merging make use of the underlying information from the document sub-collections. Braschler and Schäuble [19] compared the scores of the retrieved documents using information from the aligned documents in the individual collections. Document alignment refers to the process of matching documents from different languages in a mixed document corpus based on their similarity. The authors intended to address the issue of incomparable document scores. The authors reported comparable performance to other submissions for the TREC-7 task on multilingual collection of German, French, English and Italian.

Another related work is Lin and Chen [117], in which the merging process took into account translation qualities as well as the characteristics of individual sub-collections. That is, if a collection, for example, lacks relevant documents for a given query, its documents should not appear in the final ranked list. As a result, this approach proposed using collections weights and a translation penalty. The authors demonstrate that their previous *normalized-by-top-k* method, now with translation penalty and collection weight, outperformed other approaches such as R-R and normalized score approaches. It did, however, perform equally well with the raw-score approach.

Martínez-Santiago, Urena-López, and Martín-Valdivia's [136] method, known as 2-step RSV, groups a given query term and its corresponding translations with their respective document frequencies. As a result, document scores are re-calculated to reflect the new document frequencies for each query term. Because re-indexing all of the collections would be costly, the authors proposed two steps. First, run a traditional CLIR-like process in which a query is translated to match the supported documents and the results are retrieved from each collection. Second, re-indexing the document collections while only taking into account the newly generated concepts, which are terms and their corresponding translations. The co-created query resulted in a new query. The authors reported excellent results when compared to other MLIR merging approaches such as R-R, logistic regression, and normalized scores.

The approach taken by Nie and Jin's [149] for producing the final merged results list differs from that of the majority of studies in MLIR merging. They proposed a centralized architecture in which all of the documents in the targeted collections are combined into a single corpus. The language tag is then assigned to each document in the mixed documents collection, and both the original and translated queries are tagged with their respective languages. As a result of the mixed document corpus having a single central index, the retrieval process can proceed normally, as in classic IR. The authors modified the R-R and raw score approaches to include language tags, and their results showed that their central-

ized/tagged approach outperformed the original raw score and R-R approaches by a small margin.

2.7.2 *Other Heuristic Approaches*

The works of Chu and Komlodi [32] and Ling, Steichen, and Choulos [119] propose to display results from individual result lists separately on panels or tabs, avoiding the merging process entirely. These approaches are not particularly interesting because they are concerned with user experience.

2.7.3 *Machine Learning Merging Approaches in MLIR*

When ML approaches are used to solve ranking problems, they are known as Learning-to-Rank (L2R) [123]. Only a few studies, particularly in MLIR, have used ML for result merging, some of which are primarily based on statistical approaches for predicting the probability of a binary outcome (pure ML methods).

Le Calvé and Savoy [111] and Savoy [180] investigated the use of logistic regression in determining the probability of relevance for each document. They considered two factors: the document's score and the logarithm of its rank. The final ranked list is then generated based on each individual document's estimated probability of relevance. Savoy, for example, reports that their approach outperformed the R-R, raw-score, and normalized scores approaches. Si and Callan's [191] work also employs a logistic-based approach, yielding comparable document scores from query and language-specific logistic models. These authors also report better performance compared to heuristics approaches.

The works that typically employ L2R are those of Gao, Niu, Zhou, *et al.* [68], and Tsai, Wang, and Chen [206], [208]. This group of studies use a set of predefined features. The researchers hand-craft the features, which are then used to train the ranking/merging models.

The Gao, Niu, Zhou, *et al.*'s [68] approach to developing the feature list is to create it based on the document's similarities with the search query, where a joint relevance probability is used. The authors compared their proposed approach, where it was shown that it significantly outperformed other approaches such as SVM-MAP and RSVM.

Tsai, Wang, and Chen [208] and Tsai, Chen, and Wang [206] extract named-entity, document length, and the number of query terms. To learn the weights of these features, the authors used a L2R algorithm called FRANK [207]. The learned weights for each feature were used in combination with the BM25 ranking model scores to calculate the final ranking score for each document, then the documents were sorted based on these scores to generate the final ranked list. This approach outperformed other merging strategies it was compared with, such as logistic regression-based, R-R, raw-score, and normalized-by-top1 (or -topk).

A semi-supervised approach based on the multi-view architecture was proposed by Usunier, Amini, and Goutte [209]. The authors consider each language in the collection to be a *view* of a document. They reported that this multi-view approach produced better results than a single-view semi-supervised ranking.

2.7.4 *Machine Learning in MLIR Related Tasks*

ML is widely used in federated search, distributed IR, and/or meta-search for merging/fusing results from various information sources and/or search engines. The Gialampoukidis, Moumtzidou, Tsikrika, *et al.*'s [72] framework fuses media results using multiple modalities such as visual concepts, textual concepts, and visual descriptors. The authors claim that their approach is more effective than bi-modal approaches.

Liu, Meng, Qiu, *et al.* [122] developed the *AllInOneNews*, a meta-search engine connecting over 100 news websites across the world. To determine the rank of each document, the merging algorithm considers several factors, such as the quality of the selected search engine from the system that retrieves the results, the matching terms and the closeness of terms in the query and the title/snippet of the result, the proximity of the query terms in the title/snippet of the result, the order of query terms in the result, and the publication time of the result. Merging using these factors achieved higher effectiveness of AllInOneNews over strong baselines such as Google News and Mamma News.

Qin, Geng, and Liu [163] proposed a probabilistic approach to the ranking aggregation problem. Their probabilistic model, coset-permutation distance based stage-wise (CPS), utilizes other probabilistic models, such as *Luce and Mallows*, to provide an effective ranking model. Kozorovitsky and Kurland [108] also employs a probabilistic approach to fuse or merge document lists. The authors reported that their approach outperformed approaches that use standard retrieval ranks and scores by using inter-document similarities and traditional retrieval ranks and scores.

Rabinovich, Rom, and Kurland [165] based the fused result list on the relevance feedback provided by users. Their empirical evaluation revealed that the relevance feedback-based meta fusion strategy can improve resultant retrieval performance when compared to the use of relevance feedback in single intermediate result lists. Liang, Markov, Ren, *et al.* [114] proposed a method for ranking aggregation called Manifold, which is based on several hypotheses and makes use of inter-document similarities. The authors demonstrate a significant improvement over current fusion strategies.

Given the temporal characteristics, Liang and Rijke [115] suggests fusing microblog search results based not only on the relevance score (ranking information), but also on the publication date/time. The researchers claim that their BurstFuseX model outperforms other time-sensitive fusing strategies in terms of MAP.

In cases of reformulation, search engines will typically suggest several queries. Sheldon, Shokouhi, Szummer, *et al.* [186] suggests a λ-*merge* algorithm for fusing/merging the candidate queries into a single reformulated query. For predicting and selecting the best candidate query and thus improving query reformulation effectiveness, the λ-merge outperformed other supervised methods.

2.7.5 *Comparisons of the Merging Approaches*

Based on the studies reviewed, we can divide the MLIR merging strategies into three sub-groups based on their merging strategies. For a summary of the reviewed works, both traditional and ML-based approaches, see Table 2.6.

The first group employs document comparable scores and/or ranks; merging approaches in this group include round-robin [181], [214], raw-score [181], and normalized score [116], [162], [179], [180]. The merging approaches in this group assume a similar distribution of relevant documents across individual collections or that individual collection scores are comparable [166]. The second group aims to relax the major assumption made by the previous group's approaches by utilizing more latent information from the retrieved result lists or individual sub-collections. This group includes the weighted-score [157], [168], and sub-collection based merging techniques by Braschler and Schäuble [19], Martínez-Santiago, Urena-López, and Martín-Valdivia [136], and Lin and Chen [117]. The third group of approaches employs ML techniques like logistic regression [111], [191], supervised learning based on feature analysis [206], [208], and semi-supervised learning [209].

The primary advantage of traditional approaches is that they can be used in low-resource MLIR settings. ML approaches are typically data hungry, requiring a large amount of data to train, validate, and test. Unfortunately, not all languages have a large number of resources to train on, making heuristic approaches the only option. Some of the algorithms examined (both heuristics and ML-based) were originally designed for data and collection fusion. However, because they are all designed for working with multiple source or document collections, MLIR settings can also use them.

Table 2.6: A Summary of the results merging approaches for MLIR systems.

#	Studies	Approach	Evaluation
1.	Savoy, Le Calvé, and Vrajitoru [181], and Voorhees, Gupta, and Johnson-Laird [214]	Round-Robin (R-R)	
2.	Savoy, Le Calvé, and Vrajitoru [181]	Raw Score	
3.	Powell, French, Callan, *et al.* [162], Lin and Chen [116], Savoy [179], and Savoy [180]	Normalized scores	

... Continued on next page

Table 2.6 – continued from previous page

#	Studies	Approach	Evaluation
4.	Rasolofo, Abbaci, and Savoy [168], and Paltoglou, Salampasis, and Satratzemi [157]	Weighted Score	
5.	Lin and Chen [117]	*normalized-by-top-k* with translation penalty and collection weight	Outperformed R-R and normalized score approaches, but performed equally with raw score approach.
6.	Martínez-Santiago, Urena-López, and Martín-Valdivia [136]	2-step RSV	Outperformed R-R, logistic regression, raw-score and normalized score.
7.	Braschler and Schäuble [19]	Document alignment	Comparable performance with other submissions for TREC-7.
8.	Nie and Jin [149]	Centralized architecture	Slightly better than R-R and raw score.
9.	Le Calvé and Savoy [111], Savoy [180], and Si and Callan [191]	Logistic regression	Generally better than R-R, raw score and normalized approaches.
10.	Gao, Niu, Zhou, *et al.* [68]	Joint probability ranking	Outperformed RSVM and SVM-MAP.
11.	Tsai, Wang, and Chen [208], and Tsai, Chen, and Wang [206]	FRANK [207]	Outperformed 2-step RSV, logistic regression, R-R, raw score and normalization score approaches.
12.	Usunier, Amini, and Goutte [209]	Multi-view	Better performance than approaches that use single-view.

2.8 SUMMARY

This chapter presented key concepts in IR as well as the two stemming sub-fields of CLIR and MLIR to set the stage for further discussion in MLIR. It then provided a brief history and the driving forces that prompted the creation and growth of this MLIR research area. The driving forces examined include: publicly funded projects, specifically sponsoring the development of technologies for multilingual information access and distribution; standards bodies such as the ISO developing technologies and standards to support languages other than English; and facilitation by IR evaluation forums such as TREC and CLEF. The chapter also discussed the two main approaches for developing MLIR architecture: centralized and distributed MLIR architecture.

The background part of this chapter ended with presenting commonly used evaluation measures in IR. It includes popular measures based on binary relevance judgment, such as Precision and MAP, as well as those based on graded relevance, such as the DCG and NDCG.

This chapter's related works covered three topics: general information needs and search behaviour, code switching in Web search, and approaches for MLIR system results merging. The chapter discusses works on general human information needs and search behaviour, whether or not they are related to the Web as a source of information. According to research, an individual's information needs and search behaviour are influenced by their occupation and/or industry. Health professionals, for example, are likely to need information on diagnoses, drugs, therapies, procedures, and treatments.

The thesis then examined works on Web users' search behaviour and preferences, particularly code-switching behaviour. Several studies, mostly survey-based, have identified a number of reasons, including resource availability, search context, query formulation challenges, search topic domain, type of task, searcher language skills and proficiency, and searcher knowledge and profession. These findings were derived primarily from user interactions with monolingual search (IR) systems.

To the best of our knowledge, research on code-switching behaviour (language preferences) among polyglot Web users has been limited to monolingual IR systems, with the majority of findings based on survey data. The few works on language preferences in MLIR systems, mostly survey-based, are part of the reason we undertook this research: to investigate language preferences behaviour via click-logs analysis and use it to improve the relevance of MLIR ranked results.

The chapter examined both traditional and ML-based merging strategies using a distributed MLIR architecture. Traditional merging methods have limitations, most notably the inability to incorporate user preferences in the merging and/or re-ranking of results. These existing strategies fail to address the question of how language preferences can improve search result ranking. This contributes to the thesis's second reason for conducting research.

The following chapter will present a specific literature review on Kiswahili, describing existing solutions such as technologies, tools, and techniques that could aid in the development of a multilingual Swahili IR.

3

A REVIEW OF SWAHILI INFORMATION RETRIEVAL

Information Retrieval (IR) research on Swahili are rare; partly because it is a low resource language in terms of Web documents availability and the number of known language technology tools. However, the language is crucial in serving as a national and/or official language in East and Central Africa (see Figure 3.1), with an estimated population of over 100 million speakers [64], [221], making it the only indigenous African language with the largest number of speakers. Kiswahili, as a language, has the potential to become a regional or "continental" language, as many African countries, particularly those in the Sub-Saharan region, are increasingly adopting it [2], [106]. Thus, there is an ever increasing number of Swahili documents/resources on the Web. These documents have a wide range of diversity due to the uniqueness of the language, which evolved from Bantu languages incorporating Arabic, English, German, and a number of other colonial and pre-colonial commercial languages, such as Persian. Furthermore, as each country where the language is spoken has its own culture and education system – both of which are known to influence language use in Web search.

In this chapter, we are interested in reviewing the Natural Language Processing (NLP) and Machine Translation (MT) technologies, techniques and tools for aiding IR research in Kiswahili, with a particular focus on the solutions for Swahili Cross-Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR).

Figure 3.1: Geographic areas where Kiswahili is widely spoken.

The countries highlighted in deep green are Tanzania, Kenya, Uganda, and (Eastern) Democratic Republic of Congo. The ones in light green are Rwanda, Burundi and some northern parts of Zambia, Malawi and Mozambique. The darker green areas indicate the first language speaking area, which is mostly along the East African coast and some islands in the Indian Ocean – *Source:* Wikipedia.

Natural Language Processing (NLP) is critical to the success of an IR application such as question answering, word sense disambiguation, or information extraction [18], [213]. Machine Translation (MT) is essential for the success of CLIR and MLIR, [158], [226]. As a result, the chapter examines existing (state-of-the-art) Swahili Natural Language Processing (NLP) and MT solutions, such as techniques, technologies, and tools that may enable Swahili IR, particularly multilingual Swahili IR.

There are few surveys on Swahili language technology, particularly its application in NLP and IR systems. As shown in Table 3.1, they primarily used Swahili dictionaries and corpora.

A quantitative survey conducted by De Pauw, De Schryver, and Wagacha [48] examined four Swahili-English bilingual dictionaries. The authors evaluated the dictionaries' coverage against a large monolingual Swahili corpus and investigated their utility in the development of Machine Translation (MT). Previously, Hurskainen [84] conducted a quantitative survey in which five Swahili dictionaries were tested against a Swahili POS parser tool called SWATWOL[1] [83]. A recent survey on bilingual dictionaries by Wójtowicz [223] examined the differences and similarities at a macro structure level. The authors also investigated the trends in Swahili dictionary compilation, advocating for a modern corpus-based approach to dictionary compilation. A survey on Swahili text and speech corpora conducted by Oirere, Deshmukh, Shrishrimal, *et al.* [155] revealed the need for more efforts to improve Swahili speech corpora.

Table 3.1: Existing reviews on Swahili language technologies and tools.

No.	Article	Year	Area
1.	Hurskainen [84]	2004	Monolingual Swahili dictionaries; computational analysis of dictionary coverage.
2.	De Pauw, De Schryver, and Wagacha [48]	2009	Swahili - English dictionaries; evaluation of bilingual dictionary coverage.
3.	Oirere, Deshmukh, Shrishrimal, *et al.* [155]	2012	Swahili text and speech corpora; general review and applicability of such corpora.
4.	Wójtowicz [223]	2016	Bilingual dictionaries; analysis of the similarities and differences of Swahili bilingual dictionaries.

This chapter's review goes beyond dictionaries to cover a broader range of topics. It discusses the most recent advancements in Swahili language technology, tools, NLP, CLIR and MLIR.

1 https://www.sketchengine.eu/swatwol-swahili-part-of-speech-tagset/

The following is the chapter's organizational structure. The general introduction to Swahili language is in Section 3.1. While the literature for Swahili NLP is presented in Section 3.2, the literature for Swahili MT is presented in Section 3.3. In Section 3.4, the literature on Swahili CLIR and MLIR is presented. Finally, Section 3.5 provides a synopsis of this chapter.

3.1 THE SWAHILI LANGUAGE

Kiswahili is a Bantu language, descended from the Benue-Congo branch of the Niger-Congo language family [64], [205]. As a result, many Swahili words are derived from Bantu languages. However, as a result of early interactions between the East African coast and Arabic and Persian-speaking countries, Kiswahili has many Arabic and Persian loanwords [161], [205]. This is why some people refer to Kiswahili as a mixed language [205]. Several other languages, including English, German, and Portuguese, contribute to the vocabulary of Kiswahili [161], [182]. Kiswahili was originally written in Arabic scripts, but due to the influence of German and British colonial administration, it was later changed to Roman scripts in the mid-19th century [161], [182], [205].

Sahel (or *Sawāhil*), an Arabic word for *"of the coast"*, is the source of the word Swahili [13], [205]. At the moment, the term *Swahili* may refer to the people, culture, ethnicity, region, and so on, i.e., Swahili people, Swahili culture, and so on. However, in English, it is most commonly referred to as a language. Kiswahili (*sw*), the Swahili language, is spoken by the Swahili people. Even though Kiswahili is a noun and Swahili is an adjective [205], these two terms may be used interchangeably in this thesis.

Kiswahili has a number of dialects. Standard Swahili (*Kiswahili Sanifu*) is based on the well-known *Kiunguja* dialect. *Kimgao*, *Mtang'ata*, *Pemba*, and *Malindi* are among the other dialects [64], [139].

Kiswahili has 5 vowels and 22 consonant phonemes, according to Ethnologue [64]. The syllable structure is straightforward, with all syllables ending in vowels [144]. A vowel is usually added at the end of loanwords, especially those that end in consonants, such as chalk, which becomes *chaki*. Furthermore, unlike other Bantu languages, the language is not tonal [69].

Kiswahili, like many other Bantu languages, is an agglutinating language, which means that prefixes and suffixes can be added to the root [69], [161]. It employs a nominal class system of 15 noun classes [144] (or 18 noun classes according to Ethnologue [64]). The language also adheres to concordial agreement and has a very complex verb morphology [50], [69], [86]. Kiswahili lacks feminine and masculine gender structures [66]. For example, when translated back to English, he is coming home (*"anakuja nyumbani"*, Sw) becomes "is coming home," losing the gender identity (he). Table 3.2 summarizes a few characteristics that distinguish Kiswahili from English.

While the exact number of Swahili speakers is unknown, there are several accounts from various sources. According to Ethnologue [64], the total number of Swahili speakers is approximately 69 million (16 million as L1, and 53 million as L2). Schadeberg [182] estimated the number of Swahili speakers to be around 75 million about a decade ago. This figure may correspond to the Swahili Wikipedia

Table 3.2: Significant linguistic differences between English and Swahili.

Feature	Swahili Value	English Value	Example
Nominal plurality coding	Plural prefix	Plural suffix	***ch****umba* (room), ***vy****umba* (room**s**).
Adjective and noun order	noun–adjective	adjective–noun	*maji masafi* (clean water, lit. *water clean*).
Numeral and noun order	noun–numeral	numeral–noun	*watoto watatu* (three children, lit. *children three*).
Inflectional morphology	Mainly prefixing	Suffixing	*tulihama* (we moved).
Definite articles	Demonstrative word used as definite article	Demonstrative word differs from definite	*jengo* (a building, the building, building).
Noun phrase conjunction	And is identical to With	And differs from With	*ameleta bata* ***na*** *kisu* (s/he brought duck **and** knife), *Ben alienda* ***na*** *kisu* (Ben went **with** a knife).
Position of interrogative phrases in content questions	Not initial interrogative phrase	Initial interrogative phrase	*unaangalia TV* (you are watching a TV), *unaangalia nini?* (what are you watching? lit. *you are watching what?*).
Negation	Negative form of a verb	Particle or construction	*ninakuja* (I am coming), *siji* (I am not coming).

These characteristics are derived from the Swahili language's World Atlas of Language Structures (WALS) [60].

page [221], which estimates the number of Swahili speakers to be between 50 and 100 million.

Kiswahili is not only the most widely spoken indigenous language in Africa, but it is also the most taught language in the world from Sub-Saharan Africa, according to [54], [182]. Swahili broadcasts are available to East and Central African listeners from several international broadcasting corporations, including

the BBC[2] in the United Kingdom, DW[3] in Germany, VoA[4] in the United States, NHK World-Japan[5] in Japan, and RFI[6] in France.

Kiswahili is supported or integrated into a number of software and social media platforms in order to tap into the potentially large market of Swahili-speaking users. Microsoft (MS), for example, released Swahili versions of MS Windows and Office in 2005 [88]. In 2009, Facebook launched the Swahili version [101], and in 2018, Twitter began supporting Kiswahili, with a tweet in the language being translated to other languages and vice versa [100].

Kiswahili has also been adopted by a number of international and regional integration organizations. Kiswahili is an official language of both the African Union (AU) and the East African Community (EAC) [182]. In 2019, the Southern African Development Community (SADC) adopted Kiswahili as its fourth working language, joining English, Portuguese, and French [145]. Some SADC member states, such as South Africa and Botswana, plan to (or have already) introduced Kiswahili as a learning subject in the school curriculum [2], [106].

3.2 SWAHILI NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) uses artificial intelligence, computer science, and linguistic approaches to process (massive) amounts of natural language data. [18], [171]. NLP is critical to the development of MT, text summarization, and word decompounding, automated text summarization, and spoken language dialogue systems [79], [171]. By definition, MT refers to the computer-aided translation of text or speech from one language to another [218]. Information Retrieval (IR), on the other hand, deals with the organization, storage, representation and access (searching and retrieving) of information, which is primarily unstructured and in their natural languages [11], [174]. Since the data sources/information are in their natural languages, as opposed to controlled languages, then processing such information in IR is naturally an NLP process [213].

NLP has wide applications in IR systems, such as text summarization, question answering, decompounding, word sense disambiguation, and information extraction [79], [171], [213]. NLP is also used in IR for pre-processing [18]. In addition, the language models in IR are derived from NLP [43]. However, NLP solutions did not significantly improve the relevance of results in many IR problems [213]. Recent advances in NLP, thanks to the deep neural network, are altering how IR researchers perceive this field's contribution to IR. This comes on the heels of recent deep learning pre-trained language model successes, specifically the Bidirectional Encoder Representations from Transformers (BERT) [58] and RoBERTa [124]. Guo, Gao, Shi, *et al.* [79] and Long, Ye, Li, *et al.* [127], for example, show how effective deep NLP is for search and recommendation applications.

2 **B**ritish **B**roadcasting **C**orporation (https://www.bbc.com/swahili)
3 **D**eutsche **W**elle (https://www.dw.com/sw/idhaa-ya-kiswahili/s-11588)
4 **V**oice **o**f **A**merica (https://www.voaswahili.com/)
5 NHK is also called Japan Broadcasting Corporation (https://www3.nhk.or.jp/nhkworld/sw/news/)
6 **R**adio **F**rance **I**nternational (https://www.rfi.fr/sw/)

In this section, we visit the NLP works on Swahili, specifically looking into the Swahili stopwords, Named Entity Recognition (NER) and morphological analysis and Part-of-Speech (POS).

3.2.1 *Swahili Stopwords*

Stopwords (also known as stoplits) are words in a text of a specific language that are of little importance and can be safely removed without affecting the performance of the IR system [167]. Removing stop-words is a necessary component of text (pre-)processing in both NLP and some IR applications, such as automatically indexing [126], [167]. Kiswahili, like any other language, has stopwords that must be removed.

Table 3.3 compiles a list of several online, available Swahili stopword corpora and/or removal tools. However, to the best of our knowledge, the only work that describes how they developed a Swahili stopwords corpus is that of Masua and Masasi [137]. In addition to the stopwords dataset, Masua and Masasi [137] created a dataset of Swahili typos and slang. The researchers used a Swahili SMS dataset from the U-report SMS platform, which is a forum for young people to express themselves. The U-reporters could cover a wide range of issues, such as protection, education, health care, and community services. Their corpus contained 4 million words, over 300,000 of which were unique. Using such a dataset ensures that slang and typos among the youth are well captured. However, because the words come from a specific group of people, the dataset may not be large enough to capture as many stopwords as possible.

Table 3.3: A list of Swahili stopwords tools.

No.	Tool	Features
1	Stopword Lists [201]	Support for many other African Languages; and free downloadable dataset.
2	Multi-stopwords [146]	Specify the language(s) of interest and get a list of stopwords or remove them from an array; remove stopwords from a mixed-language sentence; and free access.
3	Stopwords Swahili [45]	The implementation code is freely available as GitHub repository.
4	Stopwords Cleaner [224]	Provides a simple interface to write Python, Scala, or NLU code to remove stopwords.

Some of the tools listed in Table 3.3 work well for NLP tasks; for example, Table 3.4 shows an example of a sentence in which all of the stopwords are successfully removed using the *Stopword Cleaner* tool [224]. Others have some drawbacks, such as the presence of words that are not stopwords. *Stopword lists*

[201] corpus, for example, contains words that are arguably not stopwords, such as *"nyumbani"* (home), *"mama"* (mother) and *"sungura"* (rabbit).

Table 3.4: Stopwords are removed using the *Stopwords Cleaner* tool for Swahili.

Sentence:	*Rais wa Marekani amewasili nchini China na kukutana na Rais Jinping.*
Output:	*Rais, Marekani, amewasili, nchini, China, kukutana, Rais, Jinping*

3.2.2 *Swahili Named Entity Recognition*

Named entities, such as people's names, locations, and organizations, are common in unstructured texts. The process of automatically extracting (locating and classifying) these named entities is known as Named Entity Recognition (NER) [134]. NER can be achieved through hand-crafted grammar-based linguistic (rule-based) or lexicon-based, ML-based (and/or semi-supervised) approaches [73], [76]. NER has several applications including, information extraction, machine translation, question-answering, semantic annotation and automatic text summarization [73]. Despite their high precision, the manual processes and computational linguists required to achieve grammar-based and lexicon-based NERs make these approaches unsuitable [151]. On the other hand, ML-based approaches require a large amount of annotated training data, which is nonetheless a barrier for low-resource languages such as Kiswahili. Additionally, Hu, Ruder, Siddhant, *et al.* [82] concluded that the stated NER approaches that work well for English may struggle to generalize for NER in Swahili.

SYNERGY, proposed by Shah, Lin, Gershman, *et al.* [185], is the most well-known Swahili NER work. The authors used English NER tools on Swahili text that had been machine translated to English, then used word alignment information to project named entity information back into Kiswahili. The SYNERGY's translation process from Kiswahili to English requires the use of a Swahili MT tool and a Swahili morphological dictionary. Using MT adds a cost to the overall process of creating named entities.

In another development, Wentland, Silberer, and Hartung [219] created, through Wikipedia, a multilingual named entity dictionary (HeiNER), which included Kiswahili. The authors found target language equivalences primarily through the use of multilingual links, disambiguation, and redirect pages. They compiled a list of nearly 3,000 Swahili-named entities. However, the number is lower when compared to highly resourced languages such as German, which produced approximately 250,000 named entities.

Steinberger, Ombuya, Kabadjov, *et al.* [198] enhanced the European Media Monitor (EMM) by adding functionality to support named entities and quotations for Swahili news. The authors report success for NER of up to 93.7% precision for person and organization name recognition using a rule-based approach and translation.

Adelani, Abbott, Neubig, *et al.*'s [1] most recent work on Swahili NER employs both supervised (via *gazetteers*) and transfer learning approaches. The authors compared three recent deep neural network NER models as baselines: Multilingual Bidirectional Encoder Representations from Transformers (MBERT) [58], CNN-BiLSTM-CRF [131], and XLM-RoBERTa (XLM-R) [38]. They discovered that their model, based on XLM-RoBERTa, achieves cutting-edge performance for the NER task in 10 African languages, including Kiswahili. Their NER model (MasakhaNER) can be found on the HuggingFace Model Hub[7].

Mueller, Andrews, and Dredze [142] also used a transfer learning approach. They used LORELEI [199] – a DARPA program that supports the development of language technology resources for underserved languages such as Kiswahili – and CoNLL datasets. Mueller, Andrews, and Dredze [142] uses the LORELEI data to train a polyglot NER model, which is then compared to fine-tuned polyglot NER models (including the polyglot Swahili NER model). They come to the conclusion that fine-tuned polyglot NER models outperform their monolingual NER model counterparts. According to their findings, polyglot models learn more by observing multiple languages.

For a summary of Swahili NER, see Table 3.5. The table shows that the research on Swahili NER development is relatively new; dating back to just a decade ago. The development approach is also migrating from the rule-based in the early works [185], [198], [219] to the ML-based in the most recent works [1], [142].

3.2.3 *Swahili Morphological Analysis and Part-of-speech Tagging*

Computational morphological analysis is concerned with (automatically) locating the smallest meaningful units of a word [47]. Morphological analysis is particularly useful in morphologically complex languages such as Kiswahili. *"anajisimamia"* (En: She/He stands up for herself/himself) is an example of a complex word. As a result, computational morphological analysis is required to decompound these words and obtain lemmas, which are sometimes required for parsing and document retrieval purposes [46]. Part-of-Speech (POS) tagging is the process of automatically assigning parts of speech (word class category e.g., verb, and noun) to words in a phrase or sentence using lexical and contextual information [183]. POS can be applied in NER, sentiment analysis, question-answering and word-sense disambiguation [102].

De Pauw, Schryver, and Wagacha [50] proposed an early and well-known work on automatic Swahili POS tagging. The authors tested four corpus-driven taggers: Trigrams'n'Tag (TnT), Support Vector Machines (SVM) Tools (SVMTools), Memory-Based Tagger (MBT), and Maximum Entropy Tagger (MET) (MXPOST tagger). They trained all of the taggers using annotated Helsinki Corpus of Swahili (HCS). The MXPOST tagger produced the most accurate results for Kiswahili, according to the authors, when compared to the other three approaches. They combined individual taggers into a tagger committee to improve performance even further. As a result, the *Swatag*[8] tagger performed the best.

7 https://huggingface.co/Davlan/xlm-roberta-large-masakhaner
8 https://www.aflat.org/swatag

Table 3.5: A summary of Swahili NER.

No.	Study (NER)	Year	Datasets	Approach
1	Wentland, Silberer, and Hartung [219] (HeiNER)	2008	Wikipedia articles	Rule-based and cross-lingual translations and transliterations.
2	Shah, Lin, Gershman, *et al.* [185] (SYNERGY)	2010	Swahili text translated to English	Using an English NER (Union NER) to the English texts translated from Swahili texts.
3	Steinberger, Ombuya, Kabadjov, *et al.* [198]	2011	Swahili news sources	Rule-based and translations.
4	Mueller, Andrews, and Dredze [142]	2020	LORELEI and CoNLL datasets	Fined-tuned neural networks models.
5	Adelani, Abbott, Neubig, *et al.* [1]	2021	Local news sources and gazetteers	Supervised learning and transfer learning.

According to De Pauw and De Schryver [47], there are two approaches for developing a computational morphological analyzer: i) rule-based *two-level formalism*, which uses a massive collection of finite-state transducers and is thus language agnostic; and ii) data-driven approaches, which are corpus-based. According to De Pauw and De Schryver [47], corpus-based approaches are the true language-independent morphological analyzers. Furthermore, De Pauw, Laureys, and Daelemans's [49] comparison of the two approaches on Dutch shows that the rule-based approach fails to outperform the data-driven approach.

Recent approaches use unsupervised Machine Learning (ML) approaches with the goal of automatically inducing morphological properties of a language. On non-annotated data/text, minimum-distance and pattern matching techniques can be used [47]. Hurskainen [86] created *Swahili Language Manager (SALAMA)*[9], a tool for POS tagging and morphological analysis, using the rule-based two-level finite-state formalism. The arguably largest Kiswahili corpus, the Helsinki Corpus of Swahili (HCS), was annotated with SALAMA [85].

De Pauw and De Schryver [47] and De Pauw, Schryver, and Wagacha [50] used a data-driven morphological analysis approach in their work. De Pauw, Schryver, and Wagacha [50] created a Swahili tagger that is morpho-syntactic. Later, De Pauw and De Schryver [47] created a 'Memory-Based Swahili Morpho-

9 http://77.240.23.241/

logical Analyzer' (MBSMA)[10] that outperformed unsupervised data-driven and rule-based approaches for Swahili morphological segmentation and lemmatization, such as *Morfessor*[11] [42] and SALAMA [86], respectively. Morfessor is an unsupervised data-driven morphological analyzer that claims to be able to work with almost any language. However, De Pauw and De Schryver [47] warn that it is not designed to work well with Bantu morphology and, in particular, Kiswahili. Another data-driven work developed a morphological parser using lexical resources from the Kamusi project (a then custom Kiswahili dictionary) and an English word list [121].

In the context of ML-based Swahili morphological analysis, Lindén [118] proposed a semi-supervised lemmatization approach for Kiswahili. They generated a probabilistic model that can guess the base forms of previously unseen words using the annotated version of the HCS. *Conv-LSTM* by Shikali, Mokhosi, Shijie, *et al.* [188] is a more recent work on Swahili POS tagging. Conv-LSTM is an extension of the authors' previous model, *WEFSE*, which is a convolutional neural network model that generates word representation vectors at the syllable level [189]. The Conv-LSTM achieves state-of-the-art POS tagging results with an(98.78%) [188]. Table 3.6 summarizes the works on Swahili morphological analysis and POS tagging. The table shows that research on Swahili morphological analysis and POS tagging was active in 2000's, then stalled until recently.

Table 3.6: A summary of Swahili morphological analysis and POS tagging works.

No.	Study	Year	Type	Approach
1	Hurskainen [86] (SALAMA)	2004	POS tagging and morphological analysis	Rule-based.
2	Creutz, Lagus, Lindén, *et al.* [42] (Morfessor)	2005	Morphological analysis	Data-driven unsupervised ML learning.
3	De Pauw, Schryver, and Wagacha [50] (Swatag)	2006	POS tagging	Corpus-driven supervised ML approach.
4	De Pauw and De Schryver [47] (MBSMA)	2008	Morphological analysis	Data-driven unsupervised ML.
5	Lindén [118]	2008	Morphological analysis	Semi-supervised ML.

... Continued on the next page

10 http://aflat.org/?q=node/241
11 http://www.cis.hut.fi/cis/projects/morpho/

Table 3.6 – continued from the previous page

No.	Study	Year	Type	Approach
6	Shikali, Sijie, Qihe, *et al.* [189] (WEFSE)	2019	POS tagging	Convolutional neural networks.
7	Shikali, Mokhosi, Shijie, *et al.* [188] (Conv-LSTM)	2021	POS tagging	Convolutional neural networks.

There are a number of tools available for POS tagging and morphological analysis of Swahili words, phrases, and sentences (See Table 3.7). Except for the *MBSMA* [47], which no longer exists, these tools are freely available online.

Table 3.7: A summary of Swahili POS taggers and morphological analyzers.

No.	Tool	Features
1	Swatag [50]	Developed using corpus-driven memory-based approach; considers word's morphological properties and the context; freely available online; and limited interpretation of each word.
2	SALAMA [86]	Developed using rule-based two-level finite-state formalism approach; considers every possible interpretation of each word-form, no word context; and freely available online.
3	Morfessor [42]	Covers a wide range of languages; and freely available online.
4	MBSMA [47]	Developed using corpus-driven memory-based approach; and no longer available online.
5	Conv-LSTM [188]	Developed using deep neural networks word embeddings at syllable level; and not available online.

The evaluation of two tools, *Swatag* and *SALAMA*, reveals that the difference lies in the use of contextual information. While the former handles new words by considering their morphological properties and context, the latter gives each word-form every possible interpretation [88]. For an example of a tagged/analyzed Swahili sentence, see Table 3.8 and Table 3.9 – *"Rais wa Marekani amewasili*

nchini China" (The United States (US) President has arrived in China) – using *Swatag* and *SALAMA*, respectively.

Table 3.8: An example of a Swahili sentence that has been tagged with *Swatag*.

Word	*Rais*	*wa*	*Marekani*	*amewasili*	*nchini*	*China*
Tag	N	GEN-CON	N	V	N	ADV

Where N – Noun, V – Verb, ADV – Adverb, and GEN-CON – genitive connector.

Table 3.9: An example of a Swahili sentence that has been tagged with SALAMA.

Word	**Tag**
"<Rais>"	"rais" N CAP 9/6-SG HUM MALE AR HUM
	"rais" CAP N 1/2-SG AN HUM CAP
	"Rais" N TITLE N 1/2-SG AN HUM CAP
"<wa>"	"wa" GEN-CON 3-SG
	"wa" GEN-CON 11-SG
	"wa" GEN-CON 1-SG
	"wa" GEN-CON 2-PL
"<Marekani>"	"Marekani" N PROPNAME SG PLACE
"<amewasili>"	"wasili" V 1-SG3-SP VFIN PERF:me [wasili] SV AR
"<nchini>"	"nchi" N 9/10-SG PLACE LOC
	"nchi" N 9/10-PL PLACE LOC
"<China>"	"China" N PROPNAME SG PLACE
	"china" CAP ADV

3.2.4 *Other Swahili Natural Language Processing Solutions*

Shikali, Sijie, Qihe, *et al.* [189] proposed WEFSE, a convolutional neural network model that generates word representation vectors. They trained it on the Newspapers corpus developed by Gelas, Besacier, and Pellegrino [69] for swahili data. The authors reported that this syllable-aware language model outperformed character aware models in terms of perplexity.

MultiSeg, another Swahili word embedding project, learns bilingual embeddings using parallel and sub-word information such as character n-grams, morphemes, and byte pair encoding [105]. Models that used sub-word information outperformed models that did not use such information.

Martin, Mswahili, and Jeong [135] recently used MBERT to analyze sentiments from Swahili data. They primarily used data from social media platforms (Twitter and YouTube) and discussion forums (JamiiForums and DW Kiswahili). The accuracy of the pre-trained MBERT model was 87.59%. Because the authors stated that their work was ongoing, there was no comprehensive comparison with other approaches for sentiment analysis on Swahili text. Furthermore, to the best of our knowledge, no other works on Swahili sentiment analysis exist.

3.3 SWAHILI MACHINE TRANSLATION RESOURCES

Machine Translation (MT) is mainly used for text and speech translation purposes. Since text and speech are all form of natural language, it is also part of NLP. In IR, it has been mainly used to enable CLIR and MLIR systems [158], [226]. CLIR is a sub-field of IR that deals with retrieving results from a collection of resources in a language other than the query [158]. MLIR, on the other hand processes information in multiple languages, i.e., handles information retrieval when either documents, queries, or both are in multiple languages [149], [158]. Since either the query or the documents are in a different language, then MT makes it easier to translate from the query language to the language of the documents and vice versa. Thus, the success of MT is critical to the success of CLIR and MLIR systems [158], [226]. MT makes it easier to translate from the query language to the language of the documents and vice versa. The availability of resources such as corpora [75] has a large impact on the quality of MT technology.

There are two broad approaches to developing MT: rule-based and corpus-based [52], [53]. For corpus-based MT, the availability of resources such as corpora and bilingual dictionaries has a large impact on the quality of MT output [75]. This section examines the existing resources for enabling Swahili-English MT, regardless of the MT approach. We are particularly interested in corpora (both monolingual and parallel) and online bilingual dictionaries. The bilingual dictionaries and parallel corpora available online are tailored to the Swahili↔English language pair. This review does not cover bilingual dictionaries or parallel corpora for Kiswahili and other language pairs.

3.3.1 *Monolingual Swahili Corpora*

A corpus is a (large) collection of electronic resources, such as text, speech, images, and so on, that are used in a variety of applications, including Machine Learning (ML), NLP, and IR [63]. The corpus may contain documents/resources that are entirely in one language (monolingual corpus) or documents in multiple languages (parallel/multilingual corpus).

The Helsinki Corpus of Swahili (HCS)[12] is the largest Kiswahili corpus developed and updated over a two-decade period. To compile the corpus, the developers primarily used Swahili newspapers, books, and government documents (including Tanzanian parliamentary Hansards) from 1953 to 2016 [85]. The corpus's second edition (HCS 2.0), which is available through the Language Bank

12 http://urn.fi/urn:nbn:fi:lb-2014032624

of Finland – *Kielipankki*, contains 25 million words of written text. It is available in two versions: annotated HCS[13] and unannotated HCS[14]. The SALAMA tagger was used to annotate the corpus, and a constraint grammar parser was used to perform morphological disambiguation of the annotated HCS version [87].

The Swahili Wikipedia edition[15] is the largest when compared to editions in other Bantu languages and the Nilo-Saharan languages as a whole. As of May 2021, this community-contributed resource had approximately 62,500 articles [220].

There are several other monolingual corpora that have been developed for specific purposes. Recently, Masua and Masasi [137] created a corpus of Swahili stopwords, slangs, and typos using the U-report SMS platform, as described in Section 3.2.1. Shikali and Mokhosi [187] created a corpus of Swahili syllables and word analogies. The non-annotated corpus contains texts derived from online media platforms on a variety of topics such as politics, family, sports, general news, and religion. Gelas, Besacier, and Pellegrino [69] developed a Swahili speech corpus for automatic speech recognition (ASR) using a crowdsourcing approach. The SYNERGY corpus, [185], is also a task-specific corpus, as described in Section 3.2.2.

3.3.2 *Parallel Swahili Corpora*

Despite the fact that English and Kiswahili are the official languages of Tanzania, Kenya, Rwanda, and Uganda, documents are not always translated in both languages. Even those that have been translated, particularly government documents, are not made public. As a result, developing a parallel English-Swahili corpus is a difficult task [51]. This explains why there are so few parallel corpora in English and Swahili.

De Pauw, Wagacha, and Schryver [51], [53] created the SAWA corpus, the most well-known annotated Swahili-English parallel corpus, which contains over 2 million aligned words. The corpus includes parallel words from religious books such as the Quran and Bible, as well as political documents, movie subtitles, investment reports, and some materials donated by local translators to the SAWA project [51]. This resource was used by authors such as Sánchez-Martínez, Sánchez-Cartagena, Pérez-Ortiz, *et al.* [176] to develop Swahili-English Statistical Machine Translation (SMT).

Kiswahili is one of over 300 languages in the JW300 parallel corpus [4]. The materials in the corpus are mostly religious, sourced from the Jehovah's Witness website and the church's magazines, *Awake!* and *Watchtower*. These multilingual documents are the result of translations from English into many other languages where the church provides services around the world. Although the documents in the corpus are overly religious, the authors claim that the content covers a wide range of topics. [4] showcased their JW300 corpus by creating a cross-lingual word embedding induction, POS projection, and MT. The JW300 corpus was also

13 http://urn.fi/urn:nbn:fi:lb-2016011301
14 http://urn.fi/urn:nbn:fi:lb-2016011302
15 https://sw.wikipedia.org/wiki/Mwanzo

used by Nekoto, Marivate, Matsila, *et al.* [147] in the development of a Swahili MT system.

GoURMET, created by Sánchez-Martínez, Sánchez-Cartagena, Pérez-Ortiz, *et al.* [176], is another English-Swahili parallel corpus created through website scraping. The corpus contains over 3.3 million English tokens and approximately 3 million Swahili tokens. This resource was used by the authors to create a neural MT for the news domain.

IARPA's Machine Translation for English Retrieval of Information in Any Language (MATERIAL)[16] is another parallel corpus. The corpus is bilingual in order to facilitate research in low-resource languages. The MATERIAL corpus was used in the works of Zbib, Zhao, Karakos, *et al.* [228] and Zhang, Westerfield, Shim, *et al.* [230], which involved neural network MT for Kiswahili.

The OPUS, one of the most important data sources for parallel corpora, houses several corpora, including two Swahili-English collections [204]. The first corpus, Tanzil[17], contains Quran translations [203], while the second, GlobalVoices[18], is a collection of news stories from the Global Voices website [203]. Table 3.10 summarizes the monolingual and bilingual/parallel corpora.

Table 3.10: A summary of the Swahili corpora that have been reviewed (both monolingual and parallel).

No.	Corpus	Features
1	HCS [85], [87]	Monolingual corpus; developed from Swahili newspapers, books and government documents; has both annotated and non-annotated versions; and has a total of 25 million words.
2	Wikipedia [220]	Monolingual corpus; has about 62,500 articles; the number of words is not known; and the texts are not annotated.
3	Stopwords, Slangs and Typos [137]	Monolingual corpus; developed from U-report users' SMS; non-annotated corpus; has a total of 4 million words; and limited to stopwords, slangs and typos.
4	Syllables and word analogy [187]	Monolingual corpus; limited to syllables and word analogy; non-annotated corpus; and covers a range of topics such as sports, religion, family, general news and politics.
5	SYNERGY [185]	Monolingual corpus; and limited to named entities

... Continued on the next page

16 https://www.iarpa.gov/index.php/research-programs/material
17 https://opus.nlpl.eu/Tanzil.php
18 https://opus.nlpl.eu/GlobalVoices.php

Table 3.10 – continued from the previous page

No.	Corpus	Features
6	SAWA [53]	Parallel (Swahili-English) corpus; has over 2 million aligned words; the text is annotated; and developed mainly from Bible and Quran.
7	JW300 [4]	Parallel (Swahili-English) corpus; and developed from Jehova's Witnesses website and magazines.
8	GoURMET [176]	Parallel (Swahili-English) corpus; developed from scraped websites; and has over 3.3 million English and 3 million Swahili tokens.
9	OPUS [204]	Parallel (Swahili-English) corpus; and has two versions, one compiled from Quran translations and the second compiled from news stories on the GlobalVoices website.

3.3.3 *Bilingual Swahili Dictionaries*

Although there are many electronic bilingual Swahili dictionaries available, it is unclear how they were created/developed. The lexicographic processes are unclear, whether they are corpus-based or based solely on the lexicographer's intuition [55]. To the best of our knowledge, De Schryver [54] is the only work that details a framework for electronic dictionary development. Their detailed description of how to create an artificially intelligent lexicographic (*aiLEX*) corpus, which resulted in the creation of the *Go Swahili-English Dictionary*[19].

Joffe, MacLeod, and De Schryver [99] introduced the TshwaneLex electronic dictionary system, which provides a platform for dictionary compilers to publish their dictionaries. This system has eight key features that enable dictionary developers/compilers to: i) localize and dynamically customize the metalanguage; ii) integrate multimedia into the dictionary; iii) integrate the dictionary into MS Word; and iv) perform customized searches quickly; v) download system updates and encryption to improve data security; vi) change the view mode and customize the styles and colors; vii) customize the pop-up help; and viii) create a desktop icon. This system is used to publish *TshwaneDJe Swahili-English Dictionary (TeDJe-SED)*[20]. Bański and Wójtowicz's [12] *FreeDict* is a platform for hosting compiled dictionaries. The FreeDict Swahili↔English[21] is hosted on this open-source architecture.

19 https://www.goswahili.org/dictionary/
20 https://tshwanedje.com/dictionary/swahili/
21 https://www.freedict.com/onldict/swa.html

The online Swahili dictionaries are summarized in Table 3.11. The table (from item 5 to 8) contains some Swahili-English dictionaries that do not have supporting papers but are available online. Except for the *Go Swahili English dictionary* [54], all dictionaries provide translations from either language, i.e., Swahili↔English translations.

Table 3.11: A list of some Swahili-English online dictionaries and translators.

No.	Dictionary	Features
1	Go Swahili-English dictionary [54]	Swahili→English dictionary; word level translation; and limited to basic single word translations and word lookup.
2	Freedict [12]	Swahili→English dictionary; word level translations; built on an extensible architecture, an open-source Freedict dictionary; and limited to basic single word translations and word lookup.
3	TshwaneDJe-SED [99]	Swahili→English dictionary; word and phrase level translations; has over 16,000 entries and phrases, and over 36,000 translation equivalents; supports morphological decomposition and corpus-based phrase examples; its interface allows cross-referencing and supports integration with MS Word processor; supports British and American English spellings; available online and as a standalone software for MS Windows operating system; and it is proprietary.
4	SALAMA [86]	Swahili→English dictionary; word level translations (Word and phrase translations for the SALAMA Translator); supports morphological decomposition and phrase examples; and gives phrase-based examples with POS tags.
5	TUKI[22]	Swahili↔English dictionary; word level translation; has more than 50,000 entries; translations include POS tags, lemmas, derivations and occasional example phrases; no supportive interface for translation purposes, resembles an in print dictionary; and limited to basic single word translations and word look up.

... Continued on the next page

Table 3.11 – continued from the previous page

No.	Dictionary	Features
6	Glosbe[23]	Swahili→English dictionary; word level translations; translations are community-contributed; provides examples of usage; and corpus-driven translations.
7	Lingvanex[24]	Swahili→English dictionary; word and phrase, sentence and paragraph level translations; and the interface resembles that of the search engines translators.
8	Web-based dictionaries and translators e.g., Bing Translator[25] and Google Translate[26]	Swahili→English dictionary; word, phrase, sentence and paragraph level translations; and translations may be accompanied by speech (machine generated pronunciation) and images.

The Go Swahili-English, SALAMA, Glosbe, TUKI, and FreeDict dictionaries only provide word-for-word translations. For example, as shown in Figure 3.2, the Go Swahili-English dictionary breaks down a phrase or sentence into a bag-of-words before translation. Such dictionaries' applications are limited to simple single-word translations and word lookups.

Some dictionaries, such as TUKI, TshwaneDJe-SED, and the SALAMA dictionary, provide POS tags and morphological decomposition, as well as corpus-based phrase and/or sentence examples. The home page of the TshwaneDJe-SED, for example, shows an example of a compound word *tunaosha* (En: We are washing), as shown in Figure 3.3. TshwaneDJe-SED has the most detailed lexicographic information on Swahili than any other online dictionary for the language, according to De Pauw, De Schryver, and Wagacha [48].

TshwaneDJe-SED, SALAMA Translator, Lingvanex, and Web-based dictionaries and translations all support word, phrase, sentence, and paragraph translations. As a result, they are appropriate for high-level translation.

22 http://www.elimuyetu.co.tz/subjects/arts/eng-swa/5.html
23 https://glosbe.com/sw/en
24 https://lingvanex.com/english-to-swahili/
25 https://www.bing.com/translator
26 https://translate.google.com/

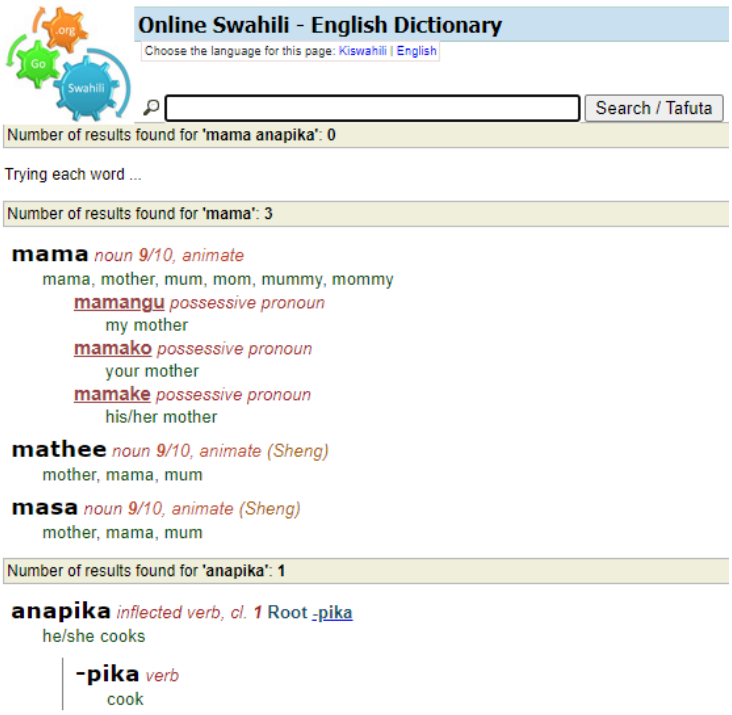


Figure 3.2: A word-for-word translation from the Go Swahili Dictionary.

Source: Go Swahili Dictionary.

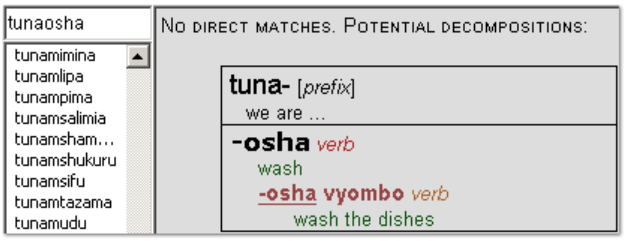


Figure 3.3: An example of a morphologically decomposed compound word translation by the TeDJe-SED dictionary.

Source: The TshwaneDJe SED.

3.4 APPLICABILITY OF SWAHILI NLP AND MT TO SWAHILI IR AND MLIR

Language technology research for low-resource languages such as Kiswahili is limited. There are a few available works, datasets, and tools in this paper, which specifically focused on two main areas – NLP and MT. In this section, we visit the state-of-the-art Swahili IR systems that are successfully developed out of the MT resources and NLP solutions. Our review does not cover applications of Swahili NLP and MT in other fields of study apart from IR and/or CLIR/MLIR. The CLIR and MLIR systems rely on (machine) translations, which traditionally rely on bilingual dictionaries and/or parallel corpora. Translation is mainly required in CLIR and MLIR problems for the purpose of translating the query to the targeted documents' languages or translating the targeted documents to match the query language [149], [158].

Parallel corpora are used in probabilistic (or statistical) (Statistical Machine Translation (SMT)). According to Arora, Shterionov, Moriya, *et al.* [8], translation using the SMT approach has not resulted in better CLIR/MLIR systems. However, MT systems' development has been impeded by the translation's lack of robustness. Problems such as out-of-vocabulary (OOV) and corpora with insufficient data result in poor-performing MT, which escalates to poor IR, and/or CLIR/MLIR systems. Current state-of-the-art approaches to MT use neural networks (Neural Machine Translation (NMT)) as an alternative to traditional corpus-based SMT, [211]. Again, training a NMT for Swahili, a low resource language in terms of parallel corpora, remains a challenge. The Swahili corpora we reviewed are insufficient to train better translation models. To avoid the OOV problem and the need for large amounts of data to train the translation model, Zbib, Zhao, Karakos, *et al.* [228] used only parallel data to train the model and compute the CLIR relevance scores. Swahili, Tagalog, and Somali were used in the evaluation because they are low-resource languages. Their NMT approach significantly improved CLIR relevance when compared to the SMT baseline.

Yarmohammadi, Ma, Hisamoto, *et al.* [226] used document representations that combined several best translations (N-best) and a bag-of-words (BoP). This SMT and NMT approach translates at the document level by utilizing a shared embedding space for both the document and the query. The authors reported that their N-Best+BoP representation improved CLIR performance on all three low-resource languages used, namely Somali, Swahili, and Tagalog. A related work to this, is that of Boschee, Barry, Billa, *et al.* [17], which approaches the translation and retrieval problem by mapping both the query and the targeted documents into a shared embedding space where they can perform the retrieval from that space.

Arora, Shterionov, Moriya, *et al.* [8] combine both the dictionary-based and the (based) SMT in their multi-modal CLIR (MMCLIR). Their goal was to use English queries to retrieve Swahili speech and text documents. While they reported excellent results for text documents, they reported poor results for speech documents. The authors acknowledged the difficulty of CLIR speech in low-resource languages.

CLIR/MLIR systems, as previously stated, include a translation step. There are several approaches to translation in the literature, including MT-based, dictionary-based, ontology-based, and corpus-based [140]. Despite the numerous dictionaries discussed above, it should be noted that none of the studies on Swahili CLIR/MLIR used dictionaries for translation. This could be due in part to the OOV issue, as dictionaries have limited entries and require exact matches [8]. The Swahili CLIR/MLIR approaches presented above all rely on MT, either SMT or NMT. The majority of the corpora were used to develop MT models, such as in the works of Sánchez-Martínez, Sánchez-Cartagena, Pérez-Ortiz, *et al.* [176] and Nekoto, Marivate, Matsila, *et al.* [147] who developed SMT systems, and Zhang, Westerfield, Shim, *et al.* [230] and Zbib, Zhao, Karakos, *et al.* [228], who developed NMT systems. The created MT models can then be used for CLIR/MLIR translation. The ontology approach for CLIR/MLIR translation has yet to be investigated.

In general, we find that works on Swahili IR, and particularly Swahili CLIR/MLIR, do not heavily rely on reviewed works and developed datasets. The stopwords, NER, and morphological analysis are not explicitly mentioned as having been

used in the development of Swahili CLIR and MLIR systems. Some of the Swahili CLIR and MLIR works use probabilistic or neural models for MT to avoid the purely dictionary-driven translation process.

3.5 SUMMARY

The second section of the literature reviewed studies on existing Swahili NLP, MT and CLIR/MLIR solutions. The chapter began with a history of Kiswahili, an East African Bantu language that is the most widely spoken indigenous language in Africa, with approximately 100 million speakers. The agglutinating nature of the language makes Natural Language Processing (NLP) applications difficult to use. As a result, we examined several works on Swahili NLP, including stopwords, Part-of-Speech (POS) tagging, computational morphological analysis, and Named Entity Recognition (NER). Despite the low resource nature of the language, tools developed from these works, such as *SALAMA* and *Swatag* taggers, produce excellent results.

To successfully develop CLIR/MLIR, a successful Machine Translation (MT) system that can automatically perform translations of either queries or documents is required. This chapter examined a variety of resources for enabling MT, including Swahili corpora and bilingual dictionaries. Several studies on Swahili MT have used monolingual corpora such as HCS and Wikipedia, as well as multilingual corpora such as the SAWA corpus.

The chapter also reviewed the existing Swahili CLIR and MLIR works. There are not many works in this genre. To overcome MT's poor performance in implementing CLIR/MLIR, these studies propose alternative approaches, such as combining dictionary-based and corpus-based approaches, using document translations, and training only the translation model before calculating relevance scores.

The reviewed works on Swahili IR and CLIR/MLIR approaches continue to leave users out of the loop by failing to consider their behaviours and preferences when interacting with IR and/or CLIR/MLIR systems. This is one of the reasons we conducted the research described in this thesis.

Part II

EXPLORING THE LANGUAGE PREFERENCES

4

UNDERSTANDING SWAHILI-SPEAKING WEB USERS' INFORMATION NEEDS AND SEARCH BEHAVIOUR

In order to develop an intervention, it is critical to first understand the problem domain and then learn about the needs and perspectives of the users. As a result, the purpose of this study was to gain a better understanding of two fundamental aspects: the search behaviour of Swahili-speaking Web users and the needs and uses of Swahili information in Tanzania. Thus, this chapter describes and presents findings from an analysis of the perspectives and practical expressions of experts representing Swahili-speaking Web users in their search for information in Tanzania. The chapter, in particular, attempts to answer the first research question (**RQ1**) of this thesis, which states that:

> *"What are the information needs and search behaviours of Tanzanian polyglot Swahili-speaking Web users?"*

To address this question, we established the following research objectives.

RO1 To learn about search language preferences and the reasons behind them from the people being investigated.

RO2 To reveal the experience of Swahili query formulation efforts, and the associated reasons from the people being investigated and the community they serve.

RO3 To determine the preferences for information language among professionals and ordinary citizens using the opinions of those being investigated.

RO4 To learn about citizens' needs for Swahili information in various sectors.

RO5 To inquire about Swahili IR and language technology and tools awareness.

The remainder of this chapter is organized as follows. The research methods used to carry out the study are presented in Section 4.1. The results are in Section 4.2, and a discussion of the results is in Section 4.3. The final section of the chapter (Section 4.4) summarizes the key findings of the study.

4.1 METHODOLOGY

4.1.1 *Research Approach*

The study employed a qualitative strategy, primarily the interview approach, with the goal of conducting in-depth interviews with key informants in the fields of information science and Kiswahili.

4.1.2 *Study Location*

This research was carried out in the United Republic of Tanzania. We chose the country to represent other Swahili-speaking countries in the region primarily because it has the greatest number of first language speakers as well as the largest number of speakers compared to its neighbouring countries [64], [144]. Also, Kiswahili is used as a teaching and learning medium in primary schools and adult education, as well as a national and official language [172]. English, on the other hand, is used in secondary schools, colleges, universities, higher courts, and the majority of official government communications [172], [205].

4.1.3 *Targeted Population*

The target group consisted of information science experts and Kiswahili specialists. This population is well-versed in the subject under investigation. Librarians/information science experts from all libraries across the country, as well as Kiswahili specialists and lecturers in information/library science and Kiswahili, qualified to participate in the study. We used expert representatives (information experts and Kiswahili specialists) primarily for convenience and due to thesis time constraints.

4.1.4 *Sampling Procedure and Sample Size*

Based on qualitative research, the study required a small number of participants who are sufficiently knowledgeable in the field. As a result, we purposefully chose six libraries, three of which were public and three of which were university. The selection of university libraries was guided by the specialization of the given university. The Sokoine National Agricultural Library (SNAL)[1] represented the field of agriculture and allied sciences, the Muhimbili University of Health and Allied Sciences (MUHAS)[2] field of medicine and allied sciences, and the University of Dar Es Salaam (UDSM)[3] library field of all disciplines. We purposefully chose the National Kiswahili Council [of Tanzania] (formally known as *Baraza la Kiswahili la Taifa (BAKITA) [Tanzania]*[4] in Kiswahili) and Kiswahili lecturers from the State University of Zanzibar (SUZA) and University of Dodoma (UDOM) Departments of Kiswahili. As stated in Table 4.1, the goal was to have at least two participants per institution.

Librarians/information scientists represent the general public of information searchers because they have in-depth knowledge of information searching practices and are well-informed about their information customers in libraries. Swahili specialists represented the general Swahili-speaking public and the language's technical aspects.

To select individuals for participation in the study, we sent request letters to the heads of institutions and departments, requesting that some of their employ-

1 https://www.lib.sua.ac.tz/
2 https://library.muhas.ac.tz/
3 http://library.udsm.ac.tz/
4 https://www.bakita.go.tz/

Table 4.1: The number of expected and actual participants in a study of Swahili-speaking Web users' information needs and search behaviour.

SN.	Institution	Expected Participants	Interviewed Participants
1	The National Library of Tanzania	2	2
2	Arusha Regional Library	2	1
3	Dodoma Regional Library	2	1
4	UDSM Library	2	0
5	SNAL Library	2	4
6	MUHAS Library	2	1
7	SUZA (Kiswahili lecturers)	2	0
8	UDOM (Kiswahili lecturers)	2	0
9	BAKITA (Kiswahili specialists)	2	2
	Total	**18**	**11**

ees be recruited. Potential participants' names and contact information were provided by the heads. We then sent invitation messages and the consent form to these participants via email and/or WhatsApp messages. Please see Appendix A for templates of request and invitation messages to heads and participants. Because we were not physically present in Tanzania, we conducted the interview entirely electronically, from the interview appointment to the interview itself, using emails, phone calls, Skype calls, and WhatsApp messaging.

While 11 participants responded to the communication and took part in the study, the remaining 7 either did not respond or did not pick up the phones on the agreed-upon interview dates, even when we tried to reach them on other days. As a result, they were unable to provide final consent to participate in the study. According to researchers in qualitative studies e.g., Fugard and Potts [67], the number of respondents can range from 2 to over 400 participants. The number of participants chosen is largely determined by the prevalence of the population theme, the number of instances of the theme desired, and the study's *power*, or the likelihood of obtaining the desired number of theme instances [20], [67]. Thus, it was reasonable to proceed with the study with the 11 available participants.

4.1.5 *Data Collection*

We used Skype[5] software to interview consenting participants after obtaining ethical clearance approval[6] to conduct research involving human subjects. The interviews with the participants took place on various dates over the course of two weeks, from May 25th to June 8th, 2018. The interviews were all conducted in Kiswahili (participants preferred the language over English), and each session took an average of 25 minutes. We used a third-party freeware tool called MP3 Skype Recorder[7] to record the conversions.

The interview schedule was divided into five major sections, namely: demographic information; search behaviour; Swahili search experiences; Swahili information needs and uses; and IR and language technology awareness, as shown in Table 4.2.

Table 4.2: Major aspects of the interview schedule used in the study on Swahili-speaking Web users' information needs and search behaviour.

Category	**Variables**	**Measurement**
Demographics Information	Job title/position	Open-ended
	Work experience	Open-ended
	Previous job	Open-ended
	Relation of the previous job to the current job	Open-ended
Search behaviour	Search language	En or Sw
	Reasons for the choice of such language	Open-ended
Experiences of Swahili Search	Time to think and formulate query	Open-ended
	Query size & difficulty	Open-ended
	Relevance of Swahili results	Open-ended

... Continued on the next page

5 https://www.skype.com/en/

6 Approved by UCT, Ethical Clearance Approval Code: FSREC 26 - 2018 (Refer to Section A.1.1)

7 https://voipcallrecording.com/ (Skype did not have a recording feature at the time of this study. They announced the recording feature in September 2018, according to the Skype Blog)

Table 4.2 – continued from the previous page

Category	Variables	Measurement
	Reasons for such results	Open-ended
Needs and uses of Swahili information	Swahili information needs among professionals	Open-ended
	Swahili information needs among ordinary citizens	Open-ended
IR and language technology	Examples of Swahili tools you know	Open-ended
	Importance and challenges of such tools	Open-ended

4.1.6 *Data Analysis*

Because there were no accurate and reliable tools for Swahili transcription available, we manually transcribed the conversations and then translated them to English. Using a qualitative content analysis procedure known as open coding [39], [200], the transcribed responses were grouped, coded, tagged, counted, and ranks were generated based on their themes.

In a spreadsheet program, we saved all of the codes for a specific question's responses and shaded the related codes in the same color. Then, as needed, we grouped codes of the same color to form a category called a theme, or we subdivided a theme into sub-themes. We counted and/or produced rankings for each theme and sub-theme. We also assigned each participant a number between 01 and 11 so that we could easily quote them.

4.2 FINDINGS

4.2.1 *Demographics Information*

As shown in Table 4.1, the study included 11 participants: 9 librarians/information scientists from three public libraries and two university libraries, as well as 2 Kiswahili specialists from BAKITA. 5 of the participants were female, while 6 were male. 3 of the 9 librarians held administrative positions such as regional librarians and section heads; 2 taught in library and information science degree programs; and 4 were librarians and library officers. The 2 Kiswahili specialists identified themselves as language investigators.

Only 1 participant had less than 5 years of work experience; 5 had up to ten years; 3 had between 11 and 20 years; and the other two had worked for more than 20 years (see Table 4.3).

Table 4.3: Work experience of interview participants.

Years worked	0-5	6-10	11-20	21-30	31+
Participants	1	5	3	1	1

The previous job experience of the participants differed from one another. While 7 participants began their careers in their current positions, 3 were teachers and 1 was a factory worker. This implies that the positions and work experiences of the participants provide confidence that they were the right informants to meet our goal – the experiences and preferences of Swahili-speaking information seekers on the Web.

4.2.2 *Behaviour of Participants during Web Search*

Table 4.4 displays various codes that represent our respondents' Web search behaviour.

Table 4.4: Participants overall behaviour during Web search.

Theme	**Code**	**Participants**
Information Source	Web	11
	Dedicated digital library	1
Language	Contents	11
	Search engine's interface	0
Language use	Both English and Kiswahili	10
	English only	1
	Kiswahili only	0

The table (Table 4.4) shows that all participants acknowledged to using search engines to find information on the Web. One participant stated that, in addition to using general search engines such as Google, they conduct some searches in dedicated library databases (directories). Participants also stated that they are unconcerned about the language of the search engine's interface as long as the results are in a language they understand. As an example, one participant stated:

> *"What I need to look at is the content itself to determine its relevance. I usually think about the [language of the] content and ignore the interface."*(03)

And another said:

> *"Language [of the results] is important because you cannot retrieve information from a language in which you are not fluent or in which the client, if you are assisting someone, does not understand."*(o8)

Except for one participant, all acknowledged to using both English and Kiswahili in their information searches. However, they indicated to prefer English over Kiswahili. One participant never searched for information on the Web in Kiswahili.

4.2.2.1 *Reasons for Preferring using English*

The majority of participants cited the following reasons for preferring English in the most of of their search sessions:

- Trust in the English information. According to one of the participants, English information is

  > *"... reviewed and standard information."*(o8)

- The speed with which English information can be found in comparison to Kiswahili information.
- Unsatisfactory results in Kiswahili. Some participants were concerned that using Kiswahili would not capture the context of the search, causing them to waste a lot of time searching.
- The dominance of English documents on the Web. Others believe that the massive number of English documents on the Web has an impact on making English the obvious language to use.

4.2.2.2 *Reasons for the occasional use of Kiswahili*

We also asked participants why they sometimes use Kiswahili while searching the Web. Here are their views:

- For information that is not related to one's profession/job. For example, the librarians mentioned preferring searching for social issues information in Kiswahili.
- The context of information. Participants mentioned the use of English especially in the local context may cause irrelevant information. Thus, it is suitable to use Kiswahili. Others said the use of Kiswahili in a particular context can help one get some background or clue regarding the region or information they want to search.
- Challenge with English terminology. Participants stated that the challenge to formulate a query in English may force one to use Kiswahili, especially when one has proper terminology in Kiswahili.

Table 4.5: Reasons for switching between English and Kiswahili when searching on the Web.

Code	Participants
Information need	7
Topic of search	5
Type of the task	4
Information context	4
Terminology	3
Trustworthy results	1
Easy to get results	1

4.2.2.3 *Reasons for Language Switching between English and Kiswahili*

Despite the stated preference for using English in most search sessions, as stated in the preceding section, there are several reasons for using Kiswahili. Table 4.5 summarizes the driving forces for such search language dynamics.

One participant contended that language switching is merely a search technique.

> *"But if you search in English and feel like you cannot get enough materials or they are almost irrelevant, especially if the search context is in Tanzania, you can simply switch to Kiswahili and see if you can get relevant information."* (08)

4.2.3 *Searching Experience in Kiswahili*

As previously stated, ten participants indicated that they search the web in both English and Kiswahili. We wanted to learn more about their experiences with Swahili query formulation and the relevance of Swahili results.

4.2.3.1 *Swahili Query Formulation Efforts*

We rate the query formulation efforts used in the study based on the amount of time it takes to think about and formulate the query, the size of the query (number of words), and the level of difficulty, as coded in Table 4.6.

While 2 participants stated that formulating a Swahili query takes the same amount of time as formulating an English query, 4 participants made no comments on this, implying that they do not notice the difference. Only 3 participants said it would take longer to think about and formulate a Swahili query than it would to think about and formulate an English query. They hinted that the reason is their experience and expertise in Kiswahili. The other 2 participants

Table 4.6: Swahili query formulation efforts.

Theme	Code	Participants
Time	No comments	4
	Much	3
	Same	2
	Less	2
Size	No difference	7
	More words	3
Difficulty	Hard	7
	No comments	2
	Easy	1

believed that creating a Swahili query was much faster. According to these participants, the nature of their jobs, search skills, and language proficiency enabled them to think quickly and formulate search queries –

> *"... because I understand the language, I simply jump to the specific topic."*(06)

In terms of Swahili query size, three participants, those who take much time to formulate Swahili queries, stated that they use more words than for an English query. One participant asserted, using a technical term as an example:

> *"English has more terminology, particularly in these technical terms. As a result, I can use a single English word rather than a slew of Swahili words."*(02)

In terms of how difficult it is to formulate Swahili queries, even those who spend little time on it believe it is difficult. This is because some users form an English query (in their minds) before writing it in Kiswahili. A query conceived in English, on the other hand, is difficult to translate into Kiswahili. One must be fluent in both languages or knows the search strategies –

> *"... with sufficient proficiency in both languages, you simply change a[n] [English] search term to Kiswahili."*(08)

4.2.3.2 *Relevance of Swahili Results*

Only 3 participants admitted to getting relevant results when searching in Kiswahili. The remaining 7 had mixed feelings about the relevance of the results, with some saying it depends on the precision of the query and others saying the results are insufficient, imprecise, and sometimes completely irrelevant.

Table 4.7: Reasons for poor relevance of Swahili results.

Code	Participants
Scarcity of Swahili documents	7
Search techniques and query formulation	5
Limited vocabulary	4
Outdated Swahili information	3
Search engines problems	1
dedicated search systems	1

We sought explanations from the seven participants who held opposing views on the relevance of Swahili results.

The reasons for their opinions are as summarized in Table 4.7.

- Scarcity of Swahili documents on the Web. The majority of participants saw this as the primary cause of unsatisfactory results. Participants linked the scarcity of Swahili documents to factors such as a small number of Swahili authors, particularly in specialized fields such as medicine. According to one participant:

  > *"... The authors and experts have not recognized the significance of publishing books in Swahili."* (07)

  Furthermore, this participant believes that:

  > *"... some writers do not publish their work online."* (07)

  The participant insisted that the few Swahili documents in the libraries reflect what is available on the Web.

  > *"Forget about Web documents; look at books. For example, in our library, there are a few Swahili books, as if we are not Swahili speakers."* (07)

  As a result, Swahili-speaking searchers have very few options to explore.

- Search techniques and query formulation abilities, particularly query precision. A participant complained that because searchers do not know what they are looking for, they create broad queries. For example,

  > *"You come across someone looking for Kitenge[8] costume design, but you don't specify which country the costumes are from! Because there are costumes from Congo, Nigeria, and other countries, but perhaps this person is only interested in Tanzanian designs."* (05)

8 A cotton fabric from East Africa, West Africa, and Central Africa that is printed in a variety of colors and designs with distinct borders and is primarily used for women's clothing. In Zambia, Malawi, and Namibia, the fabric is also known as Chitenge. Sources: Google Dictionary and the Kitenge page on Wikipedia.

According to some participants, many people lack knowledge of query formulation and experimenting with terminologies.

- Limited vocabulary. Participants stated that many Swahili speakers, particularly in Tanzania, struggle with proper terminology in both English and Kiswahili. For example, one participant stated:

  > *"If I want to search for "medication for allergies," how do I write "allergy" in Kiswahili? But I know there is allergy information out there, and I know if I search in English, I'll find it."* (10)

- Swahili information that is no longer current. Participants noted that some Swahili information sources do not receive regular updates and thus cannot be trusted. As an example, one participant stated:

  > *"Today, if you search for Tanzania's President, the search engine returns [Jakaya] Kikwete! [9]."* (08)

- Search engines provide inadequate support for other languages. Some participants believed it was a flaw in the search engines because, English is used to engineer search engines, the rules/principles used to filter English results would not work for Swahili queries. As an example,

  > *"Will you write **"NA"** [and, English] when searching for something in Kiswahili and you want to narrow down your results? The [search] engine will not understand if you use **"NA"**. It is not capable of defining your intent."* (11)

- Using dedicated search systems to conduct searches. A participant mentioned that searching for Swahili documents in dedicated sources such as library directories, which have no or only a few Swahili documents, may be unsuccessful.

4.2.4 *Perceived Needs and Applications of Swahili Information*

4.2.4.1 *Swahili Information among Professionals*

We asked our participants to describe their perspectives on the needs for online Swahili or English information in their jobs if they were professionals in other fields (such as accounting, medicine, engineering, and so on) or had attended such people. In such a case, 3 participants strongly argued that Swahili information has no place. The remaining 8 stated that they needed a combination of information in both languages, but insisted on English information, similar to the search language in Section 4.2.2 above.

The following are the most frequently stated reasons for preferring English information for professional use:

- *English is used as a medium of instruction at professional levels of education such as colleges and universities*. All colleges and universities in Tanzania offer

9 The former Tanzanian president (2005-2015); the incumbent president at the time of this study was Late John Magufuli (2015-2021).

professional programs in English. Kiswahili, on the other hand, is only used as a medium of instruction in public primary schools [172], [205]. English is used as a medium of instruction in a number of private primary schools. As a result, the majority of professionals receive their training/skills in English. As a result, it is difficult for them to learn new professional terminology in Kiswahili for their Web searches.

- *In Kiswahili, there is little to no scientific and professional information*. Because professional training and programs are in English, very few authors will be able to write such documents in Kiswahili because there will be no audience.
- *The original information is distorted.* Some participants believed that using Kiswahili would taint the originality of the information because most professional information is written in English.

Kiswahili is desirable when professional information needs to be delivered to the public – popular science – and a simple or indigenous language is required. It was also stated that it is a necessary language for jobs and professions that require writing reports in Kiswahili.

4.2.4.2 *Swahili information among ordinary citizenry*

We asked the participants what they thought about their needs for Swahili information if they were ordinary citizens. It should be noted that, in addition to non-professionals in any field, the term "ordinary citizen" as used in this study may also include anyone looking for information that is not related to their profession, especially if they do not understand the technical jargon used. In contrast to a professional guitarist, we might consider an accountant looking for entertainment information to be an ordinary citizen.

Surprisingly, none of them desired English information! In a multilingual country like Tanzania, all participants believe that Swahili information is more important and necessary for ordinary citizens than English information. One of the participants asserted:

> *"Look at the number of daily English newspapers if you want to prove this! They publish a small number of copies; if they print a large number, they will incur a loss."* (10)

Another participant added:

> *"and perhaps another indication that people [ordinary citizens] require this information [in Swahili] is the sale of booklets [in Swahili] by a few experts on poultry farming, fish farming, and so on. Those Swahili booklets sold on the side of the road are the best-selling in the country."* (02)

The participants revealed several reasons that highlight the importance of Swahili information among ordinary citizens. For someone to be competent in English in Tanzania today, they must have a high level of education. Participants were concerned that many people in the country are still uneducated or do not speak English well. According to one participant:

> *"Because of your lack of education, many of the terms you are familiar with are in Kiswahili. So you spend the majority of your time looking for Swahili information. Your English vocabulary is limited."* (01)

As a result, it is advantageous to obtain information in the language in which they are competent and fluent – the language in which they interact on a daily basis. When using such information, they will not struggle with the terminology because they can easily read and understand it. Furthermore, it is most people's mother tongue and Tanzania's national language; we assume that most people can understand Kiswahili.

4.2.5 *The Demand for Swahili Information in Various Sectors*

We also asked the participants for their thoughts on which sectors they believe require the most Swahili information. The majority of participants believe that all Tanzanians, regardless of where they live or work, require Swahili information. Participants, on the other hand, were more categorical in the following sectors as summarized in Table 4.8:

Table 4.8: Demand for Swahili information in various sectors.

Code	**Participants**
Agriculture	9
Justice	6
Health and Well-being	4
Entertainment	2
Banking	2
Entrepreneurship	2
Communication	1
Business	1
Social networking	1

- *Agricultural sector* Participants specifically suggested that farmers be provided with Swahili information on farming practices. Information such as cutting-edge farming solutions is desperately needed in a common language. Farmers require this information in order to transition from traditional farming to large-scale farming or, at the very least, agri-business. Participants mentioned how difficult it is to find Swahili information on best practices in livestock husbandry, fish farming, poultry farming, and horticulture in their local libraries.

> *"These are the Swahili information that most people want, such as how to choose the best bull, modern methods of milking and feeding, and so on."* (05)

> *"A farmer's language is usually Kiswahili. People will be delighted if they can learn about entrepreneurship, particularly poultry farming, in Kiswahili."* (07)

As an example, one participant suggested that veterinary medicine product descriptions be written in Kiswahili. –

> *"... and this is where popular science and popular language come into play."* (08)

- *Justice sector*. Participants referred to the need for judicial information in Kiswahili, citing the inability of ordinary citizens to understand and interpret court orders, legal contracts, and documents. Because these documents are all written in English, a person cannot read his or her own ruling; they must rely on a lawyer to translate and interpret for them.

- *Health and Well-being sector*. Others mentioned the need for Swahili information on human diseases and medication, citing the top-selling Swahili booklets on health tips as an example. –

  > *" For example, you can find a booklet titled* ***ujue ugonjwa wa kisukari*** *[Learn about diabetes]."*(02)

  English medical terminologies are difficult to understand and must be interpreted by a physician. Participants believe that Swahili speakers are familiar with some diseases by their local names and that they can understand their illnesses in Kiswahili.

- *Other sectors*. Participants also mentioned the entertainment, banking, business, entrepreneurship, communication, and social networking sectors as having a high demand for Swahili information.

4.2.6 *Swahili Information Retrieval and Language Technology Tools Awareness*

Finally, we asked the participants to share their experiences with any software programs they are familiar with that provide information or services in Swahili. The vast majority of participants stated unequivocally that they were unaware of any such system. A few of them mentioned Facebook and Google Swahili as programs that provide Swahili services.

Participants agreed that text processing and/or language technology tools are extremely important. The majority of participants believe Swahili technology is unavoidable. For example, when Swahili leaders attend international conferences, they require technology to help them interpret speeches and documents, such as contract documents.

Others believe that having translation technology, for example, will save time from visiting language experts for document translation and interpretation. Participants also stated that the majority of scientific and research information is

in English and that not everyone is proficient in translation. Thus, in order to deliver something to the public, one must rely on machines (Machine Translation (MT)) or visit Swahili specialists for a consultation, which is both time and money consuming.

One participant alluded to valuable documents left to rot on the shelves due to a lack of readers. He proposed that if they obtain automatic digitization and translation technology, they will be able to translate the majority of the information into Kiswahili, which will stimulate Swahili readers and help the knowledge reach many ordinary citizens.

Many participants were concerned about the accuracy of these automated text processing technologies and tools, citing Google Translate as an example of a system that does automatic translation. Their translations do not adhere to grammatical agreement.

> *"... it may translate correctly word for word, but it composes an incorrect sentence."*(06)

That is, Machine Translation (MT) services generate grammatically incorrect sentences and/or tenses while failing to recognize context. According to another participant:

> *"This one from Google [Translate] isn't very accurate; it sometimes gives completely incorrect translations. It simply introduces words that do not fit the context. For example, the actual translation of "security" as used in monetary institutions is* ***amana*** *[Deposit] or* ***dhamana*** *[Guarantee], not* ***ulinzi*** *as Google [Translate] translates."*[02]

Despite the shortcomings, participants were optimistic that the translation tools would be useful in providing an insightful view on the information; they could provide clues on a problem, allowing a human to simply do the editing and corrections. As one participant put it:

> *"I understand that these tools are not very precise, but even if they translate word for word, that is fine because editing and corrections will be done by a human."*[10]

4.3 DISCUSSION

The discussion of the results is based on the research objectives stated at the beginning of this chapter. Each of the following subsections corresponds to each of the research objectives.

4.3.1 *Search Language Preferences and the Associated Reasons*

In the first objective (**RO1**), we wanted to *"learn about search language preferences and the reasons behind them from the people being investigated."* The findings indicate that polyglot Swahili-speaking Web users use both English and Kiswahili to varying degrees. They appear to prefer searching in English rather than Kiswahili, but further discussion with the participants revealed that this statement is far too

broad; they normally code-switch. The study discovered that code-switching is influenced by several factors, including: the information need at hand; the topic and/or sector from which one wants to obtain information; the type of task one wants to complete; the context of the information one is looking for; and the searcher's competency in the language's terminology. Other considerations include the dependability of the results in a language and the speed (how quickly a searcher can get results).

Previous studies using different settings and populations, such as Aula and Kellar [10], Ling, Steichen, and Choulos [119] and Steichen and Lowe [197], identified some of these factors for code-switching in Web search. Our findings, despite the fact that we used only information experts, do not appear to differ from those of actual users.

4.3.2 *Experience of Swahili Query Formulation Efforts and the Associated Reasons*

Using **RO2**, we wanted to "*reveal the experience of Swahili query formulation efforts, and the associated reasons from the people being investigated and the community they serve.*" The majority of participants did not believe that formulating a Swahili query takes longer than creating an English query. The study discovered that a searcher's expertise, experience, and language proficiency are the primary determinants of query formulation effort. These findings are not surprising given that the majority of the participants were experts in information/library science. Participants who stated that they needed more time to formulate Swahili queries believed that this was due to the fact that they used more words than they did when forming English queries.

Tanzania's curriculum requires all public primary schools to use Kiswahili as a medium of instruction, and English as a medium of instruction at the secondary and tertiary levels [172], [205]. Because of this educational system, highly educated people are more likely to be fluent in English, making it easier for them to formulate English queries. People who are less educated may find it more difficult to formulate an English query and thus resort to phrasing their queries in Kiswahili.

In terms of whether they get satisfactory results from search engines, the majority of participants said they were frequently dissatisfied. They cited several reasons for the poor results, including a lack of Swahili documents on the Web, insufficient search techniques, query formulation skills, and limited vocabulary, outdated Swahili information on the Web, and poor support for other languages by search engines, i.e., IR systems designed to handle English and high resource language queries.

The issue of limited availability of Swahili documents, as raised by our participants, is difficult to resolve, at least from the perspective of IR. However, solutions to the remaining issues can be devised. This includes our suggested use of language preferences to increase the relevance of MLIR system results.

4.3.3 *TPreferences of Language of Information among Professionals and Ordinary Citizens*

In **RO3**, we attempted to "determine the preferences for information language among professionals and ordinary citizens using the opinions of those being investigated." The participants' perspectives on the use of English or Swahili information differ when the searcher is a professional in a particular field or an ordinary citizen. Participants stated that while most professionals may use both English and Kiswahili information, they prefer English. Professionals primarily use Kiswahili for "popular science", such as outreach and providing public services.

The participants unanimously agreed that ordinary citizens desperately needed Swahili information. One reason for this viewpoint is that most non-professional ordinary citizens struggle with English vocabulary but are proficient in Kiswahili. As a result, Swahili information is an obvious choice for them.

These findings may be related in part to Tanzania's education system, in which all high levels of education and profession training are conducted in English as a medium of instruction [172], [205]. As a result, it is understandable that professionals in a particular field would prefer English information over Swahili information. Ordinary citizens, on the other hand, who are mostly from the working class, may have a low level of education and thus require Swahili information for their daily activities. The same is true when searching for information on the Web.

4.3.4 *Demand for Swahili Information in Various Sectors*

The fourth research objective (**RO4**) sought to "*learn about citizens' needs for Swahili information in various sectors.*" Ordinary citizens control or interact with the majority of the sectors, such as agriculture and entrepreneurship, that our participants mentioned as requiring information in Kiswahili. According to the National Bureau of Statistics, the agricultural sector employs approximately 65% of Tanzania's workforce [192]. Their findings also show that the sector contributes 27.5% of GDP and 24.7% of foreign currency earnings from exports, respectively. This is consistent with the findings that the farming, agri-business, livestock husbandry, fish farming, and poultry farming sectors all require Swahili information, as this is where many non-professional ordinary citizens reside.

Justice, entrepreneurship, health, entertainment, banking, business, and communication and social network systems all interact with ordinary citizens in some way. These sectors, according to the participants, require information in Kiswahili. It should be noted that some of the sectors identified in our study correspond with those identified by Ngonyani [148], namely informal communication, worship, literature, politics, commerce, education, literature, administration, low level judiciary, and mass media.

Despite the fact that Swahili information is required in a variety of sectors (topics), some of which are critical to the economy, the participants believed that Web search engines returned unsatisfactory results in Kiswahili.

The need for Swahili information in some sectors (topics) requires a study that may effectively use the few Swahili resources (documents) available on the Web, supplemented by resources from other languages, primarily English. This may assist ordinary citizens, in particular, in consuming valuable information that they have been missing due to language barriers, such as difficulty in query formulation in English.

4.3.5 *Awareness of Swahili IR and Language Technology and Tools*

The last objective (**RO5**) sought to "*inquire about Swahili IR and language technology and tools awareness.*" Unfortunately, our study found that the majority of participants had no idea about systems and tools that can support Swahili IR or language technology. Only Machine Translation (MT) and its limitations were mentioned by a few participants. This calls for further research in Swahili NLP, IR and MT in order to develop better tools and systems.

4.3.6 *Limitations of the Study*

There were a number of constraints and restrictions in carrying out this study. The first is relying solely on information scientists and Kiswahili specialists to tell the story without balancing it with actual Swahili-speaking Web users. Using representatives may be more cost effective in terms of time, convenience, and money, but it may introduce a biased view on a subject in comparison to actual users.

Second, due to the difficulty in recruiting respondents with busy schedules, the study only used a few participants, resulting in a limited number of themes and codes from the thematic and open coding analysis. Of course, experts use the Web as a source of information too, but the small number used may not be sufficient to generalize well.

One might be interested to know whether the factors we identified, such as the topic of search, are actually causal factors, or if they are simply correlated due to other causal factors that also correlate with the topic, such as the availability of resources or the locality of information. Unfortunately, our investigation was unable to cover such a fascinating and multifaceted topic.

4.4 SUMMARY

This chapter presented a survey study of key informants in order to achieve the broader goal of developing a multilingual Swahili IR system that incorporates search behaviour, particularly language preferences. The interviews with information experts (librarians) and Kiswahili specialists shed light on the information needs and search habits of Swahili-speaking Web users in Tanzania.

We asked the participants about their preferred language for Web searches. They revealed that they use both English and Kiswahili in their searches, with an emphasis on English in the majority of cases. They do, however, dynamically

code-switch based on the search topic, task type, information context, and language proficiency.

The study discovered that the majority of participants do not believe that formulating a Swahili query differs from formulating an English query, i.e., the time to think and formulate a Swahili query is not different from doing the same in English. However, participants reported that their Web search yielded unsatisfactory Swahili results. Participants mentioned several reasons for poor Swahili results, including a lack of Swahili documents on the Web, poor search techniques, search engines' inability to handle Swahili queries, and a limited vocabulary. Due to a lack of vocabulary, users may create mixed-language queries.

The study also sought to ascertain the participants' perspectives on Swahili information usage among professionals and ordinary citizens. Participants demonstrated that, while professionals primarily use English, ordinary citizens prefer Kiswahili. Participants reported that many ordinary citizens are less educated and have a limited English vocabulary, making it difficult for them to conduct effective searches in English. Unlike less formally educated Swahili speakers, highly educated Swahili speakers successfully code-switch to English to find relevant information to meet their professional needs. Some well-educated Web users may use Kiswahili for non-professional information needs.

The study also revealed that there is a high demand for Swahili information on the Web, particularly among ordinary citizens and in many sectors such as agriculture and justice. Unfortunately, most ordinary people use English to search for information on the Web. Participants mentioned dissatisfying and irrelevant results in Kiswahili as factors motivating these Web users to use English.

Finally, the interviewees demonstrated a lack of knowledge about Swahili language technology and tools. They did, however, express a desire to see much of the search and access to information tasks automated.

The study presented in this chapter calls for additional research in two areas: one, the use of code-switching behaviour (language preferences) of Swahili search engine users in delivering more relevant results – this research attempts to contribute in this research area; and, two, making resources available for enabling Machine Translation (MT) and Natural Language Processing (NLP) – this is beyond the scope of this thesis.

5

ESTIMATING TOPIC-LANGUAGE PREFERENCES IN MULTILINGUAL SWAHILI INFORMATION RETRIEVAL

When searching for information on the Web, polyglots switch between languages. Studies on multilingual Web users, including the one presented in Chapter 4, suggest that part of the reason for such behaviour is the topic of search [10], [120], [217]. This chapter presents a carefully controlled study on a Multilingual Information Retrieval (MLIR) system as a follow-up to these works, particularly our study in Chapter 4.

The study in this chapter, however, used query and click-through logs from a guided Swahili MLIR system developed and presented to polyglot Swahili-speaking Web users, as opposed to survey-based studies. Thus, this study investigated the querying and results selection behaviour of Swahili-speaking MLIR system users, with the goal of determining how the topic of search (query) and language preferences are related. Associations can exist between the query topic and: i) the query language, and ii) the language of the results. These associations are referred to as *Topic-Language (T-L) associations/preferences* in this thesis.

Thus, the primary goal of the current chapter's research was to investigate the T-L association using query and click-through logs from a guided Swahili MLIR system. Polyglots (Swahili-speaking Web users) from Tanzania, an East African multilingual country where Kiswahili and English are both official languages, were used in the study. While a self-assessment questionnaire was used in a small portion of this study, the majority of it was a controlled study in which participants interacted with a guided multilingual search engine.

This chapter specifically attempts to answer the second research question (**RQ2**) of this thesis, which states that:

> *What are the topic-language preferences of the polyglot Swahili-speaking users of the multilingual Swahili Information Retrieval (IR) system?*

The study was guided by the following research objectives based on both the questionnaire and the users' interaction with the guided MLIR search engine:

RO1 To learn how Swahili-speaking Web users rate their use of English and Kiswahili for Web search.

RO2 To explore the preferred query language among Swahili-speaking MLIR system users.

RO3 To explore the preferred results language among Swahili-speaking MLIR system users.

RO4 To investigate the shift in topic-language preferences at various stages of MLIR searching.

The remainder of this chapter begins with an experimental setup, materials, and methods used in this study (Section 5.1). The findings are then presented in Section 5.2, followed by a summary and discussion session in Section 5.3. Finally, Section 5.4 concludes the chapter with a summary.

5.1 METHODOLOGY

5.1.1 *Development of the Topics and Queries Corpus*

This section describes how the search topics and queries used in this study were created. It is a standard practise that users of an IR system be allowed to create their own queries. Creating own queries aids in revealing actual information needs from real users. However, we controlled this step for several reasons as described below.

First, to save participants' time. As there was no monetary reward for taking part in this study, it was critical that they spend as little time as possible thinking of scenarios and queries to search from. There are several other studies in IR and MLIR, such as those by Ling, Steichen, and Figueira [120], Lowe and Steichen [129] and Yamamoto and Yamamoto [225], in which the authors prepared the tasks and topics or queries ahead of time. The prepared tasks and topics/queries demonstrate (simulated) information requirements.

Second, to avoid the problem of data skewness. By providing additional guidance to users and using a small number of users, it is possible to obtain a dataset that can be assumed to be representative. Thus, it is possible to get feedback on a wide range of topics using a small group of users, as opposed to asking the same small group of users to search for whatever topics they wanted.

Third, to avoid the machine translation problem. MLIR system requires translation of the query to the languages of the documents intended. Peters, Braschler, and Clough [158] emphasizes the importance of (machine) translation (MT) in achieving MLIR. However, this is not obviously done with MT for settings where there is an under-resourced languages, such as Kiswahili is involved. MT for low resource languages is far from perfect, in part due to insufficient parallel corpora [103].

Allowing users to create queries on the fly would thus reduce system robustness due to translation errors affecting retrieval results in one language, and thus imbalance the language preferences investigated in this study.

Tanzania, an East African country, was chosen because of its large Swahili-speaking population. The country's multilingualism, with Kiswahili and English as official languages, ensures that Web searchers will use both languages. All of the prepared topics and queries had a connection to Tanzanian Web searchers or originated in Tanzania.

To identify various topics on the Web, Web directories specific to Tanzanian websites and Google Trends[1] were used. Web directories categorize websites based on major themes (topics) and sub-themes (sub-topics) of the information they contain. A tourism theme, for example, may have websites such as tourism.

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

Alexa[2], 123Tanzania[3], Yalwa[4], and (the deprecated) WWW Virtual Library[5] were the Web directories used in this study because they had a good coverage of Tanzanian websites. Google Trends provides a high-level view of trending queries on Google search engine.

To ensure that all queries were in the geographical region of Tanzania and that the topics were not only mentioned in the Web directories but also used in the region under study, the Google Trends Explore[6] system was configured as follows: category – Web search; location – Tanzania; and duration – 2004 to 2019. Then, we ran each of the Web directories' topics through Google Trends Explore to identify the related queries in each of the topics.

We exported the topics and their associated queries from the Google Trends Explore system as comma-separated values (CSV) files, which we then combined into a single file. All single-word queries were removed due to ambiguity (word sense) [93], [177], even when translated. For example, a *apple* query in English may imply a *fruit* or a *technology company* company, but it may only translate to *tufaa* in Kiswahili. This translation only takes into account the fruit's meaning, leaving out relevant results related to the technology company's information.

We kept only topics with at least two queries, removing those with a single query because they were deemed less important to the community under study. We merged all of the queries that were related to the same topic but had slight differences in spelling, pre- and post-fixes, and information requirements. This resulted in 1184 queries covering 123 different topics.

Because we used Bing Web Search API[7] to retrieve results from the Web in the guided Swahili MLIR system, we primarily used Bing MS Translator[8] to translate all queries between the two languages. We translated queries from Kiswahili to English and vice versa. In cases where there was term ambiguity or lexical-semantic issues, we also used Google Translate[9] for verification and/or as an alternative.

5.1.2 *Data Collection Platform*

As a data collection platform, the guided Swahili MLIR system had three major sections: *demographics*, *topic and queries system*, and *search engine interface*. Before they could access the demographics page, each participant had to sign a consent form on the platform's index page[10]. The system then redirected the user to a page that requested some personal demographic information, such as gender, age group, education level, and occupation. On the same page, the system asked users to rate their use of English and Kiswahili when searching the web for information. Users could rate themselves on a scale of 1 to 5 for each language

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
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/

(Kiswahili and English), with 1 representing never, 2 representing rarely, 3 representing sometimes, 4 representing frequently, and 5 representing always using the language for Web search.

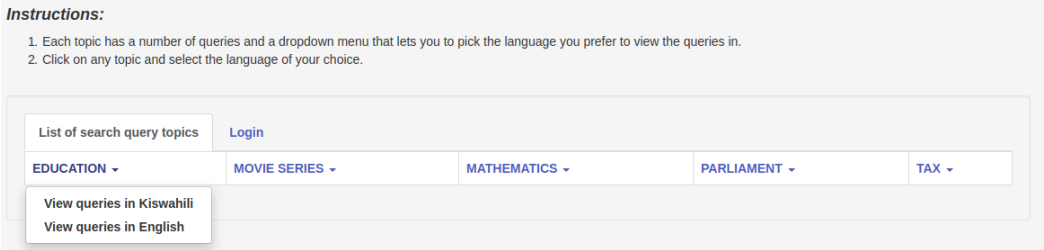


Figure 5.1: The topics interface allows users to select a topic and query language.

The topic and queries system section was divided into two pages: *topics* and *queries*. The system generated five randomly generated topics from the 123 topics mentioned above for each user (see Figure 5.1 for an excerpt of sample topics). A random display of five topics per user session ensured that users were not overburdened or confused by all 123 topics and that each topic had an equal chance of being selected. The system then instructed the user to select a topic of interest from the displayed topics, and then to select the language for viewing queries from a drop-down list attached to each of the displayed topics. The system only supported two languages: English and Kiswahili. The language chosen is referred to as a *query language*.

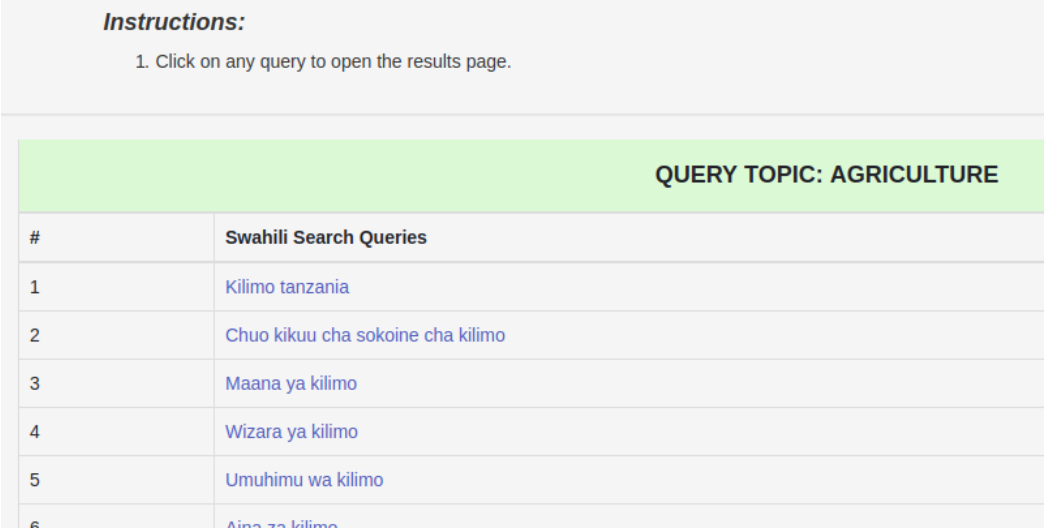


Figure 5.2: The queries interface allows users to select from the available queries by clicking on them.

After selecting the preferred query language, the queries page appeared, displaying only queries in the user's selected topic and language. To avoid the bias of users selecting only short and/or easy-to-type queries, each query included an embedded clickable link to the search engine results interface. The system instructed users to open the results page by clicking on a query of their choice

(search engine interface). For an example of displayed Swahili queries on the Agriculture topic, see Figure 5.2.

The platform's final component was the search engine results page, which used the MS Bing Web Search Application Programming Interface (API) to retrieve (multilingual) results. Regardless of the query language specified by the user when selecting the query, the system presented the search results in an interleaved round-robin style in both Kiswahili and English. An excerpt of the displayed results can be found at Figure 5.3. To counteract the effect of the position bias of results in one language always appearing first, the system randomly alternated languages of the results appearing first in each session.

Despite the fact that Ling, Steichen, and Choulos's [119] findings on multilingual display/interface preferences suggest that the panel style was mostly preferred by MLIR users, we chose not to use it for two reasons. For starters, the participants were assumed to be new to MLIR, so they are used to the monolingual style of results presentation; the look and feel of a monolingual search engine is appealing to such participants. As a result, the participants can make decisions based on the results' language rather than the layout.

Second, the use of panels has the potential to introduce layout bias. Many Web users, for example, may be drawn to the results on the left panel (the first to appear) and ignore or pay less attention to those on the right, viewing the latter as extra/additional. Consider the study by Jimmy, Zuccon, Koopman, *et al.* [96] on the use of health cards, which are displayed on the right panel of the results snippets, which discovered that users spent the majority of their time on the snippets (left panel) rather than the health cards (right panel).

Figure 5.3: Search results layout displaying an example of the Search Engine Results Page (SERP).

As shown in Figure 5.3, we concealed the search bar to prevent users from modifying or creating their own queries. We also disabled the clickable links in the results titles, so users could not navigate to a specific result or open multiple tabs to find the most relevant results. Instead, users could inspect the results based on the snippets and determine whether or not a result is relevant. Relevance judgments are frequently derived or inferred from snippets in click log data. Most studies in click models, for example, Chuklin, Markov, and Rijke [33],

and Grotov, Chuklin, Markov, *et al.* [74], as well as learning to rank, for example, Liu [123], Joachims, Swaminathan, and Schnabel [98] and Joachims, Granka, Pan, *et al.* [97], rely on relevance judgments via clicks due to snippets examination.

There was a checkbox to the left of each result that allowed users to show/check (✓) the most relevant results. This saves users time by allowing them to skim through all of the results, and complete as many search sessions as possible. After selecting all relevant or the first most relevant result(s), the user must click the submit button at the bottom of the results page. The chosen results (URLs) and search query were saved on a server in a file. This data represented the click-through log data used in this research's analysis. Aside from this data, all responses from the Bing Web Search API calls were saved to a separate file on the server.

The platform's final page included an acknowledgement message thanking a user for taking part in the research. The page also included a message asking users who still had time to repeat the search process with different randomly generated topics to do so. Participants could follow these simple procedures flawlessly, and some could perform several iterations of search on their own and complete in less than 10 minutes.

5.1.3 *Participant Recruitment*

The recruitment process targeted two groups of participants: Tanzanian Web users (non-students) and university students, specifically from Sokoine University of Agriculture (SUA), Tanzania. The general Web users were recruited using: social media such as WhatsApp messenger groups and individuals in a snowball fashion, LinkedIn, Instagram, and JamiiForums[11]; and mailing lists of Tanzanian organizations and universities, with the assistance of friends and colleagues working there.

We recruited student participants with the help of lecturers and Class Representatives (CRs) after obtaining research clearance from SUA[12]. The invitation message was sent to the CRs via WhatsApp messenger. To deliver the message, the CRs used WhatsApp class groups, a popular form of communication among students. Non-student participants interacted with our data collection platform (guided MLIR system) using their own gadgets such as smartphones and computers, while student participants interacted with our system using university computer laboratories. The invitation message, written in both English and Kiswahili, included a link to the data collection platform (URL). After signing the informed consent form on the platform's homepage, the system granted permission for the study.

5.1.4 *Data set Description*

The data collection period lasted three months, from November 6, 2019 to February 5, 2020. The system logged both queries and click-through data from user

11 https://www.jamiiforums.com/ – Tanzania's leading social networking forum/website

12 Sokoine University of Agriculture (SUA) Staff, Students and Researchers Clearance, Ref. Number SUA/DRPSG/R/126/3/97

interaction with the guided MLIR system, as explained in the data collection platform in Section 5.1.2 above. In addition, we gathered user ratings on the use of English and Kiswahili in Web searches, as well as demographic data.

The first (query) data set contained 2387 query records derived from user interactions with the 123 topics mentioned above in Section 5.1.1. Each query record contained the query as well as the topic to which it belonged. In Kiswahili, the distribution of query records per topic of search was as follows: minimum – 0, maximum – 28, and average – 9 queries. The distribution of queries per topic of search in English was as follows: minimum – 0, maximum – 23, and average – 7 queries. We calculated the aggregate counts (frequency) of query records per topic of search for each language. A Law topic, for example, had 21 and 7 query records in Kiswahili and English, respectively.

For the sake of brevity and demonstration, we grouped related topics into large groups of topics known as *super-topics*. *Computer*, *Hardware*, *Internet*, *Phones*, *Software*, *Telecommunications* and *Television*, for example, were all grouped into the *Information Technology (IT) and Electronics* super-topic. As shown in the Supplemental Materials Table B.1, the grouping produced 19 super-topics from the original 123 topics.

The second data set (click-through data) contained 3157 click-through records (or clicked URLs). The term click-through data should not be confused with data collected by clicking; rather, it refers to data that participants indicated as relevant or not via check-boxes, as explained in Section 5.1.2. Every click-through record had between 0 and 10 relevant clicked results (URLs). We removed the records in which neither Kiswahili nor English were clicked on URLs, i.e., records in which users did not find relevant results in both languages and did not click on any of the results. Each record has three columns: a language identifier, a query, and a list of URLs that have been clicked. By looking up the original topic and query corpus, we were able to associate each query with its corresponding topic.

5.1.5 *Data Analysis*

We used descriptive statistics to examine demographic information as well as user ratings on the use of Kiswahili and English when searching for information on the Web. While MS Excel was used for descriptive and exploratory statistical analyses, an online test tool[13] was used for hypothesis testing (Mann-Whitney test). The subsections that follow describe the analysis of the query and click-through logs from the user's interaction with the guided MLIR system.

5.1.6 *Estimating Query Language Preferences*

Remember that the guided MLIR system presented each user with a list of five topics and a drop-down menu for selecting the language in which to view the queries. The study regarded the language chosen as a preferred query language over the other. Because there were only two language options, the system forced users to select one of them, a practice known as a forced-choice paired preference

13 http://www.statskingdom.com/170median_mann_whitney.html *[Accessed on 30 August 2020]*

test [110], [138]. A preference test is defined by Meilgaard, Civille, and Carr [138] as one in which a respondent is forced to choose one item over another or others. The following sub-objectives were achieved by using this test to address **RO2**:

1. To estimate the overall preferred query language.
2. To estimate the preferred query language in super-topics.
3. To estimate the preferred query language in topics.

The tests were one-tailed, so that they can show that one language was preferred over the other. To avoid erroneously concluding that a preference exists, the sensitivity values α, β and P_{max} must be adjusted differently to address each of the above objectives. It should be noted that α (or α-risk) is the probability of concluding that there is a preference when, in fact, there is not (Type I error) and β (or β-risk) is the probability of concluding that there is no preference when there is (Type II error).

Meilgaard, Civille, and Carr [138] 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 in detecting a 70:30 split in preferences, $P_{max} = 70\%$ and $\beta = 0.10$. According to the rule of thumb, if $P_{max} < 55\%$, $55\% \leqslant P_{max} \leqslant 65\%$ and $P_{max} > 65\%$, there is a small, medium, and large deviation from equal intensity, respectively.

Using the formula by Meilgaard, Civille, and Carr [138]: the values of α can be calculated based on the number of responses n and the minimum number of common responses x, such that:

$$\alpha = 1 - \text{BINOMDIST}(x - 1, n, P_0, 1) \tag{5.1}$$

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

$$\beta = \text{BINOMDIST}(x - 1, n, P_{max}, 1) \tag{5.2}$$

The P_{max} is calculated using the probability of common guess (P_0 – probability that a random guess will result in a significant difference between the two objects) and the proportion of distinguishers (P_d – maximum number of users/population who distinguish between two objects) such that:

$$P_{max} = P_d + P_0(1 - P_d) \tag{5.3}$$

Setting P_0 and P_d to 0.5 each results in a $P_{max} = 75\%$. This value is large enough to ensure that user preferences are clearly divided between English and Kiswahili. Some topics, in particular, had higher values of β than the maximum desired (i.e., 0.20), and were thus omitted to avoid large Type II errors. This was caused in part by a small number of respondents/responses in those topics. As a result, only 47 of the 123 topics were eligible for analysis.

The minimum number of common responses x required to conclude that there is a significant difference between the two objects involved, can be calculated as follows:

$$x = (n/2) + z\sqrt{n/4} \tag{5.4}$$

Where n is the total number of responses in a super-topic or topic, and $z = 1.645$ is the significance level for a one-tailed test with $n \leqslant 30$. For $n > 30$, different values of z were obtained from Table 17.3 in Meilgaard, Civille, and Carr [138]. If the observed number of common responses c is greater or equal to x, then there is a preference for a language with common responses and no preference otherwise.

5.1.7 *Estimating Preferences for Language of Results*

We use the previously discussed method for estimating query language preferences to estimate the preferred language of results, treating each URL as an independent choice/response from a user. Then, using the sub-objectives listed below, we attempted to address the third research objective (**RO3**).

1. To estimate the overall preferred language of results.
2. To estimate the preferred language of results in super-topics.
3. To estimate the preferred language of results in topics.

To avoid committing large Type II errors, we set $\texttt{beta} \leqslant 0.20$. The analysis excluded topics with higher values of β than the desired value, i.e., $\beta > 0.20$. 66 of the 123 topics were eligible for analysis. It should be noted that a 1% margin of error was allowed to accommodate some topics with (beta)-values that were closer to the desired value.

5.2 RESULTS

5.2.1 *Demographic Information*

The experiment included 676 participants, 65.1% of whom were male, 34.5% of whom were female, and 0.4% who did not disclose their gender (Figure 5.4a). The majority of participants were young and middle-aged, aged 18-24 (40.4%), 25-34 (42.6%), and 35-44 (12.4%). Only 4.6% of participants were over the age of 45, as shown in Figure 5.4b.

As shown in Figure 5.4c, the majority of the participants had (or were pursuing) tertiary level of education: 63.0% with bachelor's degree, 13.0% with diploma, 10.7% with master's degree, 5.2% with doctorate degree, and 1.2% with professional degree, such as veterinary medicine doctor. While 0.3% had a certificate level of education, 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).

In terms of occupation (Figure 5.4d), 46.6% were students, 35.2% were employed, and 11.2% were self-employed, entrepreneurs, or business-people. Meanwhile, 5.6% of participants were unemployed but looking for work, and 1.0% were unemployed but not looking for work. There were 0.3% of participants who were retired, and none who were disabled to the point of being unable to work.

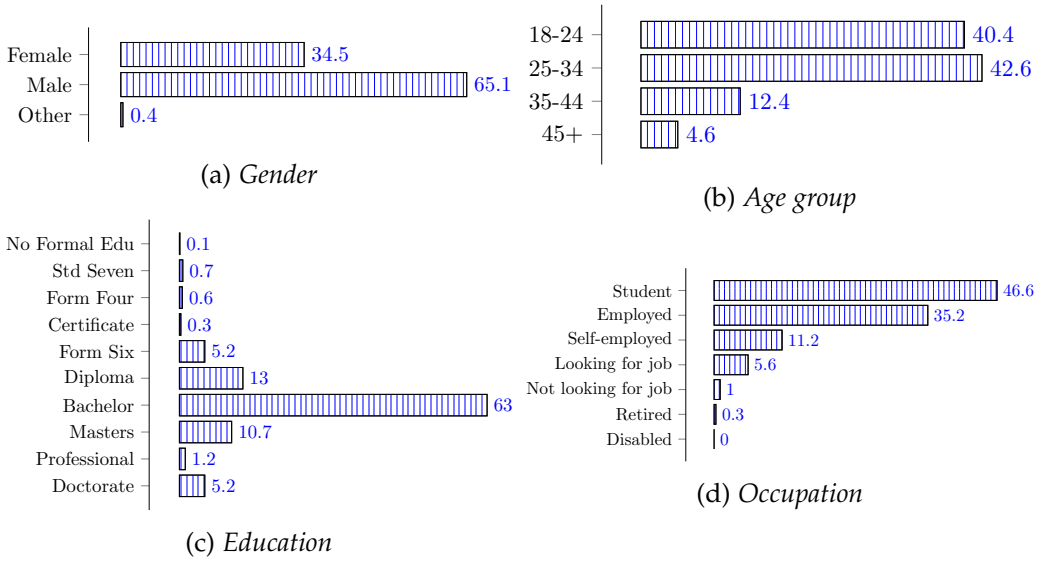


Figure 5.4: Demographics information of our multilingual guided search engine (**N=676**).

Standard (Std) Seven represents the highest level of primary school education in Figure 5.4c, while Form Four and Form Six represent the ordinary and advanced levels of secondary (high) school education, respectively.

5.2.2 *The Use of English and Kiswahili in Web Searches*

The self-assessment of participants' use of English and Kiswahili in Web search reveals that the majority of our participants rated themselves as "always" (41.9%), "often" (28.3%), and "sometimes" (25.1%) using 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 5.5). 39.9 percent of participants rated themselves as "sometimes" using Kiswahili in their Web searches. 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 searches.

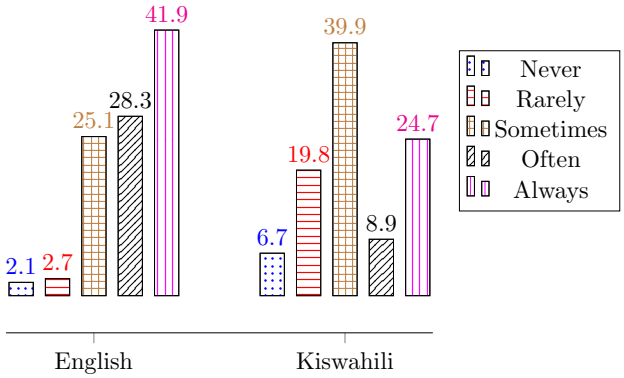


Figure 5.5: Participants ratings on their language use (Kiswahili and English) on the Web search **(N=676)**.

Participants rated Kiswahili and English use in Web search with a mode of 3 and 5, respectively, according to the descriptive statistics in Table 5.1. These modes may imply that participants occasionally use Kiswahili while always use English to search the Web for information.

Table 5.1: Ratings of our participants on their language use on the Web search **(N=676)**.

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

We run a statistical test under H_0*: there is a difference in the medians of the ratings for English and Swahili use*. Using a Mann-Whitney U test with $\alpha = 0.05$ to compare the ratings of Kiswahili and English, the null hypothesis H_0 is rejected. This means that the difference in median English and Swahili usage is statistically significant (U=142098.5, p=0.0000). According to the findings, English is significantly preferred as a language for Web search.

5.2.3 *User Interaction with the Topic and Queries System*

Figure 5.6 shows how users interact with the topic and queries system. While some searched only once, presumably in a single topic, others searched multiple times in different topics, either using the same language across all topics or switching between English and Kiswahili as the topics changed. 23.6% and 17.7% of users searched in a single topic using only Kiswahili and English, respectively.

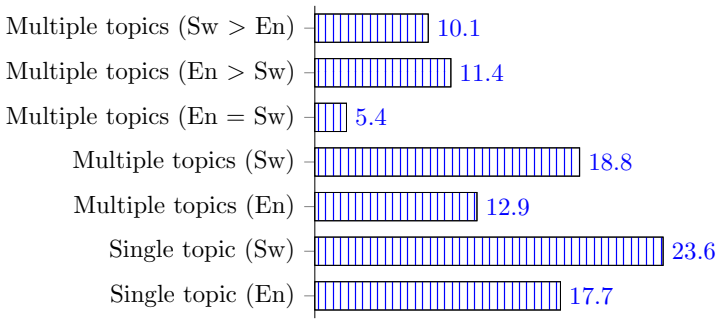


Figure 5.6: User behaviour when interacting with topics and queries, i.e., how many topics did they select and in what language, for example, 17.7% of users selected only one topic and used English as a query language.

Users who searched in multiple topics were divided into two groups: i) those who used one language regardless of topic – 18.8% for Kiswahili and 12.9% for English; and ii) those who used both languages across topics. The latter are classified into three categories: a) those who use both languages equally (5.4%), e.g., 2 topics in English and 2 topics in Kiswahili; b) those who use more topics in English than Kiswahili (11.4%), e.g., 5 topics in English and 2 topics in Kiswahili;

and c) those who use more topics in Kiswahili than English (10.1%), e.g., 7 topics in Kiswahili and 3 topics in English.

The aggregated counts of responses per super-topic are represented by the bar chart in Figure 5.7. The number of queries in a particular language is represented by the responses. The figure shows that the bars in each super-topic are not equal, indicating that English and Kiswahili were used in different super-topics. The figure shows that Swahili bars are taller than English bars in the majority of super-topics. These visual observations can be used to formulate three sub-questions, which are addressed in the following subsections.

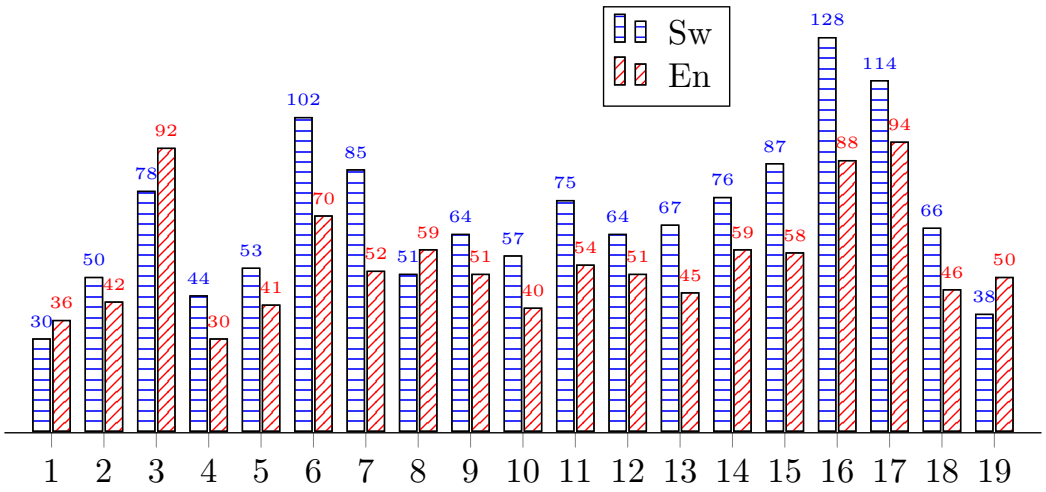


Figure 5.7: Frequency/counts of query language vs super-topic of search i.e., how many times a topic, say, Justice was searched for in English?

Where 1 = Religious Faith, 2 = Higher education, 3 = IT and Electronics, 4 = Justice, 5 = Tourism, 6 = Health and Facility, 7 = Education, 8 = Earth and Environment, 9 = Human Resource(s) Management (HRM) and Training, 10 = Lifestyle, 11 = Agriculture and Food, 12 = Transportation, 13 = Business, 14 = Economic development, 15 = Society and Culture, 16 = Sports and Entertainment, 17 = Government, 18 = Family and gender, and 19 = Engineering and Construction super-topic.

5.2.3.1 *What is the generic preferred query language?*

There were 2387 responses across all topics, with 1329 and 1058 in favor of Kiswahili and English, respectively. Using Equation 5.4, a minimum of 1250 common responses were required to conclude that one of the languages is preferred. Because Kiswahili received 1329 responses to English's 1058, it can be concluded that there was a significant preference for Kiswahili as a generic query language, at the calculated (using Equation 5.1 – 5.3) sensitivity values of $\alpha = 0.01$, $\beta = 0.00$ and $P_{max} = 75\%$.

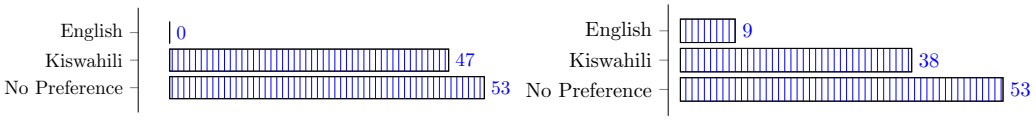
5.2.3.2 *What is the preferred query language in super-topics?*

The Supplemental Materials Table B.2 (Appendix B) detail the tests for query language preferences in all the 19 super-topics, tested at different sensitivity values using Equation 5.1 – 5.3 such that: $0.04 \leqslant \alpha \leqslant 0.08$; $0.0000 \leqslant \beta \leqslant 0.0033$; and $P_{max} = 0.75$.

In 9 of the 19 super-topics, there was a statistically significant preference for Kiswahili as a query language (47%, Figure 5.8a). In the remaining 10 super-topics, there was no preference (ties) for language (53%, Figure 5.8a). Table 5.2 displays the list of Swahili preferred super-topics as well as super-topics with no preference for query language.

Table 5.2: A list of super-topics that are preferred in Kiswahili and those that do not have a preference for query language.

Swahili Preferred	**No Preference**
Justice, Health and Facilities, Education, Lifestyle, Agriculture and Food, Business, Society and Culture, Sports and Entertainment, and Family and Gender.	Religious faith, High education, IT and Electronics, Tourism, Earth and Environment, HRM and Training, Transportation, Economic development, Governance, and Engineering and construction.



(a) (b)

Figure 5.8: Proportion of query language preferences in: (a) 19 super-topics; and (b) 47 topics.

In contrast to the visual observation in the bar chart (Figure 5.7), where the English bars were marginally taller than the Swahili bars in four super-topics, English is not a significantly preferred query language in any of the super-topics. Kiswahili is significantly preferred as a query language in only 9 super-topics, where Swahili bars are taller than English counterparts. Users used Kiswahili and English as query languages equally because visual differences in the sizes of the bars in 10 super-topics are not statistically significant to conclude preferences for any language.

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

The details of the tests for 47 topics that passed the requirement of a minimum number of responses and the sensitivity values can be found in Appendix B, Supplemental Materials Table B.3. These values were calculated using Equation 5.1 – 5.3 such that: $0.03 \leqslant \alpha \leqslant 0.09$; $0.02 \leqslant \beta \leqslant 0.21$; and $P_{max} = 0.75$.

There was a significant preference for Kiswahili as a query language in 18 topics (38%, Figure 5.8b). Only 4 topics (9%, Figure 5.8b) had English as a significantly preferred query language. There was no preference for language in 25 topics (53%, Figure 5.8b), including phones, the environment, training, and fashion. Table 5.3 presents a list of topics that were significantly preferred in Kiswahili and English for query language.

Table 5.3: A list of topics that are preferred in English and Kiswahili for query language.

Swahili Preferred	**English Preferred**
Clinic, School, Law, University, National park, Christianity, Management, HIV/Aids, Waste management, Development, Industry, Society, Award, Music, Photographs, Female, Education, and Agriculture.	Religion, Computer, Heart, and Water.

5.2.4 *User Interaction with the Results Page*

The SERP displayed 20 results per search query, with each language receiving an equal share of the results, i.e., 10 each. In both English and Kiswahili, each query had a minimum of 0 to a maximum of 10 relevant clicked results (URLs).

A descriptive statistical summary in Table 5.4 from 809 sessions shows that the mean number of URLs clicked per query is 2.14 for English and 1.77 for Kiswahili, respectively. Both English and Kiswahili had modes and medians of 1 URL per query. The modes of 1 indicate that the majority of queries had at least one clicked relevant answer in both English and Kiswahili.

Table 5.4: Query-URL 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

Users' interaction with the SERP can be divided into two categories: users who clicked on results in only one language (English (35.6%) and Kiswahili (24.1%); and users who clicked on results in both languages. In this group, some users clicked on an equal number of results in both languages (10%), while others clicked on more results in English (15.9%) and others on more results in Kiswahili (14.3%) (Figure 5.9).

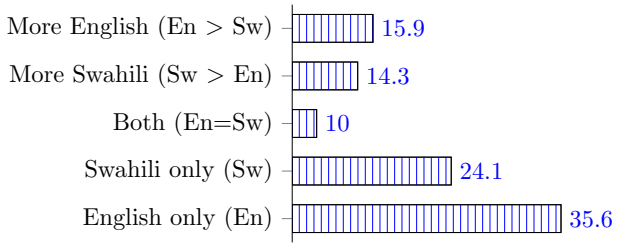


Figure 5.9: User behaviour when interacting with the SERP, i.e. the proportion of users who wanted results in only one language or both languages.

In Figure 5.10, the frequencies of clicked URLs/results are plotted. Visually, the length of the bars varies within super-topics, which implies the same in topics, leading to three sub-questions addressed in the following subsections.

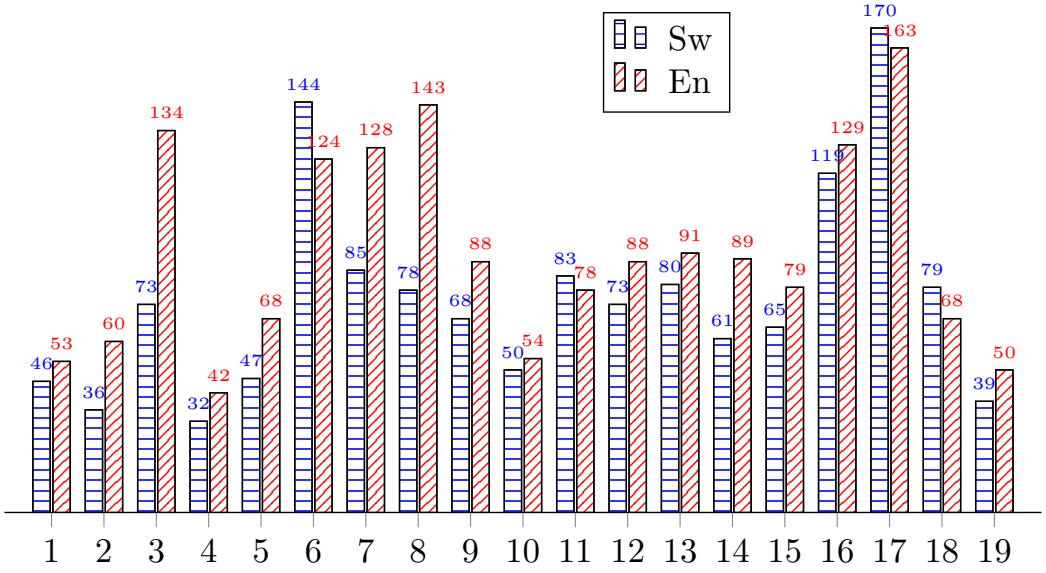


Figure 5.10: Total URLs/results clicked in each language vs super-topic of search results.

Where 1 = Religious Faith, 2 = Higher education, 3 = IT and Electronics, 4 = Justice, 5 = Tourism, 6 = Health and Facility, 7 = Education, 8 = Earth and Environment, 9 = HRM and Training, 10 = Lifestyle, 11 = Agriculture and Food, 12 = Transportation, 13 = Business, 14 = Economic development, 15 = Society and Culture, 16 = Sports and Entertainment, 17 = Government, 18 = Family and gender, and 19 = Engineering and Construction super-topic.

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

There were 3157 URLs/results clicked, with 1729 in English and 1428 in Kiswahili. Using Equation 5.4, a minimum of 1644 common clicked results (responses) were required to conclude that one language is preferred over the other. Given the calculated (using Equation 5.1 – 5.3) sensitivity values of $\alpha = 0.01$, $\beta = 0.00$ and $P_{max} = 75\%$, there is a generic significant preference for English as a language of results.

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

Refer to Table B.4 in the Supplementary Materials (Appendix B) for the details of the preference tests in each super-topic at different sensitivity values calculated using Equation 5.1 – 5.3 such that: $0.04 \leqslant \alpha \leqslant 0.06$; $0.0000 \leqslant \beta \leqslant 0.0024$; and $P_{max} = 0.75$.

The findings show that there was no significant preference for language in 12 (63%, Figure 5.11a) super-topics. In the remaining 7 (37%, Figure 5.11a) super-topics, English was significantly preferred as a language of results. The super-topics that were significantly preferred in English as a results language and those that had no preference for results language are tabulated in Table 5.5.

Table 5.5: A list of super-topics that are preferred in English and those that do not have a preference for results language.

No Preference	English Preferred
Agriculture and Food, Business, Society and Culture, Sports and Entertainment, and Family and Gender, Justice, Health and Facilities, Religious faith, Lifestyle, High education, Transportation, Governance, and Engineering and construction	Higher education, IT and Electronics, Tourism, Education, Earth and Environment, HRM and Training, Economic development

Kiswahili was not significantly preferred in any of the super-topics, in contrast to the visual observation in Figure 5.10, where Kiswahili was marginally preferred to English as a query language in four super-topics: Health and facility, Agriculture and Food, Government, and Family and Gender.

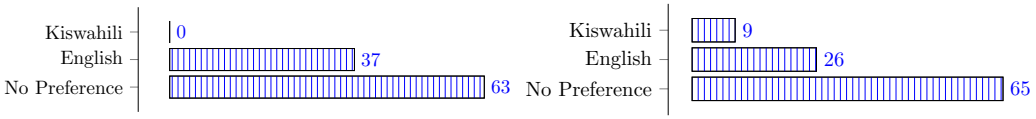


Figure 5.11: Proportion of preferences of language of results for: (a) 19 super-topics; and (b) 66 topics.

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

In the Supplementary Materials (Appendix B), Table B.5 describes the tests for preference for the language of results in 66 topics that passed the sensitivity values criteria. The values were calculated using Equation 5.1 – 5.3 as follows: $0.03 \leqslant \alpha \leqslant 0.09$; $0.00 \leqslant \beta \leqslant 0.21$; and $P_{max} = 0.75$.

Table 5.6: A list of topics that are preferred in English and Kiswahili for results language.

English Preferred	Swahili Preferred
University admission, Scholarship, Law, Energy, Chemistry, Computer, Hardware, Heart, Environment, Water, Training, Food, Transport, Accounting, Industry, Social, and Government	HIV/Aids, Livestock, News, Music, Railway, and Election

The findings show that there was no significant preference for language of results in 43 topics (65%, Figure 5.11b). In 17 topics (26%, Figure 5.11b), there was a significant preference for English as a language of results. The findings

also show that Kiswahili was significantly preferred as a language of results in 6 topics (9%, Figure 5.11b). Table 5.6 presents a list of topics that were significantly preferred in English and Kiswahili for results language.

5.2.5 *Language Preferences Change*

According to Figures 5.8 and 5.11, there were changes in users' language preferences during search at the stage of querying the system (query language preference) and results selection (results language preference). Table 5.7 displays the percentage of super-topics and topics whose preferred languages have shifted. It should be noted that only 36 topics with a perfect match or alignment are presented in this work, i.e., topics that existed in both the analysis of query language preferences (Section 5.2.3.3) and language of results preferences (Section 5.2.4.3).

The preferred language shifted from Kiswahili to English in only 1 super-topic (5%). In 8 super-topics (42%), the preferred language shifted from Kiswahili to no preference. Furthermore, the preferred language shifted from no preference to English in 6 super-topics (32%). There was no change in 4 super-topics (21%) that had no preference for language.

The language preference changed from Kiswahili to English in only 1 topic (3%). In 9 topics (25%), the preferred language shifted from Kiswahili to no preference. The shift in language preference from English to no preference occurred for 1 topic (3%). The same percentage of change, 3%, occurred for no preference to Kiswahili. The switch from no preference to English occurred in 6 topics (17%). While the preference for Kiswahili remained the same in two topics (6%), the preference for English remained the same in 3 topics (8%). Furthermore, there were no change in 13 topics (36%) that had no preference for language.

5.3 DISCUSSION

This section discusses the findings; the discussion is primarily based on the stated objectives stated in chapter's overview.

5.3.1 *Swahili Speakers' Perceptions of the Use of English and Kiswahili for Web Searches*

In the first objective (**RO1**), we wanted to "*learn how Swahili-speaking Web users rate their use of English and Kiswahili for Web search*". The findings indicate that the two languages are used in a mixed manner when conducting Web searches. However, the majority of users indicated that they always and occasionally use English and Kiswahili, respectively.

There could be a number of reasons why users rated themselves as always using English in Web search. There were no follow-up questions to delve into the reasons, as that was beyond the scope of this study. However, the massive amount of information available in English on the Web compared to a low-resource language (Kiswahili) may be part of the reason. Wikipedia, for example, had 6,317,055 English articles as of June 14, 2021, compared to only 63,301 Swahili articles [220]. Tanzania uses English as a medium of instruction from sec-

Table 5.7: Changes in language preferences in both super-topics (N=19) and topics (N=36) from query to results languages.

Change	%	Super-topics	%	Topics
sw → np	42	Justice, Health and Facility, Lifestyle, Agriculture and Food, Business, Society and Culture, Sports and Entertainment, Family and Gender	25	University, Clinic, National Park, Education, Development, Award, Industry, Waste Management, Agriculture.
np → en	32	Higher education, IT and Electronics, Tourism, Earth and Environment, HRM and Training, and Economic development	17	University admission, Computer hardware, Environment, Training, Food, Social.
np → np	21	Religious faith, Transportation, Government, and Engineering and Construction	36	Christianity, Software, Phones, Mathematics, School, Human resource, Fashion, Movies series, Culture, Public service, Government, Family, Child.
sw → en	5	Education	3	Law.
en → en	0		8	Computer, Heart, Water.
sw → sw	0		6	HIV/AIDS, Music.
en → no	0		3	Religion.
np → sw	0		3	Election.
en → sw	0		0	

where sw – Kiswahili, en – English and np – no preference.

ondary school through university, including all vocational colleges [172], [205]. This implies that the majority of Web users in Tanzania received their training/education in English, and they are more likely to seek professional information in English and non-professional information in Kiswahili, as revealed by our study presented in Chapter 4.

5.3.2 *Preferred Query Language*

In the second objective **RO2**, we wanted to "*explore the preferred query language among Swahili-speaking MLIR system users*". According to the results of the query log analysis, users either had no preference for language or preferred Kiswahili as a query language. Kiswahili was a significantly preferred query language in super-topics and in roughly one-third of the topics, almost equal to the proportion of "no preference". On a smaller scale, English was significantly preferred, particularly in topics.

According to these findings, the smaller the topic granularity, the clearer the query language preference. We discovered that, while statistically significant, generalizing that Kiswahili is an overall preferred query language may be misleading, as no preference (ties) or English had a fair share of preference in super-topics and topics. For example, at the topic level, 53% of topics had no preference for query language, while 9% of topics had a significant preference for English.

The term "language use in Web search", as used in the questionnaire section of the study, was broad in the sense that it could refer to the language used to query the Web (query language) or the language of search engine results. Thus, these findings may be related to the "occasional" use of Kiswahili found in user opinion in Section 5.3.1, as the query language preference is biased towards Kiswahili or no preference, as opposed to English.

5.3.3 *Preferred Results Language*

Our third objective (**RO3**) attempted to "*explore the preferred language of results among Swahili-speaking MLIR system users*". The findings show that at a higher generic level, English was the significantly preferred language of results, but this was not the case as the level of granularity changed to super-topics and topics.

In contrast to the querying behaviour of the same users described in the previous section, the findings reveal that English results appear to be significantly preferred in more than one-third of the super-topics and roughly one-quarter of the topics. The majority of preferences were "no preference" (ties), with both Kiswahili and English having an equal chance of being used as the language of results. On the other hand, Kiswahili was not significantly preferred in any of the super-topics and was only significantly preferred in a few topics.

5.3.4 *Shift in Language Preferences*

In the last objective (**RO4**), we wanted to "*investigate the shift in topic-language preferences at various stages of MLIR searching*". The findings show that language preferences shift during the querying and results selection stages. For super-topics and topics with preferences, there was a significant shift from Kiswahili as a preferred query language to English as a results language. For example, while users significantly preferred English as a query language in only 9% of the topics, this proportion increased to 26% for the results. Kiswahili, on the other hand, which was significantly preferred as a query language in 38% of the topics, dropped to only 9% as the results language.

The high proportion of Swahili preference as query language demonstrates that Swahili-speaking Web users prefer their native language over English. This is consistent with several studies, including Wang and Komlodi [216], Lowe and Steichen [129] and Vassilakaki, Garoufallou, Johnson, *et al.* [210], which discovered that Web users prefer or prioritize searching in languages they are more familiar with.

The justification for the major shifts in language preferences from query to results language is outside the scope of this work. However, it could be related to the problem of a lack of online documents in a low resource language (Kiswahili), as reported in Chapter 4. and several other studies involving nearly identical settings, such as Kralisch and Mandl [109], Aula and Kellar [10] and Wang and Komlodi [216]. Other reasons could be circumstantial, such as document availability (quantity) [10], [59], [216] and quality of documents [10], [59], [128].

It is also worth noting that, despite the scarcity of online Swahili documents, users were able to find the information they needed to meet their information needs. This is due to the high proportion of "no preference", in which results in Kiswahili and English were equally used or clicked (over two third of the topics).

5.3.5 *Topic-Language Association*

We can see that language preferences exist only in a few super-topics and topics in general. A topic (and/or super-topic) can be associated with the language used to query or consume the results in such cases. This means that such topics have an association with the language of the query or results, which is known as *Topic-Language (T-L) association or preferences*.

Some reasons for each T-L-association may be obvious, given the languages involved and their Web resources, but others may have no rational justification. For example, why was English the query language for water and Kiswahili the query language for a related topic, waste management? There could be several reasons for this, including prior knowledge about whether there are enough documents in a particular language, user time constraints, and whether or not high recall is desired. The investigation of these and other factors influencing language preferences was not within the scope of this study.

5.3.6 *Limitations of the Study*

The type of participants is one limitation of this study that may affect making proper generalizations to the overall Swahili speaking Web users community in Tanzania. According to Figure 5.4, the majority of participants were between the ages of 18 and 34, with the majority of them being bachelor's degree students and some working adults. In terms of practical application, it would be advantageous to have a balanced representation of the Swahili speaking Web users community in terms of education, age, and occupation.

Another constraint is the limited number of participants. Most query log research, for example, in the Learning-to-Rank (L2R) studies, employs a large number of users and/or datasets, such as the Yahoo! Weboscope dataset [26]. Unlike other studies that use commercial search engine query logs and are essentially

limited 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.

Nonetheless, for the self-assessment of language use for Web search, users were not given space to explain why they rated the way they did and/or under what conditions they use a specific language. This made finding correlations between their explicit ratings and their implicitly inferred language preferences difficult.

Furthermore, only the results snippets were used to determine (perceived) relevance. Despite the fact that there is a strong correlation between perceived relevance in clicks and absolute relevance from actual users, Liu [123] argues that snippets must be treated with caution when inferring relevance. There is a substantial body of work on user modeling (click models) that rely on snippets to make relevance judgments, for example, Guo, Liu, Kannan, *et al.* [77], Craswell, Zoeter, Taylor, *et al.* [41], and Dupret and Piwowarski [61]. Other user modeling approaches, such as Dynamic Bayesian Network (DBN) [28], believe that snippets are insufficient to infer relevance of a result, and that users must visit a page.

Allowing system users to create their own queries may be important because it helps to reveal actual information needs and relevance judgments from real users. Users were forced to make relevance judgments on queries that they did not create under the current experimental design.

5.4 SUMMARY

Understanding user behaviour and preferences is critical for the development of effective MLIR systems. Survey studies may be incapable of capturing the complexities of IR and/or MLIR users' preferences. Query and click-through logs have proven to be a valuable source of implicit information, especially when it comes to user-system interaction behaviour. As a result, the carefully controlled study in this chapter investigated user preferences, specifically language preferences related to search topic and results.

Inspired by polyglots' code-switching behaviour when interacting with information retrieval systems, specifically the search engine, this study investigated the topic-language preferences in MLIR. The goal is to help with the development of new MLIR solutions based on user behaviour and preferences.

To the best of our knowledge, this is the first study to use multilingual query and click-through logs to investigate topic-language preferences and how preferences change during an MLIR search. The work is part of an effort to support "fair" multilingual information retrieval in low-resource languages like Kiswahili, which are spoken by a large number of users. The study focused on Swahili-speaking Web users in Tanzania.

In a small portion of this study, a questionnaire was used to learn how Swahili-speaking Web users evaluate themselves in terms of how they use English and Kiswahili to search for information on the Web. The majority of this study involved participants interacting with a guided multilingual search engine, with query and click-through logs analyzed to estimate Topic-Language (T-L) preferences/associations.

According to the findings of the questionnaire study, users frequently use English and occasionally use Kiswahili in their daily Web searches.

The results of the controlled multilingual search engine show that the level of topical granularity influences the Topic-Language (T-L) association. Preferences shifted slightly from the abstract (super-topic) to the fine level (topic). At the querying stage, for example, there was no preference for language in 53% of the super-topics, preference for Kiswahili in 47% of the super-topics, and none for English super-topics. However, topic analysis revealed that 9% of topics were preferred in English and 38% in Kiswahili, while the remaining 53% of the topics had no preference.

It was also discovered that the Topic-Language (T-L) associations depend on (change depending on) the stage of Web search, i.e., whether at the querying or results selection stage. At the querying stage, for example, users significantly preferred querying in English and Kiswahili for 9 percent and 38% of the topics, respectively. However, when it came to selecting results, they significantly preferred English and Kiswahili for 26% and 9% of the topics, respectively.

The findings also suggest that the Topic-Language (T-L) association was not present in all topics; for example, at the results selection stage, while there were language preferences in only 35% of the topics (9% for Kiswahili and 26% for English), there were no language preferences in the remaining 65%.

The findings of this study open up new avenues for MLIR research in the direction of developing better systems by shedding light on topic-language associations and preferences, and they can be used as a foundation for developing a better information retrieval system to assist users in specific scenarios. The scenarios could include populating the SERP (via re-ranking) with more (or top) results in the preferred language rather than equally interleaving results from both languages. This is the primary focus of the research in this thesis, in order to assist users who prefer a specific language for a particular topic, despite the low number of documents on the Web.

Part III

UTILIZING THE TOPIC-LANGUAGE PREFERENCES

6

USING TOPIC-LANGUAGE PREFERENCES IN MULTILINGUAL SWAHILI INFORMATION RETRIEVAL

Multilingual Information Retrieval (MLIR) systems display the final merged result list retrieved from various source collections in the original languages of the document collections [158]. An MLIR system merges the individual result lists to produce this final result list, ensuring that the merged results are relevant not only to the user's query, but also to the user's information needs.

Several approaches for merging exist in the literature to achieve this goal, as reviewed in Chapter 2. Traditional approaches mostly rely on the relevance scores of each document to the query, such as raw-score [181] and normalized-by-topk [116], and others that ignore the scores, such as round-robin [181], [214]. Other approaches to merging make use of ML techniques such as logistic-regression-based [111] and feature-based [206].

The study in Chapter 5 revealed that users have preference for certain languages when searching for information in certain topics – T-L associations/preferences. The study found that users have language preferences for some topics based on query and click-through logs from multilingual Swahili information retrieval system users. For example, while users significantly preferred results in Kiswahili for the *Music* topic, but significantly preferred results in English for the *Computer* hardware topic.

However, to the best of our knowledge, the current MLIR results merging approaches do not incorporate T-L preferences in search result ranking. As a result, the system hides potential relevant results further down the list, and users either miss them or expend extra effort to find them. Users are more likely to be satisfied if they see top-ranked results in the language they want/prefer right away, rather than having to scroll.

In this chapter, we propose a method for merging or re-ranking results that essentially ignores the relevance scores from individual result lists and instead uses the users' language preferences. The proposed method is called *T-L-based* algorithm. The approach primarily re-ranks a small set of the top most results before presenting them to the users, while the remaining results are interleaved. Because users are typically interested in the first few results, so it is critical to ensure that the first few results present the most relevant documents [123].

The following is the chapter's organizational structure. The proposed T-L-based algorithm is presented in Section 6.1. In Section 6.2, we show how our proposed approach performs better or worse in various scenarios. Section 6.3 provides a summary of the chapter.

6.1 T-L-BASED APPROACH

According to Algorithm 1, for topics with language preferences, the T-L-based algorithm pushes/promotes the preferred language results to the top of the results

list. The number of promoted results (batch size) n can be varied (from 1, 2, 3,..., 10) i.e., $n \in \mathbb{Z} : n \in [1, 10]$. The remaining results in the preferred language and those in the non-preferred language are then interleaved in a round-robin style. In other words, after a predetermined batch size of results in a preferred language is pushed to the top of the results list, the remaining results in that language and those in a non-preferred language are interleaved in a round-robin fashion until the result lists are exhausted. The T-L-based approach aims to present (more) results in a preferred language for a query in a specific topic first.

Algorithm 1 Topic-Language-Based Approach (High-Level)

i. Initialization i.e., topic name, starting language for interleaving, preferred language, batch sizes

ii. For topics with no language preferences:
   a. choose the language to start with
   b. interleave the results using round-robin-based approach with a batch size of 1
   c. Terminate and return the merged list

   Result: A list of interleaved results

iii. For topics with preference
   a. specify the language to start with
   b. specify the batch size e.g. 3
   c. specify the topic name;
   d. push the results in a specified batch size from the preferred language to the top of the merged list
   e. Interleave the remaining results between the ones in a preferred language and the non-preferred in a round-robin, based on a language chosen to start with
   f. Iterate through the lists until all the results are exhausted
   g. Terminate and return the merged list

   Result: Re-ranked Results per T-L Preferences

It should be noted that this definition refers to an MLIR system that only supports two languages – Kiswahili and English, where if one language is preferred for a specific topic, the other language is referred to as a non-preferred language. Because there are two languages involved, the starting language for interleaving must be determined. The choice of starting language in the interleaving process results in two variants of the T-L-based algorithm – T-L$_{En}$ and T-L$_{Sw}$, where English and Kiswahili were the starting languages for interleaving, respectively.

Consider an example in Figure 6.1, where the results are re-ranked using the T-L-based approach. The illustration assumes that users preferred a topic T in

Kiswahili, the batch size is 3, and the starting language for interleaving is English, i.e., T-L3_{En}.

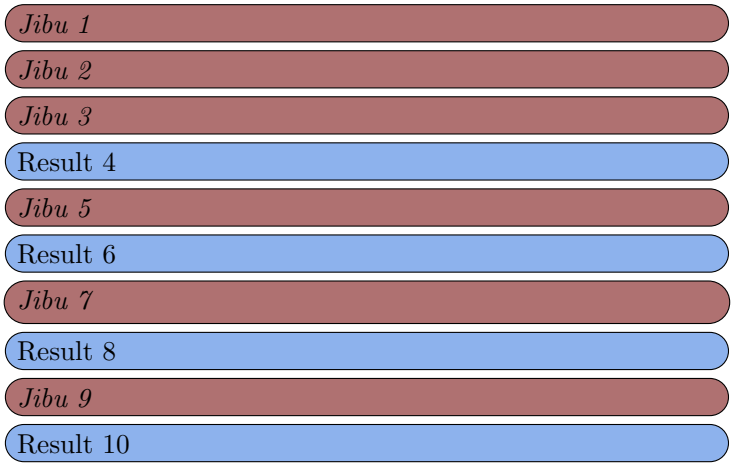


Figure 6.1: An illustration of how an T-L-based approach re-ranks and presents the first 10 MLIR results

6.2 FORMULATIONS AND ANALYSIS

In this section, we provide an analysis to show the cases where the T-L-based approach can and cannot outperform other systems that do not consider users' language preferences when ranking. We use an example of a system where the retrieval model retrieves two lists of results to be merged. That is, we use a case of an MLIR system that supports two languages A and B. Normally, each list contains a mix of relevant and irrelevant results.

The T-L-based approach promotes a specific number of results in a preferred language, while interleaving the rest of the results with those in the non-preferred language. There are two possible scenarios for result lists in the preferred language. First, the preferred language's list has either equal or more relevant top n results than the non-preferred language's list. Second, when compared to the non-preferred language list, the preferred language list has few or no relevant top n results.

6.2.1 *Equal or More Relevant Top Results in the Preferred Language*

Suppose that language B is preferred for topic G and all of the top n results from language B's list are relevant. Suppose also that for language A's list, we do not know which results are relevant and which ones are irrelevant. We use the proposed T-L-based approach to push the top n results from language B's list to the top of the merged results list. Calculating the performance using IR evaluation metrics, such as AP, can demonstrate that our T-L-based approach produces better performance for the merged MLIR results than another system S for merging results.

Assumption: If language B is preferred over language A for a specific topic G; and if the top n results in language B's list are relevant, while we do not know which results are relevant in language A's list; pushing the top n results from language B' list to the top of the merged list guarantees that the T-L-based approach T achieves better MLIR system ranking performance than any other system S, which does not promote results in a preferred language.

Using examples, we want to show that for T to promote top n results from the preferred language implies that the overall performance, in terms of AP, of the merged list is better than that of another system S, which does not promote results. Specifically,

$$AP(T) \geqslant AP(S) \tag{6.1}$$

We calculate the Precision at each rank position and then calculate AP@k. Since we do not know what top n results on language A's list are relevant and what are not, there are mainly three possibilities. First, all the top n results are relevant; second, all the top n results are irrelevant; and third, top n is a mixture of relevant and irrelevant results.

In the first case, if all of the top n results from language A's list are relevant, and given that all of the top n results from language B's list are also relevant, then

$$AP(T) = AP(S) \tag{6.2}$$

In the second case, if all of the top n results from language A's list are irrelevant, and given that all of the top n results from language B's list are relevant, then

$$AP(T) > AP(S) \tag{6.3}$$

In the third case, if top n results from language A's list is a mixture of relevant and irrelevant results, and given that top n results from language B's list are relevant, then

$$AP(T) \geqslant AP(S) \tag{6.4}$$

See Table 6.1 for an example, where we assume that each list has two top n results. We assume that the results in approach S are interleaved. In the best-case scenario, both approaches yield the same AP@4 score of 1.00, implying that

$$AP(T) = AP(S) \tag{6.5}$$

In the worst-case scenario, approach S yields an AP@4 score of 0.50, while the T-L-based approach T yields an AP@4 score of 1.00. As a result,

$$AP(T) > AP(S) \tag{6.6}$$

This example case generally means that the performance of the T-L-based approach will always be better or equal to that of any system S, i.e.,

$$AP(T) \geqslant AP(S) \tag{6.7}$$

The analysis demonstrates that the T-L-based approach improves relevance of ranked results of the MLIR system if the preferred language's results list contains equal or more relevant results at the top than the non-preferred language list.

	List A	List B	S	Precision	T	Precision
Best case	✓	✓	✓	1.000	✓	1.000
	✓	✓	✓	1.000	✓	1.000
			✓	1.000	✓	1.000
			✓	1.000	✓	1.000
			AP	**1.000**	**AP**	**1.000**
Worst case	X	✓	X	0.000	✓	1.000
	X	✓	✓	0.500	✓	1.000
			X	0.667	X	0.333
			✓	0.500	X	0.500
			AP	**0.500**	**AP**	**1.000**

Table 6.1: A demonstration for performance of the T-L-based in case the preferred language's list has either equal or more relevant top n results than the non-preferred language's top n.

6.2.2 *Few or No Relevant Top Results in the Preferred Language*

Suppose that language B is preferred for topic G and all the top n results from language B's list are irrelevant (we take the worst case). And suppose, for language A's list, we do not know which results are relevant and which ones are irrelevant. We can use the proposed T-L-based approach to push the top n results from language B's list to the top of the merged list. Calculating the performance using IR evaluation metrics such as AP can prove that our T-L-based approach T produces poor performance for the merged MLIR results than another system S of merging results.

Assumption: If language B is preferred than language A for certain topic G; and if the top n results in language B's list are irrelevant, while we do not know which results are relevant in language A's list; pushing the top n results from language B' list to the top of the merged list guarantees that the T-L-based approach T gets poor MLIR system ranking performance than any other system S, which does not promote results in a preferred language.

Using examples, we want to show that for T to promote top n results implies that the overall performance, in terms of AP, of the merged list is worse than that of another system S, which does not promote results. Specifically,

$$AP(T) \leqslant AP(S) \tag{6.8}$$

We calculate the Precision at each rank position and then calculate AP@k. Since we do not know which top n results on language A's list are relevant and which

are not, there are primarily three possibilities. First, all of the top n results are relevant; second, all of the top n results are irrelevant; and third, top n is a mixture of relevant and irrelevant results.

In the first case, if all of the top n results from language A's list are relevant, and given that all of the top n results from language B's list are irrelevant, then

$$AP(T) < AP(S) \tag{6.9}$$

In the second case, if all of the top n results from language A's list are irrelevant, and given that all of the top n results from language B's list are also irrelevant, then

$$AP(T) = AP(S) \tag{6.10}$$

In the third case, if top n results from language A's list is a mixture of relevant and irrelevant results, and given that all of the top n results from language B's list are irrelevant, then

$$AP(T) \leqslant AP(S) \tag{6.11}$$

See Table 6.2 for an example, where we assume that each list has two top n results. We assume system S interleaves the results. In the best case scenario, approach S yields an AP@4 score of 0.833 and the T-L-based approach T yields an AP@4 score of 0.417, resulting in

$$P(T) < P(S) \tag{6.12}$$

In the worst case scenario, both approaches yield the same AP@4 score of 0.000, implying that

$$P(T) = P(S) \tag{6.13}$$

This generally shows that the performance of an T-L-based approach will always be inferior or equal to that of other systems that do not promote results in a preferred language, i.e.,

$$AP(T) \leqslant AP(S) \tag{6.14}$$

Therefore, this analysis demonstrates that the T-L-based approach does not improve relevance of results an MLIR system if the preferred language's results list contains a few or no relevant results at the top compared to the non-preferred language list.

6.3 SUMMARY

The T-L-based algorithm was presented in this chapter as a method for merging or re-ranking MLIR results from the perspective of the users' language preferences. The approach is based on the findings presented in Chapter 5, in which users of MLIR systems demonstrated language preferences for specific topics. The proposed T-L-based approach incorporates these language preferences into the results ranking in order to improve MLIR performance.

We demonstrate analytically, using examples, in which scenarios the proposed T-L-based approach can provide better and/or worse MLIR performance than any system that does not include language preferences in the ranking. We primarily

	List A	List B	S	Precision	T	Precision
Best case	✓	X	✓	1.000	X	0.000
	✓	X	X	0.500	X	0.000
			✓	0.667	✓	0.333
			X	0.500	✓	0.500
			AP	**0.833**	**AP**	**0.417**
Worst case	X	X	X	0.000	X	0.000
	X	X	X	0.000	X	0.000
			X	0.000	X	0.000
				0.000	X	0.000
			AP	**0.000**	**AP**	**0.000**

Table 6.2: A demonstration for performance of the T-L-based in case the preferred language's list has a few or no relevant top n results than the non-preferred language's top n.

demonstrated that if there are equal or more relevant results at the top of the preferred language's result list compared to the non-preferred language's result list, then the T-L-based approach is guaranteed to improve the relevance of the results in the merged list. However, if there are no or only a few relevant results at the top of the preferred language's result list as opposed to the non-preferred, the T-L-based approach underperforms when compared to other systems that do not promote results.

7 EVALUATION OF THE T-L-BASED APPROACH

This chapter tries to demonstrate how using the Topic-Language (T-L) preferences can improve the relevance of ranked results in MLIR. The chapter primarily presents the evaluation results of our proposed T-L-based approach for merging multilingual Swahili IR in Chapter 6.

Thus, this chapter addresses the third research question (**RQ3**) of this thesis, which asks:

> *How can topic-language preferences improve the relevance of ranked results in a multilingual Swahili Information Retrieval (IR) system?*

We answer this question through the following research objectives:

RO1 To assess the overall performance of the proposed T-L-based approach in T-L association-sensitive topics.

RO2 To examine the factors that influence the performance of the T-L-based approach in T-L association-sensitive topics.

RO3 To determine whether there is a minimum threshold of promoted results that can provide optimal performance of the T-L-based approach in T-L association-sensitive topics.

The rest of the chapter is organized as follows. The following section (Section 7.1) is an experimental setup and data analysis for evaluating the T-L-based algorithm. The evaluation results are presented in Section 7.2, followed by a discussion in Section 7.3. The chapter is summarized in the final section (Section 7.4).

7.1 EXPERIMENTAL SET UP AND DATA ANALYSIS

7.1.1 *Dataset*

The dataset used to evaluate the proposed T-L-based merging strategy comes from a carefully controlled study that is thoroughly described in Chapter 5.

To summarize, we created a Swahili multilingual search engine that users interacted with using a prepared set of queries. Users could access non-simulated results by using the Microsoft (MS) Bing Web Search API. For each user, the system displayed the results in an interleaving style, randomly alternating the language to begin with – English or Kiswahili. The instructions directed users to click (check on the most relevant result(s) based on the snippet assessment) using a checkbox on the left-hand side of each result. Logs for queries and click-through were collected. We only used the click-through log data, which consisted of 3493 query records, for this evaluation.

After pre-processing, the click-through logs contained information about the query name, the topic to which it belongs, the language of each clicked result,

and a list of the clicked and non-clicked URLs – which were treated as relevance judgements. We assumed that the users' judgments were absolute, so the relevance judgements were purely binary, i.e., a click implies a relevant result, otherwise not. We removed all queries and their associated clicked results for topics that did not have language preferences because users did not have language preferences for all of them.

Pre-processing and removing topics with no language preference reduced the click-through logs data to 99 and 45 unique query records for English and Swahili preferred topics, respectively.

7.1.2 *Performance Measures*

The performance measures used in the evaluation were precision-based (Average Precision (AP) and Mean Average Precision (MAP)) and gain-based (Normalized Discounted Cumulative Gain (NDCG)). The AP and MAP measures both assess the average performance of the system in which the result list is ranked [150].

Despite the fact that the NDCG measure is mostly suitable for scaled/graded relevance judgements [94], it can also be used with binary labels. Thus, it was used in this study because it works on any ranked data/results, because it has an ability to take into account the position of each of the ranked results. The NDCG measure assumes that highly relevant documents are more useful than marginally relevant documents; thus, the lower a relevant document's ranked position, the less useful it is for the user because it is less likely to be examined. As a result, a discount function gradually diminishes the significance of relevant documents found further down the ranked results list.

7.1.3 *Baseline – Interleaving Approach*

The baseline approach obtains the final merged list by interleaving a *single* result from each of the (intermediate) result lists until all intermediate result lists are exhausted – *Round-Robin (R-R)* merging approach.

Unlike data fusion and other merging applications, using the algorithm in an MLIR environment requires the inclusion of an extra parameter – the language to begin with. In our case, the algorithm must begin the interleaving process with either English or Kiswahili. When English and Kiswahili are used as starting languages, the algorithm produces two instances, R-R$_{En}$ and R-R$_{Sw}$, respectively.

Figure 7.1 depicts an illustrative example SERP of an MLIR system in which Kiswahili is used as the starting language for interleaving (R-R$_{Sw}$).

Other traditional results merging strategies, such as raw score, normalized score, and weighted score merging, are represented by the R-R algorithm baseline. They essentially assume that the relevant documents are distributed uniformly across the individual result lists [166]. The R-R algorithm differs only in that it does not require relevance scores. We needed access to the relevance score values of each result from individual lists if we were going to use other measures as baselines. Because we did not have access to such scores in this study's setting, we used the round-robin as the only baseline.

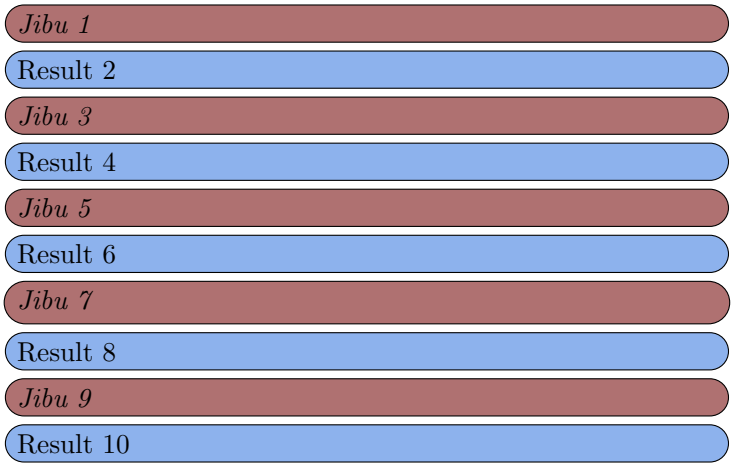


Figure 7.1: An example of how an R-R approach re-ranks and displays the top ten MLIR results.

7.1.4 *Notations and Analysis*

The notations R-R$_{En}$ and R-R$_{Sw}$ represent the Round-Robin (R-R) approach, where English and Kiswahili were the starting languages for interleaving. T-Ln_{En} and T-Ln_{Sw}, on the other hand, represent the T-L-based algorithm, with English and Kiswahili as the starting languages for interleaving, respectively, where n represents the number of promoted results, where $n \in \mathbb{Z} : n \in [1, 10]$.

The proposed T-L-based method is essentially a modified R-R method with promoted results in a preferred language. While the R-R algorithm produces no promoted results, the T-L-based algorithm promotes at least one (1) result(s) in a preferred language. As a result, our comparison employs the concept of percentage change (%), where we observe an increase or decrease in the performance of the T-L-based from R-R algorithm.

To evaluate the overall effectiveness of our T-L-based algorithms and the R-R, we averaged the MAP and NDCG scores obtained when English and Kiswahili were the starting languages for interleaving. For example,

$$\frac{\text{R-R}_{En} + \text{R-R}_{Sw}}{2} \tag{7.1}$$

The purpose of this analysis is to counteract the effect of starting language in the interleaving process. For R-R and T-L-based algorithms, we use the notations R-R$_{Av}$ and T-Ln_{Av}, where n is the number of promoted results in a preferred language.

The query-level analysis took into account users' query-clicking behaviours. The click behaviour shows how the query topic is linked to the preferred language of the results (T-L association). The T-L association in the queries means that some queries had their results clicked purely in one language, others had their results clicked more in one language than the other, and still others had their results clicked in a language other than the one revealed as the preferred language.

In Chapter 5, we discovered the following user interaction behaviours when users clicked on the MLIR results:

- clicking solely on English results;
- clicking solely on Swahili results;
- clicking an equal number of English and Swahili results, for example, 1 English result and 1 Swahili result;
- clicking more English than Swahili results, for example, 2 English results and 1 Swahili result;
- and clicking more Swahili than English results, for example, 1 English result and 3 Swahili results.

7.2 RESULTS

The evaluation results of our proposed T-L-based approach are presented in the following three subsections. The first subsection is dedicated to English preferred topics, the second to Swahili preferred topics, and the final to query level analysis of the T-L-based approach's performance. We only present an T-L-based algorithm with a maximum of 5 promoted results for space and presentation reasons. Other batch sizes (i.e., 6–10) are not included in the tables.

7.2.1 *T-L-Based Approach in English Preferred Topics*

7.2.1.1 *Interleaving with English as the starting language*

Using MAP metric, Table 7.1 shows that the T-L-based algorithm performed better than the R-R algorithm. The T-L-based approach, in particular, outperformed the R-R algorithm by up to 5.9% and 4.8% for MAP@5 and MAP@10, respectively, by promoting two English results (i.e., T-L2_{En}).

The NDCG scores show that the T-L-based approach performed comparably to the R-R approach. NDCG@10 scores, in particular, show that the performance of T-L-based and R-R$_{En}$ algorithms is the same (0.69). The performance of the T-L-based algorithm for NDCG@5 slightly deteriorated as the number of promoted results increased. For example, when the algorithm promoted 4 or 5 English results (T-L4_{En} or T-L5_{En}), the performance dropped by 8.8%.

Table 7.1: The MAP@n and NDCG@n scores for English preferred topics.

	MAP@ (% Change)		**NDCG@** (% Change)	
	5	**10**	**5**	**10**
R-R$_{En}$	0.69	0.65	**0.64**	**0.69**
T-L1$_{En}$	0.72 (+5.7%)	0.67 (+3.7%)	0.64 (0.0%)	**0.69** (0.9%)
T-L2$_{En}$	**0.73** (+5.9%)	**0.68** (+4.8%)	0.64 (-0.5%)	**0.69** (+0.9%)

... Continued on next page

Table 7.1 – continued from previous page

	MAP@ (% Change)		NDCG@ (% Change)	
	5	**10**	**5**	**10**
T-L3$_{En}$	0.72 (+5.5%)	0.67 (+3.8%)	0.64 (-0.5%)	**0.69** (+0.7%)
T-L4$_{En}$	0.71 (+3.9%)	0.67 (+3.6%)	0.59 (-8.8%)	**0.69** (+0.7%)
T-L5$_{En}$	0.71 (+3.9%)	0.67 (+3.7%)	0.59 (-8.8%)	**0.69** (+0.1%)
R-R$_{Sw}$	0.53	0.51	0.60	**0.69**
T-L1$_{Sw}$	0.69 (+29.7%)	0.65 (+27.4%)	**0.64** (+7.0%)	**0.69** (0.0%)
T-L2$_{Sw}$	0.72 (+37.1%)	0.67 (+32.1%)	**0.64** (+7.0%)	**0.69** (+0.9%)
T-L3$_{Sw}$	**0.73** (+37.4%)	**0.68** (+33.5%)	**0.64** (+6.4%)	**0.69** (+0.9%)
T-L4$_{Sw}$	0.72 (+36.9%)	0.67 (+32.3%)	**0.64** (+6.4%)	**0.69** (+0.7%)
T-L5$_{Sw}$	0.71 (+34.8%)	0.67 (+32.0%)	**0.64** (+6.4%)	**0.69** (+0.7%)
R-R$_{Av}$	0.61	0.58	0.62	**0.69**
T-L1$_{Av}$	0.70 (+16.1%)	0.66 (+14.1%)	**0.64** (+3.4%)	**0.69** (+0.4%)
T-L2$_{Av}$	**0.73** (+19.5%)	**0.68** (+16.8%)	**0.64** (+3.1%)	**0.69** (+0.9%)
T-L3$_{Av}$	0.72 (+19.4%)	**0.68** (+16.9%)	**0.64** (+2.8%)	**0.69** (+0.8%)
T-L4$_{Av}$	0.72 (+18.3%)	0.67 (+16.2%)	0.61 (-1.5%)	**0.69** (+0.7%)
T-L5$_{Av}$	0.71 (+17.4%)	0.67 (+16.2%)	0.61 (-1.5%)	**0.69** (+0.4%)

The R-R$_{En}$ and R-R$_{Sw}$ are the R-R approach with English and Kiswahili used as the starting languages for interleaving, respectively. The T-L1$_{En}$, ..., T-L5$_{En}$ and T-L1$_{Sw}$, ..., T-L5$_{Sw}$ are the T-L-based approach with different number of promoted results ranging from 1 to 5 and English and Kiswahili are the starting languages for interleaving, respectively. The R-R$_{Av}$ and T-Ln$_{Av}$ stand for the averaged R-R and T-Ln scores, respectively. It should be noted that the MAP and NDCG values in this table are rounded to two decimal places.

7.2.1.2 *Interleaving with Kiswahili as the starting language*

The MAP@5 and MAP@10 scores show that the T-L-based algorithm outperformed the R-R$_{Sw}$ algorithm by a wide margin (Table 7.1). For example, for MAP@5 and MAP@10 scores, the T-L3$_{Sw}$ improved performance by 37.4% and 33.5%, respectively.

The NDCG@5 and for NDCG@10 scores show that the T-L-based algorithm outperforms the R-R$_{Sw}$ algorithm. However, the improvement is subtle in most cases, and the performance is roughly the same for NDCG@10.

7.2.1.3 *Averaging the Scores*

Table 7.1 also shows that the averaged MAP scores indicate that the T-L-based approach performed better overall for both MAP@5 and MAP@10. Using the NDCG measure, the T-L-based approach outperformed the R-R approach slightly or performed equally. However, as the number of promoted results grew, the NDCG@5 scores dropped by 1.5%.

Using the MAP measure, Table 7.1 generally shows that when English is used as the starting language for interleaving, the T-L-based approach outperformed the R-R approach. However, the improvement in performance was minor, only up to 5.9% above baseline. When the starting language for interleaving was changed to Kiswahili, the T-L-based algorithm outperformed the R-R algorithm with a huge performance improvement of up to 37.4%. Averaging the results of starting with English and Kiswahili revealed that the T-L-based approach has a better overall MLIR performance improvement of up to 19.5%.

Using the NDCG measure, Table 7.1 shows that the T-L-based algorithm outperformed the R-R algorithm only when Kiswahili is used as the starting language for interleaving. This was also a minor improvement, up to 7.0% from the baseline. When English is used as the starting language for interleaving, the algorithm performed poorly, dropping up to -8.8% from the baseline. Averaging the two resulted in a slight improvement in the performance of the T-L-based approach, which can range between 3.4 and -1.5% above the baseline.

7.2.1.4 *At Query Level*

For the assessment of the T-L-based approach in queries, we use only the AP@10 and MAP@10 scores for demonstration. The scores of queries for both English and Swahili preferred topics can be found in Table 7.2. The scores shown in this table are the averaged AP scores when English and Kiswahili are used as the starting languages for interleaving, respectively. For example, to get T-L2_{Av}, calculate

$$\frac{\text{T-L2}_{En} + \text{T-L2}_{Sw}}{2} \tag{7.2}$$

The table also shows the sample queries categorized based on how their results were clicked by users, i.e., whether the clicked results were purely in English (En), more clicked English results than Swahili results (<En), results from both English and Kiswahili clicked equally (Equal), more clicked Swahili results (<Sw) than English results, and clicked results were purely in Kiswahili (Sw).

55 of the 99 queries from the English preferred topics had the clicked results purely in English, and 19 had more clicked English results than Swahili results. 12 queries, on the other hand, had an equal number of clicked English and Swahili results, 6 queries had more clicked Swahili than English results, and 7 queries had only clicked Swahili results.

Table 7.2: The AP@10 for a few selected queries for English preferred topics.

	QID	R-R$_{Av}$	T-L1$_{Av}$	T-L2$_{Av}$	T-L3$_{Av}$	T-L4$_{Av}$
En	2	0.75	**1.00** (+33.3%)	**1.00** (+33.3%)	**1.00** (+33.3%)	**1.00** (+33.3%)
	29	0.67	0.92 (+37.5%)	**1.00** (+50.0%)	**1.00** (+50.0%)	**1.00** (+50.0%)
	69	0.29	0.42 (42.9%)	**0.5** (71.4%)	**0.5** (71.4%)	**0.5** (71.4%)
<En	7	0.46	0.51 (+9.6%)	**0.54** (+18.1%)	**0.54** (+17.9%)	0.52 (+13.4%)
	40	0.29	0.42 (+42.9%)	**0.5** (+71.4%)	**0.5** (+71.4%)	**0.5** (+71.4%)
	75	0.63	0.84 (33.2%)	0.96 (52.7%)	**1.00** (59.3%)	**1.00** (59.3%)
Equal	8	0.67	**0.73** (+8.8%)	0.68 (+2.5%)	0.65 (-1.8%)	0.63 (-4.9%)
	43	0.75	**1.00** (+33.3%)	**1.00** (+33.3%)	**1.00** (+33.3%)	**1.00** (+33.3%)
	81	**0.56**	0.43 (-22.4%)	0.42 (-25.4%)	0.39 (-29.9%)	0.35 (-37.5%)
<Sw	17	**0.60**	0.56 (-7.5%)	0.54 (-10.3%)	0.52 (-14.3%)	0.49 (-18.5%)
	89	0.63	**0.64** (1.6%)	0.59 (-6.1%)	0.55 (-11.7%)	0.53 (-15.9%)
	94	**0.67**	0.43 (-35.0%)	0.33 (-50.6%)	0.28 (-58.5%)	0.26 (-61.6%)
Sw	1	**0.67**	0.43 (-35.0%)	0.33 (-50.6%)	0.27 (-59.9%)	0.23 (-66.2%)
	64	**0.51**	0.36 (-28.9%)	0.29 (-42.3%)	0.24 (-52.8%)	0.20 (-61.0%)
	86	**0.28**	0.23 (-17.6%)	0.20 (-29.5%)	0.16 (-41.7%)	0.13 (-52.1%)

En – only English URLs were clicked, *<En* – More clicks in English URLs, *Equal* – *Sw* and *En* clicked URLs were equal, *<Sw* – More clicks in Kiswahili URLs, and *Sw* – only Swahili URLs were clicked. The *QID* represents the query ID. It should be noted that the MAP and NDCG values in this table are rounded to two decimal places.

According to Table 7.2, with AP@10 of sample queries and Figure 7.2 with MAP@10 of all queries from the English preferred topics, the T-L-based algorithm outperformed the R-R algorithm in two scenarios.

i. For queries with purely clicked English results (En).

ii. For queries with more clicked English results than Swahili results (<En)

Table 7.2 shows that for these queries, the T-L-based approach, improved QID 2 and QID 40 performance by 33.3% and 71.4%, respectively, by promoting 2, 3, or 4 English results. Sometimes, the performance of an T-L-based approach increased with the number of promoted results, as in QID 75, where promoting 1 and 3 result(s) resulted in a 33.2% and 59.3% increase, respectively.

The performance of the T-L-based approach is both positive and negative for queries with an equal number of clicked results in both English and Kiswahili

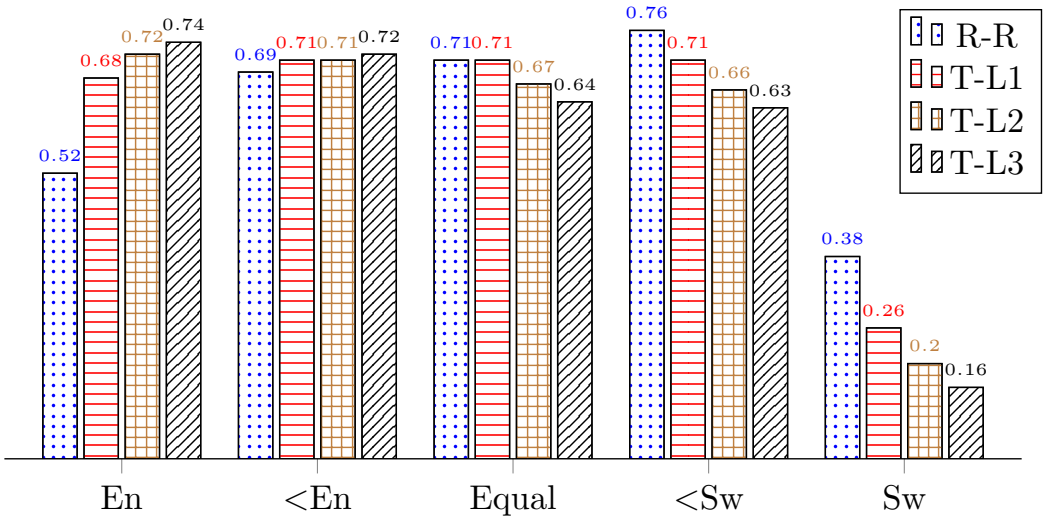


Figure 7.2: The MAP@10 scores for queries from English preferred topics were grouped based on the order in which their results were clicked.

(Equal). For example, while promoting 4 results improves the relevance of results for QID 43 by 33.3%, it cannot improve the relevance of results for QID 81, where performance drops by 37.5% when 4 English results are promoted. As the number of promoted English results increased, the T-L-based approach improvement for other queries in this category deteriorated, for example, QID 8.

The performance of the T-L-based approach was generally poor for queries with more clicked Swahili results than English results (<Sw) or queries with purely clicked Swahili results (Sw). As an example, when 4 English results were promoted (T-L4_{Av}) in QID 64, performance dropped by up to 66.0%. As the number of promoted results increased, so did the performance of the T-L-based algorithm deteriorated.

7.2.2 *T-L-Based Approach in Swahili Preferred Topics*

7.2.2.1 *Interleaving with English as the starting language*

In terms of the MAP measure, Table 7.3 shows that the T-L-based algorithm outperforms the R-R approach. T-L-based approach improves by up to 22% and 20.1% for by promoting 5 Swahili results (T-L5_{Sw}) in MAP@5 and MAP@10.

The results of the NDCG measure show that the T-L-based and R-R approaches performed nearly equally in most cases. For example, the NDCG@10 scores for R-R and all T-L-based are 0.61. For NDCG@5, T-L2_{En} and T-L3_{En} outperformed R-R$_{En}$ by 2.7%, respectively. However, as the number of promoted Swahili results increased, T-L-based performance declined by up to 22.0% (i.e., T-L5_{En}).

Table 7.3: The MAP@n and the NDCG@n scores for Swahili preferred topics.

	MAP@ (% Change)		**NDCG@** (% Change)	
	5	**10**	**5**	**10**
R-R$_{En}$	0.53	0.53	**0.51**	0.61
T-L1$_{En}$	0.63 (+17.6%)	0.61 (+15.1%)	**0.53** (+2.7%)	**0.61** (0.0%)
T-L2$_{En}$	**0.65** (+21.8%)	0.62 (+17.1%)	**0.53** (+2.7%)	**0.61** (+0.5%)
T-L3$_{En}$	**0.65** (+21.2%)	0.62 (+18.1%)	0.50 (-3.4%)	**0.61** (+0.5%)
T-L4$_{En}$	**0.65** (+21.6%)	**0.63** (+20.1%)	0.50 (-3.4%)	**0.61** (+0.7%)
T-L5$_{En}$	**0.65** (+22.0%)	**0.63** (+20.1%)	0.40 (-22.0%)	**0.61** (+0.7%)
R-R$_{Sw}$	0.63	0.61	**0.53**	**0.61**
T-L1$_{Sw}$	**0.65** (+3.5%)	0.62 (+1.9%)	0.52 (-0.4%)	**0.61** (0.3%)
T-L2$_{Sw}$	**0.65** (+3.1%)	0.62 (+2.6%)	0.50 (-5.9%)	**0.61** (+0.5%)
T-L3$_{Sw}$	**0.65** (+3.4%)	**0.63** (+4.4%)	0.50 (-5.9%)	**0.61** (+0.7%)
T-L4$_{Sw}$	**0.65** (+3.8%)	**0.63** (+4.4%)	0.40 (-24.0%)	**0.61** (+0.7%)
T-L5$_{Sw}$	**0.65** (+3.8%)	**0.63** (+3.9%)	0.40 (-24.0%)	**0.61** (+0.7%)
R-R$_{Av}$	0.58	0.57	0.52	**0.61**
T-L1$_{Av}$	0.63 (+9.0%)	0.61 (+7.5%)	**0.53** (+1.1%)	**0.61** (+0.1%)
T-L2$_{Av}$	**0.65** (+11.7%)	0.62 (+9.4%)	0.51 (-1.7%)	**0.61** (+0.5%)
T-L3$_{Av}$	**0.65** (+11.6%)	**0.63** (+10.8%)	0.50 (-4.6%)	**0.61** (+0.6%)
T-L4$_{Av}$	**0.65** (+12.0%)	**0.63** (+11.7%)	0.45 (-13.8%)	**0.61** (+0.7%)
T-L5$_{Av}$	**0.65** (+12.2%)	**0.63** (+11.4%)	0.40 (-23.0%)	**0.61** (+0.7%)

The R-R$_{En}$ and R-R$_{Sw}$ are the R-R approach with English and Kiswahili used as the starting languages in the interleaving process, respectively. The T-L1$_{En}$, ..., T-L5$_{En}$ and T-L1$_{Sw}$, ..., T-L5$_{Sw}$ are the T-L-based approach with different number of promoted results ranging from 1 to 5 and English and Kiswahili are the starting languages in the interleaving process, respectively. The R-R$_{Av}$ and T-Ln$_{Av}$ stand for the averaged R-R and T-Ln scores, respectively. It should be noted that the MAP and NDCG values in this table are rounded to two decimal places.

7.2.2.2 *Interleaving with Kiswahili as the starting language*

When the starting language for interleaving is changed to Kiswahili, Table 7.3 shows that the T-L-based algorithm outperforms the R-R algorithm for both MAP@5 and MAP@10 scores. However, the improvement is marginal, ranging between 3.1% and 3.8% for MAP@5 and 1.9% and 4.4% for MAP@10.

In terms of the NDCG measure, the T-L-based approach fell short of improving performance by up to 24% for NDCG@5. The scores for NDCG@10 are the same for the T-L-based and R-R approaches.

7.2.2.3 *Averaging the Scores*

For the averaged MAP, Table 7.3 shows that the T-L-based approach outperformed the R-R approach. The performance of the T-L-based algorithm for the averaged NDCG scores was generally poor for NDCG@5, where T-L1_{En} narrowly improved the performance by 1.1%. Promoting more results reduced the performance of the T-L-based approach by up to -23% (T-L5_{En}). Both the T-L-based and the R-R algorithms had uniform performance for NDCG@10.

When using the MAP measure, Table 7.3 shows that the T-L-based approach outperformed the R-R-based approach by a wide margin when English is the starting language for interleaving. However, when Kiswahili is used as the starting language for interleaving, the performance improvement was slight low. Averaging the two scores revealed that the T-L-based approach outperforms the baseline.

Using the NDCG measure and beginning with either English or Kiswahili for interleaving yields relatively poor results for the T-L-based approach. The performance ranged between 2.7% and -24.0%. Averaging the scores from the two sets had no effect on the T-L-based approach's performance improvement, which remained poor.

7.2.2.4 *At Query Level*

The AP@10 scores of sample queries for both English and Swahili preferred topics can be found in Table 7.4. The scores shown in this table are the averaged AP scores when English and Kiswahili are used as the starting languages for interleaving, respectively.

The table, as well as the Figure 7.3 (with MAP@10 for all queries from Swahili preferred topics) categorizes the queries based on how their results were clicked, i.e., whether the clicked results were purely in Kiswahili (Sw), more clicked Swahili results than English results (<Sw), equal number of clicked results in both English and Kiswahili (Equal), more clicked English results than Swahili results (<Sw), and the clicked results were purely in English (En).

15 out of 45 queries from the Swahili preferred topics had their only clicked Swahili results, and 16 queries had more clicked Swahili results than English results. Furthermore, 6 queries had an equal number of clicked Swahili and English results, 1 query had more clicked English results than Swahili results, and 7 queries had only clicked English results.

The table shows that the T-L-based algorithm outperformed the R-R approach for queries with purely clicked Swahili results and for queries with more clicked Swahili results than English results. For example, the T-L-based approach outperformed the R-R algorithm by up to 77.8% and 56.1%, by promoting 4 Swahili results in QID 11 and 4, respectively.

Table 7.4: The AP@10 for a few selected queries for Swahili preferred topics.

	QID	R-R$_{Av}$	T-L1$_{Av}$	T-L2$_{Av}$	T-L3$_{Av}$	T-L4$_{Av}$
Sw	11	0.37	0.45 (+21.6%)	0.54 (+45.5%)	0.61 (+63.7%)	0.66 (+77.8%)
	28	0.29	0.42 (+42.9%)	0.50 (+71.4%)	0.50 (+71.4%)	0.50 (+71.4%)
	36	0.48	0.60 (+25.9%)	0.66 (+38.5%)	0.67 (40.2%)	0.66 (38.1%)
<Sw	4	0.56	0.67 (+20.2%)	0.72 (+29.5%)	0.79 (42.4%)	0.87 (+56.1%)
	16	0.63	0.70 (+11.5%)	0.71 (+13.3%)	0.71 (+13.1%)	0.69 (+9.8%)
	37	0.73	0.85 (+16.7%)	0.90 (+24.4%)	0.90 (+24.8%)	0.88 (+21.9%)
Equal	6	0.67	0.58 (-12.5%)	0.54 (-18.8%)	0.48 (-28.8%)	0.43 (-35.0%)
	14	0.27	0.27 (+0.2%)	0.28 (+4.9%)	0.29 (+8.2%)	0.29 (+6.9%)
	18	0.33	0.36 (+10.3%)	0.38 (+16.6%)	0.37 (+11.8%)	0.36 (+8.0%)
<En	21	0.86	0.81 (-6.0%)	0.76 (-11.2%)	0.73 (-14.7%)	0.71 (-18.0%)
En	26	0.33	0.27 (-18.8%)	0.23 (-31.5%)	0.20 (-40.6%)	0.17 (-47.6%)
	29	0.29	0.23 (-22.9%)	0.18 (-37.5%)	0.15 (-46.9%)	0.13 (-54.1%)
	42	0.75	0.42 (-44.4%)	0.29 (-61.1%)	0.23 (-70.0%)	0.18 (-75.6%)

Sw – only Swahili URLs were clicked, *<Sw* – More clicks in Swahili URLs, *Equal* – *Sw* and *En* clicked URLs were equal, *<En* – More clicks in English URLs, and *En* – only English URLs were clicked. The *QID* represents the query ID. It should be noted that the MAP and NDCG values in this table are rounded to two decimal places.

The performance of the T-L-based approach is both positive and negative for queries with an equal number of clicked results in English and Kiswahili. For example, for QID 6 the performance deteriorated by up to 35%, for QID 14 the performance improved by up to 6.9%.

The performance of the T-L-based approach failed to outperform that of the R-R approach for queries with clicked results that were entirely in English or had more clicked English than Swahili results. In QID 42, performance dropped by up to 75.6%. The more Swahili results were promoted for these two cases, the worse the T-L-based approach performed.

Figure 7.3 displays the MAP scores for the grouped queries based on the order in which they were clicked. The graph shows that promoting more Swahili results for queries whose results were entirely clicked in Kiswahili and for queries with more clicked Swahili results improves relevance significantly.

The T-L-based approach performed worse for queries with results that were: i) equally clicked in both Kiswahili and English (Equal), ii) more English results

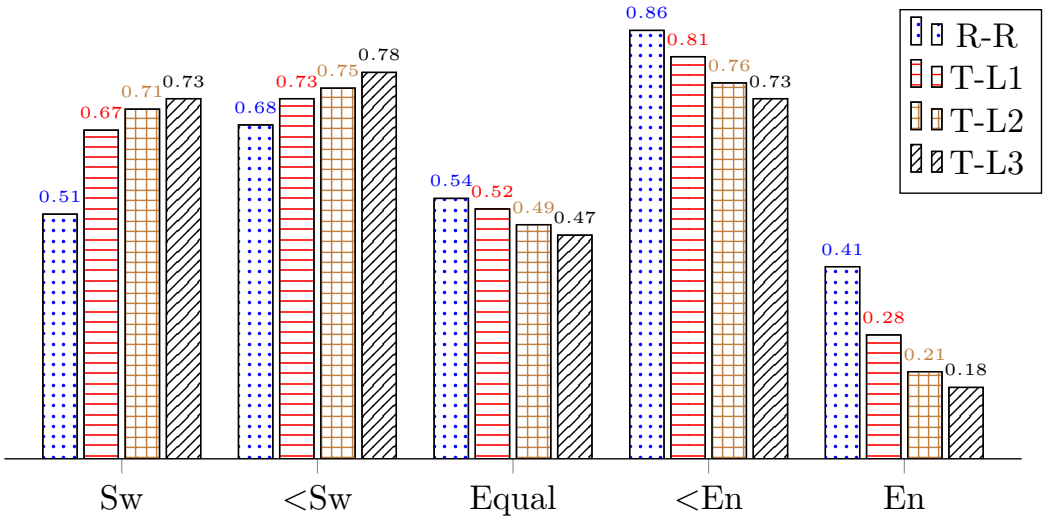


Figure 7.3: The MAP@10 scores for Swahili preferred topics were grouped based on how their results were clicked.

(<Sw), and iii) entirely in English (En). Furthermore, by promoting more Swahili results, the performance for these queries was reduced even further.

From Table 7.4 and Figure 7.3, it is seen that the T-L-based approach improved performance in two scenarios.

i. For queries with only clicked Swahili results.

ii. For queries with more clicked Swahili results than English results.

The T-L-based approach failed to improve results relevance for queries with an equal number of clicked English and Swahili results, more clicked English results than Swahili results, and queries with entirely clicked English results.

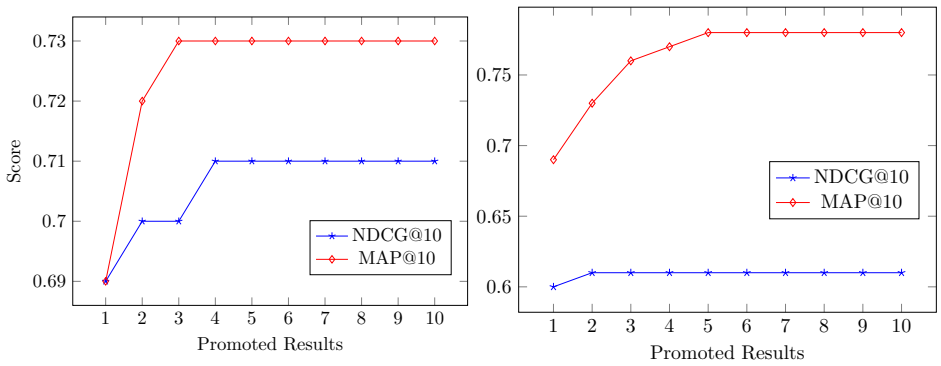
7.2.3 *How many results should be promoted?*

The T-L-based approach improves relevance of results for queries that conform to the estimated T-L association by promoting a certain number of results in a preferred language. However, the exact number of results to be promoted or a minimum threshold of results to be promoted is not yet known. To calculate this figure, we take the AP for queries with only clicked results in the estimated preferred language and those queries with more clicked results in the estimated preferred language than the non-preferred. There were 74 and 31 queries from English and Swahili preferred topics, respectively.

The graph in Figure 7.4 depicts the point at which increasing the promoted results does not improve the T-L-based algorithm performance any further. For English preferred topics (Figure 7.4a), the MAP and NDCG scores show that the T-L-based performance do not improve any further when the number of promoted English results are 3 and 4, respectively.

For Swahili preferred topics, Figure 7.4b shows that the MAP and NDCG scores stabilize when the number of promoted Swahili results are 5 and 2, respectively.

At this stage, the performance of T-L-based approach does not improve any further.



(a) *English Preferred Topics* (b) *Swahili Preferred Topics*

Figure 7.4: The required minimum number of promoted results for optimal T-L-based algorithm performance.

7.3 DISCUSSION

The proposed T-L-based approach is founded on the assumption that promoting more results in a preferred language can potentially improve the relevance of MLIR results. Unlike heuristic approaches such as R-R, our proposed approach takes the T-L association into account and uses it to re-rank/merge the MLIR results.

This section discusses the overall performance of the proposed T-L-based approach in comparison to the R-R approach as a baseline. The discussion is guided by the research objectives.

7.3.1 *Performance of the Proposed T-L-based Approach*

In the first research objective (**RO1**), we wanted to "*assess the overall performance of the proposed T-L-based approach in T-L association-sensitive topics*". The findings are divided into two categories: performance of the T-L-based approach per individual preferred language in topics; and performance of the T-L-based approach in individual queries.

7.3.1.1 *Performance at Topic Level*

The findings show that the T-L-based approach generally improves MLIR performance by outperforming the baseline in almost all cases for both English and Swahili preferred topics. However, there were minor variations in performance, owing primarily to the metric used. The MAP scores, in particular, show a significant improvement from the baseline, as opposed to the NDCG scores. In Table 7.1, for example, while the MAP@10 score improved the performance by 33.5%, the NDCG@10 score improved the performance by only 0.9% for T-L$3s_w$.

The T-L-based approach's blurred improvement when using the NDCG measure could be attributed to the nature of our dataset. MAP is better suited to the type of dataset we were working with (binary relevance judgement) than NDCG, which was originally designed for graded relevance [94], [150].

Our findings also revealed slight differences in results based on the starting language for interleaving. For example, Table 7.1 shows a minimal difference between T-L2$_{En}$ and T-L2$_{Sw}$ MAP@5 scores, which were 0.73 and 0.72, respectively. One reason for such findings could be that the starting language for interleaving may be the same as the preferred language; in this case, the starting language for interleaving acts as a continuation of the preferred language. This means that, unlike the R-R algorithm, the performance of the T-L-based algorithm remains stable regardless of the starting language for interleaving. This means that once the results have been promoted in a preferred language, the starting language for interleaving is no longer relevant.

7.3.1.2 *Performance at Query Level*

The performance improvement provided by the T-L-based approach was minor in some cases, particularly when the NDCG measure was used, as previously discussed. This necessitated a closer examination of the details, i.e., analysis at the query level rather than relying on analysis of queries bundled in a topic. To evaluate the T-L-based approach in queries, we sorted and grouped queries based on how users clicked on their results/URLs, as explained in Section 7.1.

The results show that the T-L-based approach performs well for queries that have all clicked results in the estimated preferred language, as well as those that have more clicked results in the estimated preferred language. The T-L-based approach performed poorly for queries with an equal number of clicked results from the two languages, more clicked results in the other language than the estimated preferred language, and all results in the other language than the estimated preferred language.

Assume the estimated preferred language for a specific topic is Kiswahili; the T-L-based approach performs better if all or most of the results clicked for a given query are in Kiswahili; otherwise, the performance is poor. The presence of queries with an equal number of clicked English and Swahili results, more clicked English results than Kiswahili results, and completely clicked English results for Swahili preferred topics resulted in poor T-L-based performance.

7.3.2 *Factors Influencing the Performance of the T-L-based Approach*

In the second research objective (**RO2**), we wanted to: "*examine the factors that influence the performance of the T-L-based approach in T-L association-sensitive topics.*" The findings, particularly at the fine grained level of query, suggest that the performance of the T-L-based approach is primarily determined by the proportion of clicked results/URLs in the estimated preferred language. That is, if the proportion of clicked results in the estimated preferred language is high, the T-L-based approach performs well; otherwise, it performs poorly.

The T-L-based approach's poor performance for queries that do not conform to the estimated T-L associations is understandable. Because the T-L-based approach is premised on the notion of T-L association. That is, it works best for topics and, more specifically, queries with such associations.

However, in Chapter 5, we presented statistical estimates of the T-L association that did not take into account click statistics in individual queries, but rather aggregated the click statistics under a topic. The presence of queries in topics with opposite click behaviour than the estimated T-L associations, such as having more clicked results in the other language than the estimated preferred one, harmed the performance of the T-L-based algorithm.

This observation indicates that calculating T-L-based performance without taking into account the fact that individual queries have varying click behaviour is a bad idea. That is one of the reasons why the T-L-based approach improved slightly for some cases at the topic level. As a result, the T-L-based approach's performance should be calculated for individual queries based on their click behaviour.

As a result, the performance improvement of the T-L-based approach is primarily determined by the strength of the T-L association for a query. The strength of the T-L association is determined by whether a query contained:

- all the clicked results in the estimated preferred language – *very strong T-L association*;
- more clicked results in the estimated preferred language than the non-preferred language – *strong T-L association*;
- equal number of clicked results in the estimated preferred language and the non-preferred language – *neutral T-L association*;
- more clicked results in a non-preferred language than the estimated preferred language – *weak T-L association*;
- and all the clicked results in a non-preferred language – *negative T-L association*.

For example, according to Table 7.2, there is a very strong T-L association in queries with QID 2, 29, and 69; a strong T-L association in QID 7, 40, and 75; a neutral T-L association in QID 8, 43, and 81; a weak T-L association in QID 17, 89, and 94; and a negative T-L association in QID 1, For example, in the case of a very strong T-L association, the T-L-based approach improved performance by up to 71.4% (QID 40 and 69). The T-L-based approach degrades performance by up to -66.2% for negative T-L association (QID 94).

To demonstrate that the stronger the T-L association, the stronger the T-L-based approach to improving MLIR performance, we separate queries with T-L associations (very strong, and strong) from queries without T-L associations (neutral, weak, and negative). Table 7.5 shows that, while the T-L-based approach vastly improves relevance of results for queries with actual T-L associations (very strong, and strong), the same approach vastly degrades relevance of results for queries without T-L associations.

Table 7.5: The MAP@10 scores for queries with T-L associations vs. queries without T-L associations for English and Swahili preferred topics.

	R-R$_{Av}$	T-L1$_{Av}$	T-L2$_{Av}$	T-L3$_{Av}$	T-L4$_{Av}$	T-L5$_{Av}$
			English Preferred Topics			
with	0.60	0.69 (+14.8%)	0.72 (+19.2%)	**0.73** (+20.7%)	**0.73** (+21.0%)	**0.73** (+20.9%)
without	**0.62**	0.56 (-9.30)	0.51 (-16.9)	0.48 (-22.8)	0.45 (-27.2)	0.44 (-28.3)
			Swahili Preferred Topics			
with	0.59	0.70 (+17.8%)	0.73 (+23.1%)	0.76 (+27.4%)	0.77 (+30.5%)	**0.78** (+31.3)
without	**0.61**	0.53 (-11.9%)	0.49 (-19.1%)	0.46 (-24.1%)	0.44 (-28.2%)	0.42 (-31.0%)

The *with* and *without* stand for queries with T-L association and without T-L association, respectively.

7.3.3 *Minimum Threshold of Promoted Results for Optimal T-L-based Approach Performance*

The third research objective seeks to "*determine whether there is a minimum threshold of promoted results that can provide optimal performance of the T-L-based approach in T-L association-sensitive topics*". Our results in Section 7.2.3 show that the threshold for achieving optimal performance of the T-L-based approach varies depending on the preferred language as well as the evaluation measure used. However, the results suggest that promoting at least three to four results for English preferred topics and queries and two to five results for Swahili preferred topics and queries could ensure the best T-L-based performance.

7.3.4 *Applications of the T-L-based Approach to MLIR*

We observe that the T-L-based approach, which is intended to improve the relevance of ranked results for topics with language preferences, performs better for queries with strong T-L associations. The T-L-based merging approach improves the relevance of ranked MLIR results under the following conditions.

- The query must come from a T-L association-sensitive topic.
- The query in the T-L association-sensitive topic must have a strong T-L association.

This conclusion implies that, in order for the T-L-based approach to improve results relevance, the T-L associations must be known. In Chapter 5, we proposed

using statistical approaches (at the topic level) from historical click-through logs – the *implicit approach*. By taking individual click behaviour into account, this study revealed another way to supplement the T-L association estimation at a query level. This is an implicit estimate as well.

We also recommend asking MLIR users, via the search engine interface, what language preferences they have for a query – *explicit approach*. A user stating their language preferences for a query may indicate that there is a strong T-L association for such a query.

Since we evaluated our proposed MLIR merging approach on the Swahili-speaking click-through dataset, we can claim that our proposed T-L-based approach improves the relevance of the ranked multilingual Swahili IR results for any query or set of queries with known T-L associations.

7.3.5 *Limitations of the Study*

Despite the fact that we used data from actual users interacting with the MLIR system, our evaluation was entirely based on IR evaluation metrics such as MAP and NDCG. Due to time and financial constraints, we were only able to conduct a system evaluation. It would have been interesting to see how actual users evaluate the effectiveness of our proposed T-L-based approach in terms of relevance improvement.

7.4 SUMMARY

This chapter presented the results of an evaluation of a proposed T-L-based approach for merging multilingual Swahili IR. It presents the experimental setup by describing the baseline, specifically the R-R, the dataset used for evaluation, the performance measures employed, and the notations and analysis strategies used.

We evaluate the performance of the proposed T-L-based approach in both topics and queries. The overall performance of the T-L-based approach against the baseline was good. Changing the starting language for interleaving had no effect on overall performance. This implies that, after promoting a certain number of results in a preferred language, one can choose English or Kiswahili at random without affecting the overall performance of the T-L-based approach. It should be noted that the T-L-based approach only works for T-L association-sensitive topics and queries.

The analysis at the query level revealed that the T-L-based approach outperforms the baseline by a wide margin for queries with strong T-L association. For queries with no T-L association, the baseline outperforms the T-L-based algorithm. That is, the T-L-based approach performs well for queries that have all clicked results in the preferred language and queries that have more clicked results in the preferred language than the non-preferred language; otherwise the performance is poor.

The evaluation results, both at the topic and query levels, indicate that an T-L-based approach can improve the relevance of ranked results in multilingual Swahili IR for a specific set of topics – T-L association sensitive topics – and queries with strong T-L association. A strong T-L association can be seen in queries where

the results are entirely in the estimated preferred language or have more results in the estimated preferred language than the non-preferred one. As a result, in order for the T-L-based approach to yield better results, the language preferences (T-L associations) must be known. We can estimate them by using historical click data or by explicitly asking MLIR users to state them when interacting with the system.

Using the dataset in this study, we can see that there is a certain number of promoted results in a preferred language after which no further improvements can be made. This criterion is determined by the preferred language. For example, to achieve the best T-L-based performance for English and Swahili preferred topics, respectively, at least 3 and 2 promoted results are required. While confirming the thesis's main argument, the findings show that using the latent T-L association in MLIR can significantly improve the relevance of results in MLIR, particularly in a multilingual Swahili IR.

Part IV

CONCLUSIONS

8 CONCLUSION

When using traditional IR systems, multiple language speakers (polyglots) using low resourced languages face a number of challenges, the most significant of which is document scarcity. Because native languages are underrepresented, they choose to use foreign languages such as English, which is a highly resourced language on the Web. However, not all polyglots are fluent in multiple languages; some have difficulty formulating queries. Thanks to MLIR systems, it is possible to search the Web using a query in one language, such as the native language, and receive results in multiple languages, such as English and native language. However, one challenge with MLIR is that it is based on the same concept as IR systems, which aims to ensure topical relevance while overlooking the fact that monolingual and multilingual speakers differ; polyglots have different preferences and expectations.

The purpose of this thesis was to investigate and apply a language preferences perspective to improve the relevance of ranked results in a multilingual Swahili IR system. This study was carried out in three stages. First, we looked into the information needs and search behaviours of Swahili-speaking Web users. The goal was to gain an understanding of the problem space and, in particular, to investigate the factors influencing code-switching between English and Kiswahili among Swahili-speaking Web users during Web searches. Second, we investigated the relationship between search topic and preferred language of search, where a search topic was identified as one of the factors for code-switching among Swahili-speaking Web users in the previous stage. Finally, we used the discovered association between search topic and preferred language to develop a re-ranking algorithm and evaluate its impact on improving the relevance of MLIR results.

This chapter concludes the thesis and is structured as follows. It begins with a general summary of the thesis, highlighting the answers to the research questions and the major findings in Section 8.1. Section 8.2 outlines the contributions of this research to the body of knowledge and, in particular, MLIR research. Section 8.3 discusses potential works and directions for future works on Swahili IR/MLIR. The final section (Section 8.4) contains the thesis's final concluding remarks.

8.1 THESIS SUMMARY AND DISCUSSIONS

8.1.1 *Summary of the Thesis*

The primary goal of this study was to investigate and apply language preferences, specifically Topic-Language (T-L) preferences, in order to improve the relevance of ranked results in multilingual Swahili Information Retrieval (IR) system. We argued that considering the human aspect of language preferences when merging

the final MLIR results improves the relevance of the ranked results. The primary research question posed below guided the investigation.

> *How can the association between topic and language (T-L association) in Multilingual Information Retrieval (MLIR) be estimated, and how can it improve relevance of ranked results in a multilingual Swahili Information Retrieval (IR) system?*

This primary question was divided into three research questions, which are summarized in Table 8.1.

Table 8.1: Research questions and the chapters in which they are addressed.

SN.	Question	Chapter
RQ1	What are the information needs and search behaviour of polyglot Swahili-speaking Web users in Tanzania?	Chapter 4
RQ2	What are the topic-language preferences among the polyglot Swahili-speaking users of the multilingual Swahili Information Retrieval (IR) system?	Chapter 5
RQ3	How can topic-language preferences improve relevance of ranked results in a multilingual Swahili Information Retrieval (IR) system?	Chapter 6; Chapter 7

In this section, we provide a summary of what was covered in each chapter while also demonstrating how each of the research questions listed above was addressed.

1. Chapter 1 presented the motivating reasons and rationale for looking into yet another angle to improve the performance of MLIR, especially when one language is a low-resource language and there is a clear difference in how the languages are used for daily communication and business. In addition, the chapter presented the research problem, questions as well as the methods and materials used to answer them.

2. Chapter 2 provided background information on IR and MLIR, as well as a detailed literature review on information needs and search behaviour, MLIR, and evaluation measures.

3. The literature review was expanded by Chapter 3 to include Swahili specific state-of-the-art solutions for Swahili IR and CLIR/MLIR, such as NLP and MT.

4. To answer the first research question (**RQ1**), Chapter 4 presented a survey-based study to better understand the information needs and search behaviours of Swahili-speaking Web users. The study involved 11 information science experts and Swahili specialists from Tanzania. The main findings are as follows.

- **Result 1**: Because of the topic of the search, the type of task, the context of the information, and language proficiency, there is a dynamic code-switching between English and Kiswahili. Furthermore, most users do not believe that formulating a query in Kiswahili differs from doing so in English.
- **Result 2**: The search engines return unsatisfactory Swahili results, owing to a lack of Swahili resources on the Web, poor search techniques, and search engine failure to handle Swahili queries.
- **Result 3**: While professionals largely use English information, ordinary citizens primarily use Kiswahili. Professionals typically have a higher level of education, and as a result, they can successfully code-switch or have a large enough vocabulary to locate and consume English information, as opposed to less educated ordinary citizens. Because the majority of the Swahili-speaking community consists of ordinary citizens, there is a great need for more relevant results in Kiswahili.
- **Result 4**: In many fields, such as agriculture, justice, and health and well-being, there is a high demand for relevant Swahili information on the Web.
- **Result 5**: The level of awareness of Swahili language technology and tools is generally low.

5. To address the second research question (**RQ2**), Chapter 5 covered users' self-assessment of using Kiswahili and English in their daily Web search, as well as an analysis using query and click-through logs from the carefully controlled multilingual search engine to explore the search behaviour of Swahili-speaking Web users. We focused on the relationship between the search topic and the preferred language of the query and language of the results (T-L association/preferences). The study was open to any Swahili-speaking Web user, and 676 people took part in it by using a guided multilingual search engine (supporting Kiswahili and English). The following are the major outcomes.
   - **Result 1**: Without users' specific topic in mind, Swahili-speaking Web users' self-ratings suggest that they often use English and occasionally use Kiswahili for daily Web search.
   - **Result 2**: The T-L association/preferences do not exist in all topics during the querying or results selection stages. For example, users significantly preferred querying in Kiswahili and English for 38% and 9% of the topics, respectively; and in the remaining topics users could query using either English or Kiswahili equally.
   - **Result 3**: The T-L association is likely to change during the course of a Web search, from the querying stage to the results selection stage. In other words, the query language preferences differ from results language preferences.
   - **Result 4**: The more granular the level of topic, the clearer the T-L association. That is, it is too broad to conclude that one language is

preferable to another when more topics are aggregated into a super-topic, but preferences become more apparent when the topics are fragmented. For example, while Kiswahili was generally significantly preferred as a query language, topic analysis revealed that there was a significant preference for Kiswahili in only 38% of the topics, with 9% in English, and the remaining 53% had no preference for language.

6. To address the third research question (**RQ3**), Chapter 6 described the proposal and development of our proposed T-L-based algorithm, which used the T-L associations identified in Chapter 5, to re-rank MLIR results. Analytically, we demonstrated the following.
   - The T-L-based approach outperforms any system that does not push results in a preferred language to the top of the results when the top of the preferred language's result list contains equal or more relevant results than the non-preferred language's list.
   - The T-L-based approach's performance is worse if there are no or only a few relevant results at the top of the preferred language's result list compared to the non-preferred language's list.

7. Chapter 7 responds to **RQ3** by presenting an evaluation of the proposed T-L-based algorithm for re-ranking multilingual Swahili IR results. This proposed method was compared to a widely used interleaving technique (Round-Robin (R-R)). The click-through log dataset obtained from the guided multilingual search engine described in Chapter 5 was used in the evaluation. The major findings are as follows.
   - **Result 1**: For topics and queries with T-L associations, the T-L-based approach outperformed the baseline, except for cases when there were more clicked results in the non-preferred language.
   - **Result 2**: The T-L-based approach to improving the MLIR is heavily reliant on the strength of the T-L association in individual queries. That is, the T-L-based approach performs well for queries with all clicked results in the preferred language and queries with more clicked results in the preferred language than in the non-preferred language; otherwise, performance is poor.
   - **Result 3**: The performance of the T-L-based approach is less affected by the starting language for interleaving, i.e., starting with interleaving in either English or Kiswahili is unimportant.
   - **Result 4**: The ideal number of results for optimal T-L-based algorithm performance may differ depending on the preferred language. For example, in order to achieve the best T-L-based performance for English and Swahili preferred topics using the analysed dataset, at least 3 and 2 promoted results are required, respectively.

8. The current chapter (Chapter 8) provides the thesis's concluding remarks, highlighting the significant results obtained, the limitations in carrying out the results, and future research opportunities.

The primary research question comprised of two parts: how to estimate the T-L association and how such an T-L association can improve the relevance of ranked results in a multilingual Swahili IR. The answers to the first part of this question show that T-L associations can be implicitly estimated from query and click-through logs using some statistical estimations for preference test as well as directly observing the proportion of the clicked results language for each individual query. The findings to the second part of the primary question show that incorporating Topic-Language (T-L) preferences improves the relevance of the ranked results in a multilingual Swahili IR system. Specifically, the approach is effective for topics with language preferences, and especially for queries with a strong Topic-Language (T-L) association.

8.2 THESIS CONTRIBUTIONS

The contributions of this thesis are divided into theoretical and empirical. The principal contribution of this thesis lies in the identification and use of language preferences for improving relevance of ranked results in multilingual Swahili IR systems. This major contribution is augmented by the empirical studies, which produced several empirical contributions to knowledge in the MLIR field as summarized below.

8.2.1 *Theoretical Contribution*

- Contribution 1 – *Language preferences for improving relevance of multilingual Swahili IR.* Aside from topical relevance, this thesis provides another perspective for improving the relevance of ranked results in MLIR. The thesis argued and demonstrated through testing that using T-L preferences can improve the relevance of results in language preference sensitive queries/-topics. The use of a language preferences perspective in MLIR results ranking has been ignored, and as a result, users have struggled to locate results in their preferred language for a query/topic, which may be hidden further down the list.

- Contribution 2 – *Workflow for achieving the same results in different MLIR systems using a different set of languages.* The methodology used in the study enables other researchers to replicate it in other MLIR systems with different sets of languages and benefit from the T-L preferences on relevance improvement.

  The workflow begins with understanding the users' search behaviour, which can be accomplished through a simple user survey, then rolls out an MLIR system for actual users to use, collects their query and click-through log data, estimates/determines the T-L preferences/associations, and uses such associations in the re-ranking model for topics with preferences.

8.2.2 *Empirical Contributions*

- Contribution 3 – *Evidence for the need to develop multilingual Swahili IR systems*. The empirical study to understand the information needs and search behaviour of Swahili-speaking Web users provided evidence and extent of the need for a better system for information access and retrieval. The study identified distinct sectors and people who require different types of information in different languages.

- Contribution 4 – *Implicit feedback to study the topic as a factor for code-switching in Web searches in polyglots.* Surveys in which Web users explicitly state the reasons for their language change have been used in studies to understand the factors for code-switching in Web search in polyglots. Our guided multilingual search engine generated query and click-through logs, which were used to study and deduce implicitly that the topic of a user's search can be associated with the language they use.

- Contribution 5 – *Methods for estimating T-L associations.* The study established methods for inferring the T-L associations/preferences in topics/-queries through statistical analysis and click behaviour analysis. In contrast to survey-based studies, the click-logs study demonstrated in great detail how language preferences become more apparent as the level of granularity increases.

- Contribution 6 – *The topic of search is only one of many factors that influence code-switching*. The findings that T-L associations occur in only a few topics imply that other factors for code-switching, such as document availability and search context, are important reasons for language switching during Web search.

- Contribution 7 – *T-L associations can be investigated for improved MLIR system results*. Using the case of multilingual Swahili IR, we show that the T-L association can improve the relevance of ranked results in the MLIR system.

Overall, we demonstrated in this thesis that taking into account the search topic and the preferred language can significantly improve the relevance of the ranked results for some specific cases in a multilingual Swahili IR. We provided empirical evidence that search topic is associated with search language, which we referred to as the Topic-Language (T-L) association or preference.

Topical relevance for IR problems has been paramount in providing users with relevant results. However, multilingual web users may come from completely different backgrounds in terms of culture, education, and language use. People in Tanzania, the country used as a case study for this study, speak both Kiswahili and English. English is only used for professional communication, judicial and government documentation, and higher learning studies. In other words, it is a language of documentation or records; it is not spoken on the streets. Even government officials may hold meetings in Kiswahili but document in English.

We chose a different path to supplement topical relevance in MLIR and meeting users' information needs for Tanzania's Swahili-speaking Web users community

for two reasons. First, the people who interact with IR are polyglots – people who speak both Kiswahili and English. Second, there is a clear separation in how the languages are used in Tanzania's Swahili-speaking community. There are some communities around the world where all languages are equally used in all aspects of life. This is not the case in the Swahili-speaking community, where it is uncommon to encounter people chatting in English (or both languages) in the streets or office corridors. In other words, English is merely a record language, which means that all official documents must be recorded or written in that language.

This state adds a layer of complexity to information search and consumption, where topical relevance alone may not be sufficient. As a result, this study demonstrates that using a different ranking algorithm, at least for specific topics, may yield better results than topical relevance alone. Thus, the central argument of this thesis is that providing better MLIR results requires understanding that the way users interact with information is complex. And this complex mechanism for how people interact with information has to do with factors such as the societies in which people live, as well as educational systems. That is, when people have been trained in specific ways and live in specific societies, they do not interact with information in a simple monolingual manner. As a result, their perception of the relevance of IR/MLIR results may differ, and some users may be discouraged and frustrated by systems that only focus on topical relevance.

So, in this multilingual interaction with information, certain complexities can be exploited to provide people with better results. As a result, this study delves into one of these complexities: T-L preferences. We also contend that the complexities of information interactions differ from one community or country to the next. This means that when the same study is conducted in another country, such as one that is tri-lingual, the results may be unique.

8.3 RECOMMENDATIONS FOR FUTURE RESEARCH

The proposal to incorporate language preferences in the ranking of multilingual Swahili IR for improving relevance of ranked results opens up other avenues and challenges worth investigating and/or extending.

8.3.1 *Investigating other Factors for Improving Relevance of MLIR*

The findings in Chapter 4 suggested several factors for code-switching, including the topic of the search, the context of the information sought, language proficiency, and the type of task the searcher wishes to perform. This thesis concentrated on a single factor, *topic of search*, which demonstrated significant improvement in improving the relevance of ranked results for some cases on a multilingual Swahili IR system. An intriguing study could look into the effect of the other factors, either individually or in combination, on improving the relevance of MLIR ranked results.

8.3.2 *Investigating using other Display Styles*

According to Ling, Steichen, and Choulos [119], MLIR users prefer the panel layout style of presenting results. That is, each language list's results are displayed in a separate panel. In this thesis, we used the traditional interleaving display of results. We propose expanding the study to a panel to see if the T-L associations can be influenced by results presentation style and how this affects the overall performance of the T-L-based algorithm.

8.3.3 *Investigating other Implicit Information*

The current thesis relied solely on click-data to infer T-L preferences. Another study may collect or decode a lot more information for more perspectives into user T-L preferences, such as asking users for graded relevance judgements and reasons for their choices/judgements, and collecting more parameters such as dwell time, eye movements, and scroll rate when transitioning between Swahili and English results. Users could also be allowed to visit the actual pages and be asked to save and rate whether or not they found it useful. The rating can be given on any scale, such as 4 for highly relevant, 3 for relevant, 2 for somewhat relevant, and 1 for completely irrelevant. Analysis of post-click activities such as dwell-time, amount of scrolling on the clicked page, and cursor movements. The studies by Liu, Miao, Zhang, *et al.* [125] and Buscher, White, Dumais, *et al.* [21] suggest that the use of these post-click behaviours is significantly more effective in implicitly estimating relevance of the results than relying on a single factor alone. It would be useful to have data on the amount of time required to complete relevance assessments for each user/query. In addition to better understanding user behaviour, this may be useful in removing some noise, such as users who were too fast or random users, but also in determining whether there are easy/hard topics and whether this varies by language.

8.3.4 *Considering other Ways to Collect Click-data*

We propose another study that considers the use of crowd-sourcing approaches for data collection in order to obtain diverse Web users. The data produced and used in Chapters 5 and 6 came primarily from a small group of users who exhibited similar behaviour. The demographics of participants in Chapter 5 show that the study was dominated primarily by student participants. This could have an effect on generalizability to other groups of Swahili-speaking Web users. Another study with more diverse groups of users and/or more data on the topic of language preferences and language preference changes may provide more insights.

8.3.5 *Investigating the Language Preferences in the Input Query*

We propose to investigate whether the language of the input query language (as determined by query data) is a predictor of overall language preference. For

evaluation purposes, this study was limited to the language of the results (click data), ignoring the query data.

8.3.6 *Investigating the Inter-assessor Agreement*

It is interesting to investigate the aspect of the same query being evaluated by multiple users and calculating the inter-annotator agreement (e.g., Krippendorff's alpha) among them. This could help to understand if there is a relationship between language preference, such as whether there is better agreement on documents in English or Kiswahili.

8.3.7 *Investigating the Human-Computer Interaction Perspective*

It might be interesting to test our proposed T-L-based approach with actual users. That is, create an experiment with two groups: a control group that receives the R-R ranked results and an experimental group that receives the T-L-based ranked results. Instead of relying solely on IR evaluation measures, this study may reveal actual user behaviour and satisfaction.

8.3.8 *Exploring the Machine Learning Perspective*

Because of the limited data, the current thesis used a simple, yet effective, approach to incorporate language preferences in the ranking of results for T-L-sensitive topics in order to improve the relevance of their results. Considering the identified code-switching factors as features for training the re-ranking model, we recommend investigating the use of Learning-to-Rank (L2R) approaches. Tsai, Chen, and Wang [206] and Tsai, Wang, and Chen [208], for example, handcrafted several features to develop a merge model for MLIR, as did Usunier, Amini, and Goutte [209], who treated each language as a view of a document. Thus, adding T-L preferences as a feature to Tsai, Chen, and Wang's [206] work or treating T-L as a different view from Usunier, Amini, and Goutte's [209] could potentially incorporate ML.

8.4 FINAL REMARKS

Kiswahili, as a language, has the potential to become a regional language, as many African countries, particularly those in the Sub-Saharan region, are increasingly adopting it. This means that we may see an increase in the number of documents/resources on the Web; documents have a wide range of diversity due to the uniqueness of the language, which evolved from Bantu languages and incorporates Arabic, English, German, and a number of other colonial and pre-colonial commercial languages, such as Persian. Furthermore, as each country where the language is spoken has its own culture and education system – both of which are known to influence language use in Web search – the behaviours of the language's speakers will continue to vary. As a result, more research is needed to deal with such diversity and the increasing demand for Swahili or a

combination of language information per specific community of Swahili speakers.

The research in this thesis focused on assisting in the development of a better MLIR system for Tanzania's Swahili-speaking Web users. It incorporated user search behaviour, particularly topic-language preferences, into the ranking to improve the ranking in multilingual Swahili IR. Despite the limited scope of the experiments, the research yields promising results on improving the relevance of results in MLIR for language preference sensitive topics/questions.

As a result, the thesis contributes to the development of better MLIR systems that consider user preferences in their ranking, so that users do not have to struggle to find relevant results in their preferred language.

Part V

APPENDIX

A

A SURVEY ON INFORMATION AND KISWAHILI EXPERTS

This appendix primarily contains the materials used to conduct the preliminary study on Swahili Web users described in Chapter 4.

The ethical clearance approval from University of Cape Town (UCT) is contained in Section A.1, as are the recruitment materials, such as a letter template (email) sent to the heads/directors of the institutions targeted for interview participants, the consent form for the participants, and the invitation message (template) to the participants selected by their heads/directors of institution. The interview schedule can be found in Section A.2.

A.1 PARTICIPANTS RECRUITMENT

A.1.1 *UCT Ethical Clearance Approval*

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Faculty of Science
University of Cape Town
Rondebosch
South Africa 7701

Tel: +27 21 650 2866/7
E-mail: Rachel.Wynberg@uct.ac.za

11 May 2018

Mr Joseph Telemala
Department of Computer Science

RE: Information Retrieval System for Swahili Language

Dear Mr Joseph Telemala

I am pleased to inform you that the Faculty of Science Research Ethics Committee has approved the above-named application for research ethics clearance, subject to the conditions listed below.

- Implement the measures described in your application to ensure that the process of your research is ethically sound; and
- Uphold ethical principles throughout all stages of the research, responding appropriately to unanticipated issues: please contact me if you need advice on ethical issues that arise.

Your approval code is: **FSREC 26 - 2018**

I wish you success in your research.

Yours sincerely

A/Prof Rachel Wynberg
Chair: Faculty of Science Research Ethics Committee

Cc: Prof Hussein Suleman (Supervisor)

A.1.2 *Recruitment Email Template to Heads/Directors of Institution*

Subject: Request for Research Participants from your Institution

Habari «...»?

My name is Joseph Philipo Telemala, and I am an Assistant Lecturer in the Department of Mathematics, Informatics, and Computational Sciences (or the former Department of Informatics) at Sokoine University of Agriculture (SUA), as well as a PhD student in the Department of Computer Science at the University of Cape Town in South Africa. I'm just starting out in my studies, and I'm looking for solid evidence to back up my "statement of the problem" by conducting a survey on the subject.

My research interest is in Information Retrieval (IR). I intend to apply IR techniques to the problem of retrieving information or documents from the web using Swahili queries. Commercial search engines like Google are doing well, but there are some challenges that I believe are more specific to Swahili-speaking information seekers.

Therefore, I'm writing to request your assistance in locating information experts from your institution. As «*information experts*»/«*Kiswahili specialists*» working at the «*Name of the institution*», I believe they can assist me with their knowledge and insights into the information needs and search behavior of Swahili-speaking information seekers. I only need two or three people to participate.

Because I am not currently in Tanzania, I have prepared some interview questions that I will ask during an online interview (Skype or WhatsApp Messenger).

I have attached the "informed voluntary consent form" as part of the University of Cape Town's research ethics procedure.

Finally, I will request that you provide me with the contact information or email addresses of those who agree to participate in my study (those who sign the attached consent form) so that I can make personal arrangements with each of them.

Kind regards,
Joseph.

A.1.3 *Consent Form*

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF CAPE TOWN
PRIVATE BAG X3
RONDEBOSCH 7701
SOUTH AFRICA

RESEARCHER: Joseph P. Telemala
TELEPHONE: +27-729-611492
E-MAIL: tlmjos001@myuct.ac.za
URL: https://www.cs.uct.ac.za/

Informed Voluntary Consent to Participate in Research Study

Project Title: Information Retrieval System for Swahili Language

Invitation to participate, and benefits: You are invited to participate in a research study conducted with information specialists and Swahili language experts. The study aims at developing an understanding of the challenges that the Swahili speakers face when searching and accessing information using Kiswahili on the web. I believe that your experience would be a valuable source of information, and hope that by participating you may gain useful knowledge.

Procedures: During this study, you will be asked to respond to questions via interview.

Recording: We may record the audio/video conversation for the purpose of close follow up in the analysis of the data you have provided. If you object to this, please indicate this below.

Risks: There are no potentially harmful risks related to your participation in this study.

Disclaimer/Withdrawal: Your participation is completely voluntary; you may refuse to participate, and you may withdraw at any time without having to state a reason and without any prejudice or penalty against you. Should you choose to withdraw, the researcher commits not to use any of the information you have provided without your signed consent. Note that the researcher may also withdraw you from the study at any time.

Confidentiality: All information collected in this study will be kept private in that you will not be identified by name or by affiliation to an institution. Confidentiality and anonymity will be maintained as pseudonyms will be used.

What signing this form means:

By signing this consent form, you agree to participate in this research study. The aim, procedures to be used, as well as the potential risks and benefits of your participation have been explained verbally to you in detail, using this form. Refusal to participate in or withdrawal from this study at any time will have no effect on you in any way. You are free to contact me, to ask questions or request further information, at any time during this research.

I agree to participate in this research (tick one box) ☐ Yes ☐ No _____ (Initials)

I agree to be photographed/audio-recorded/video-recorded (strikethrough as applicable)

☐ Yes ☐ No _____ (Initials)

Name of Participant	Signature of Participant	Date
Joseph P. Telemala		**14/05/2018**
Name of Researcher	Signature of Researcher	Date

A.1.4 *Invitation Message Template to Participants*

A.1.4.1 *English Version*

Hello «*Name*»

I have received your contact information from the head/director of your department/institute, «*Title & Name*», and you have signed the **Informed Voluntary Consent Form**, and you are now ready to participate in my interview study. First and foremost, please accept my heartfelt gratitude for your assistance in completing this preliminary study. I can assure you that the interview will be brief (less than half an hour).

Because I am currently in South Africa, we will conduct a virtual interview. Therefore, I require some information from you.

1. Do you prefer to conduct the interview via WhatsApp or Skype? If it's a WhatsApp call, please share the number, and if it's a Skype call, please share the Skype ID.
2. What day and time are you available for the interview? You may choose any day in the next two weeks that is convenient for you.

Thank you very much, and have a wonderful time.

Joseph P. Telemala,
University of Cape Town, South Africa.

A.1.4.2 *Swahili Version*

Habari «*Name*»?

Nimepewa mawasiliano na mkuu/mkurugenzi wa idara/taasisi yako, «*Cheo & Jina*», na kwavile umetia saini **Fomu ya Kukubali Kushiriki kwa Hiari**, na kwamba upo tayari kushiriki kwenye usaili (interview) ya utafiti wangu. Kwanza nashukuru sana kwa utayari wako kunisaidia kufanikisha utafiti wangu huu mdogo. Napenda kukuhakikishia kuwa usaili huu hautachukua muda wako mwingi (chini ya nusu saa).

Kwavile kwa sasa nipo Afrika Kusini na ili kufanikisha hili, tutahitaji kufanya huu usaili kwa njia ya mtandao. Hivyo, nahitaji kufahamu vitu vifuatavyo toka kwako.

1. Utapenda nikupigie kwa WhatsApp au Skype? Kama ni WhatsApp naomba unitumie namba unayotumia au kama ni Skype naomba unitumie Skype ID.
2. Ni siku na muda gani mzuri ambao una nafasi ili nikupigie? Unaweza kuchagua siku yoyote ndani ya wiki hizi mbili zijazo.

Nashukuru sana na ninakutakia wakati mwema.

Joseph P. Telemala,
Chuo Kikuu cha Cape Town, Afrika Kusini.

A.2 MATERIALS

A.2.1 *Interview Schedule*

A.2.1.1 *English Version*

1. What is your current title (position)?
2. How long have you been at your current job?
3. Have you ever worked in a different job before this one?
4. Was the job in information/library (or Kiswahili) related?
5. Do you use the Web (search engines) to find information?
6. Do you care about the search engine's interface language?
7. Do you care about the language of the results/answers, or do you care about the relevance of the results regardless of language?
8. What language do you primarily use when querying/searching for information?
9. If you make a query/question in Swahili (as opposed to an English question):
   a) Is it time-consuming to think about a query/question?
   b) In terms of words, how long does it take you to formulate a query/question?
   c) Do you get the expected results/answers?
   d) How much time do you spend searching for the correct information/answer among the returned results; do you have to sift through pages of returned answers?
   e) What do you believe the reasons are?
10. If you were a professional in a specific field, such as an accountant or a doctor, would you need information in Swahili from the Web for your job?
11. If you were an ordinary citizen, such as a small-scale farmer, would you need Swahili information from the Web?

12. Do citizens in a multilingual country like Tanzania require information in Swahili on the Web?
13. Can you give some examples of industries/sectors that require the most Swahili information?
14. Could you comment on the number of Swahili documents available on the Web?
15. What difficulties do you face in providing your services to a Swahili community when the majority of the information/documents are in English?

A.2.1.2 *Swahili Version*

1. Unafanya kazi kwenye nafasi gani kwa sasa (Cheo chako)?
2. Una miaka/muda gani kwenye hii nafasi ya kazi?
3. Uliwahi kufanya kazi nyingine kabla ya kujiunga hapa?
4. Je, kazi hiyo ilikuwa inahusiana na habari/ukutubi (au Kiswahili)?
5. Je, unatumia mtandao/wavuti wa kutafuta habari (search engine) ili kutafuta/kupata habari zozote?
6. Je, huwa unazingatia lugha ya kiolesura (interface) ya huo mtandao?
7. Je, huwa unazingatia lugha ya matokeo ya kile ulichotafuta au huwa unaangalia tu kama matokeo yanaendana na kile ulichotafuta bila kujali lugha iliyotumika?
8. Je, ni lugha gani sana unatumia kwenye kuandika swali/swala au kutafuta habari mitandaoni?
9. Kama unatumia Kiswahili kuuliza swali (kulinganisha na Kiingereza):
   a) Inakuchukua muda zaidi kuwaza swali?
   b) Inakuchukuaje kutunga swali, hasa upande wa kiwango cha maneno unayotumia kuunda/kutunga swali?
   c) Je, unapata matokeo/majibu kama unavyotarajia?
   d) Inakuchukua muda gani kupata jibu/majibu ya kile unachotafuta katika majibu yanayoletwa na mfumo (search engine); je, inakuhitaji kuvinjari zaidi kurasa/majibu zinazoonyeshwa?
   e) Je, unadhani nini sababu yake?

10. Kama ungekuwa mtaalamu kwenye fani fulani, mfano uhasibu au daktari wa binadamu, ungehitaji taarifa kwa Kiswahili toka kwenye wavuti kwenye kazi yako?
11. Kama ungekuwa siyo mtaalam wa fani fulani yaani mwananchi wa kawaida, mfano mkulima mdogo, ungpenda taarifa za Kiswahili toka mitandaoni (kwenye wavuti)?
12. Kwenye nchi yenye lugha nyingi kama Tanzania, unadhani wananchi wanahitaji taarifa kwa Kiswahili toka mitandaoni (wavuti)?
13. Unaweza kutaja ifano ya sekta amabazo zinahitaji zaidi taarifa kwa Kiswahili?
14. Unasemeaje idadi ya nyaraka na taarifa za Kiswahili mitandaoni (kwenye wavuti)?
15. Unakutana na changamoto gani kwenye kutoa huduma kwenye jamii ya Waswahili na mazingira ambayo taarifa na nyaraka nyingi zipo kwa Kingereza?

B

EXPLORING TOPIC-LANGUAGE PREFERENCES

This appendix contains the materials, methods, diagrams, and tables that were used to conduct the study for exploring topic-language preferences in multilingual Swahili information retrieval, as described in Chapter 5 of this thesis.

Section B.1 contains recruitment materials such as ethical clearance, email and message (templates) that were sent out to solicit user participation in the study. Materials, specifically the experiment protocol, are provided in Section B.2. Finally, in Section B.3, extra tables detailing the statistics and results presented in Chapter 5 are shown.

B.1 PARTICIPANT RECRUITMENT

B.1.1 *UCT Ethical Clearance Approval*

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Faculty of Science
University of Cape Town
Rondebosch
South Africa 7701

E-mail: shari.daya@uct.ac.za
Tel: 021 650-2880

5 November 2019

Mr. Joseph P. Telemala
Department of Computer Science

Investigating Topic-Language Preferences for Improving Multilingual Swahili Information Retrieval

Dear Mr. Joseph P. Telemala

I am pleased to inform you that the Faculty of Science Research Ethics Committee has approved the above-named application for research ethics clearance, subject to the conditions listed below.

- Implement the measures described in your application to ensure that the process of your research is ethically sound; and
- Uphold ethical principles throughout all stages of the research, responding appropriately to unanticipated issues: please contact me if you need advice on ethical issues that arise.

Your approval code is: **FSREC 103 - 2019**

I wish you success in your research.

Yours sincerely

Dr Shari Daya
Chair: Faculty of Science Research Ethics Committee

Cc **A/Prof. Hussein Suleman (supervisor)**

B.1.2 *Sokoine University of Agriculture (SUA) Staff, Students and Researchers Clearance*

CLEARANCE PERMIT FOR CONDUCTING RESEARCH IN TANZANIA

SOKOINE UNIVERSITY OF AGRICULTURE
OFFICE OF THE VICE-CHANCELLOR
P.O. Box 3000 CHUO KIKUU, MOROGORO, TANZANIA
Phone: 255-023-2640006/7/8/9, Direct VC: 2640015;
Email: vc@sua.ac.tz;

Our Ref. SUA/DRPSG/R/126/3/97 **Date:** 18 December, 2019

TO WHOM IT MAY CONCERN
SOKOINE UNIVERSITY OF AGRICULTURE
MOROGORO

Re: UNIVERSITY STAFF, STUDENTS AND RESEARCHERS CLEARANCE

The Sokoine University of Agriculture was established by University Act Number 7 of 2005 and SUA Charter of 2007 which became operational on 1st January 2007 repealing Act Number 6 of 1984. One of the mission objectives is to generate and apply knowledge through research. For this reason the staff and researchers undertake research activities from time to time.

To facilitate the research function, the Vice-Chancellor of the Sokoine University of Agriculture (SUA) is empowered to issue research clearance to both staff, students and researchers of SUA onbehalf of the Government of Tanzania and the Tanzania Commission for Science and Technology.

The purpose of this letter is to introduce to you **Joseph Philipo Telemala**, a bonifide PhD Student of University of Cape Town and an employee of Sokoine University of Agriculture. **Joseph Philipo Telemala** has been granted clearance to conduct research in the country. The title of his research is "***Language Preferences for Improving Multillingual Swahili Information Retrieval***"

The period for which the permission has been granted is from **January, 2020 to April, 2020**. The research will be conducted in Sokoine University of Agriculture.

Should some of these areas/institutions/offices be restricted, you are requested to kindly advice the researcher(s) on alternative areas/institutions/offices which could be visited. In case you may require further information on the researcher please contact the University.

We thank you in advance for your cooperation and facilitation of this research activity.

Yours sincerely,

Prof. Peter R. Gillah
FOR: VICE-CHANCELLOR

VICE CHANCELLOR
SOKOINE UNIVERSITY OF AGRICULTURE
P. O. Box 3000
MOROGORO, TANZANIA

Copy to: Joseph Philipo Telemala - **Student**

B.1.3 *Invitation Email Template to Participants*

B.1.3.1 *English Version*

Subject: Invitation to Participate in a Research

Hello,

I'm a PhD student at the University of Cape Town conducting research to investigate the effect of language preferences in improving the relevance of results in a multi-language information search (or multilingual information retrieval). I'm inviting you to participate in this study because I believe your experience using multiple languages (Kiswahili and English) while searching the Web will be a valuable source of information.

Simply click on the following link and follow the simple instructions.

`http://simba.cs.uct.ac.za/~joseph/`

You will begin by completing the consent form and a few demographics questions, followed by the search exercise using our guided multilingual search system. We will provide you with topics and queries (search questions); you will not need to compose any search queries.

Please contact me (Telemala, Joseph) if you have any questions or need assistance with this research, via tlmjoso01@myuct.ac.za (or WhatsApp +27 729 611492).

Thank you for your time and assistance.

Kind regards,
Joseph Telemala,
Department of Computer Science, University of Cape Town, South Africa.

B.1.3.2 *Swahili Version*

Subject: Invitation to Participate in a Research

Habari?

Mimi ni mwanafunzi wa Shahada ya Uzamivu (PhD) katika Chuo Kikuu cha Cape Town, nafanya utafiti unaoangazia uhusiano wa upendeleo wa lugha (language preferences) katika kukuza uwezekano wa kupata matokeo yanayoendana zaidi na swali kwenye mazingira ya utafutaji habari kwa kutumia lugha zaidi ya moja. Nakukaribisha kushiriki kwenye utafiti nikiamini kwamba uzoefu wako kwenye kutumia lugha zaidi ya moja (Kiswahili na Kingereza) wakati wa kutafuta taarifa mitandaoni (kwenye wavuti) ni chanzo muhimu katika kufanikisha utafiti huu.

Cha kufanya ni kubofya kwenye kiungo (link) ifuatoyo na fuata maelekezo rahisi yaliyomo ndani.

http://simba.cs.uct.ac.za/~joseph/

Utaanza na kujaza fomu ya kukubali kushiriki kwa hiari na kisha utajaza taarifa zako muhimu. Baadaye utaendelea na zoezi la kuatfuta taarifa mtandaoni kupitia mfumo wetu maalumu (guided multilingual search system). Tutakupatia mada na maswali utakayotumia kuuliza kwenye wavuti; huhitaji kutunga maswali yako.

Kama una swali lolote au unahitaji msaada kuhusu utafiti huu, tafadhali usisite kunifikia mimi, Joseph Telemala, kupitia baruapepe ya tlmjoso001@myuct.ac.za (au WhatsApp +27 729 611492).

Asante sana kwa muda na msaada wako.

Wako,
Joseph Telemala,
Department of Computer Science, University of Cape Town, South Africa.

B.1.4 *Invitation Message Template to Participants (WhatsApp and other Social Media)*

B.1.4.1 *English Version*

Invitation to Participate in a Research Study.

This research involves searching using a prototype search engine that supports two languages, allowing you to get results in both English and Kiswahili.

It is very simple to participate; simply fill out the consent form and a few demographics information, and you are ready to begin searching. Remember to follow the instructions for each step.

To make things even easier, we have prepared the questions (queries), so you don't have to come up with your own for searching.

Click this link (`http://simba.cs.uct.ac.za/~joseph/`) and then follow the instructions.

Please keep in mind that you may search as many times as you like.

If you have any questions or require assistance, please contact me, Joseph P. Telemala, via email tlmjoso01@myuct.ac.za or WhatsApp +27 729 611492.

Thank you so much for your time and assistance.

Joseph P. Telemala,
PhD student,
Department of Computer Science, University of Cape Town, South Africa.

AFTER PARTICIPATING PLEASE REMEMBER TO SHARE WITH YOUR WHATSAPP GROUPS AND FRIENDS

B.1.4.2 *Swahili Version*

Naomba ushiriki wako kwenye utafiti.

Utafiti huu unahusisha kutafuta taarifa (search) kwenye mfumo wa majaribio (prototype search engine) inayoruhusu lugha MBILI yaani unapata majibu kwa lugha mbili: English na Kiswahili.

Kushiriki ni rahisi sana; utajaza fomu ya kukubali kushiriki kwa hiari kwenye utafiti (consent form) na kujaza taarifa chache za awali, kisha utaendelea na zoezi la ku-search. Fuata maelekezo rahisi yaliyopo kwenye kila hatua.

Kufanya mambo yawe rahisi zaidi, tumeandaa maswali yote, hamna haja ya kujitungia swali lako la ku-search.

Bonyeza kiungo (link) hiki (`http://simba.cs.uct.ac.za/~joseph/`) na kisha fuata maelekezo.

Zingatia: Unaweza kurudia zoezi la kusearch mara nyingi uwezavyo.

Kama unaswali au unahitaji msaada usisite kunitafuta mimi, Joseph P. Telemala kwa baruapepe ya tlmjoso01@myuct.ac.za au Whatsapp +27 729 611492.

Nashukuru sana kwa muda na msaada wako.

Joseph P. Telemala,
PhD student,
Department of Computer Science, University of Cape Town, South Africa.

BAADA YA KUSHIRIKI SAMBAZA KWENYE MAGROUP MENGINE NA RAFIKI ZAKO PIA

B.1.5 *Consent Form*

Informed Consent

Dear participant,

My name is Joseph P. Telemala, a PhD student at the University of Cape Town, South Africa. I'm currently conducting a research entitled "Investigating Topic-Language Preferences for Improving Multilingual Swahili Information Retrieval."

I cordially invite you to participate in a research study conducted with persons who use both Kiswahili and English to search for information from the Web. The study aim is to investigate the association between topic of search and the preferred language (topic-language association) in a multilingual search engine to enhance relevance of multilingual search results. I believe that your experience using multiple languages (Kiswahili and English) in the Web search would be a valuable source of information.

How to participate: During this study, you will be asked to use a list of prepared queries to search the Web using a multilingual search engine, assess the relevance of the retrieved results according to your language preferences finally submit your relevance judgement/choices.

Risks: There are no potentially harmful risks related to your participation in this study.

Feedback: You will receive feedback about the results of this research in the journal paper and/or conference proceeding and my PhD thesis.

Disclaimer/Withdrawal: Your participation is completely voluntary; you may refuse to participate, and you may withdraw at any time without having to state a reason and without any prejudice or penalty against you. Should you choose to withdraw, the researcher commits not to use any of the information you have provided without your signed consent. Note that the researcher may also withdraw you from the study at any time.

Confidentiality: All information collected in this study will be kept private in that you will not be identified by name or by affiliation to an institution. Confidentiality and anonymity will be maintained as no identity-revealing data such as name, age, IP address will be used in the research.

What signing this form means: By signing this consent form, you agree to participate in this research study. The aim, procedures to be used, as well as the potential risks and benefits of your participation have been explained verbally to you in detail, using this form. Refusal to participate in or withdrawal from this study at any time will have no effect on you in any way. You are free to contact me, to ask questions or request further information, at any time during this research.

I agree to participate in this research (tick one box) ☐ Yes ☐ No _____ (Initials)

_____	_____	_____
Name of Participant	Signature of Participant	Date

B.2 MATERIALS

B.2.1 *Experiment Protocol*

1. **Overview**

   We will conduct a controlled lab study in this experiment to investigate the presence of topic and language preferences in multilingual search before devising some automated methods to determine and/or use them to improve search results.

2. **Aim**

   The primary goal of this experiment is to determine whether or not there are topic-language preferences in a multilingual Swahili information retrieval system.

3. **Requirements/Tools**

   a) A guided multilingual search engine.
      - Bing Web Search API (and subscription keys).

   b) A corpus of topics.

   c) A corpus of Queries.

   d) Online machine translation services (Bing Microsoft Translator and Google Translate).

4. **Procedures**

   a) Informed Voluntary Consent Form

      A consent form will be available on the guided multilingual search engine's home page. When a participant is willing to participate, he or she will click **YES**, and a new text field will appear for her or him to fill in her or his names (or initials). If the potential participant clicks **NO**, she or he will be asked to confirm their participation.

   b) Demographics Information

      A participant will be asked to fill out her/his demographic information after successfully signing the consent form (by registering). These details include: gender, age group, education level, and occupation. A participant will also be asked to rate herself/himself on a 5-point scale for how she/he used Kiswahili and English to search for information on the Internet.

   c) Topic and Language Selection

      The system will choose and display 5 topics at random from among all of the prepared topics. A participant will be required to choose a topic from the ones displayed, then choose a language from a drop-down menu in which to view the queries.

d) Query Selection

Queries in the participant's preferred language will be displayed. Each query contains a URLs to the search engine. The results page will be displayed once the user clicks.

e) Search Results Assessment and Submission.

Following the display of the results, a participant will be asked to carefully inspect them and select, via a checkbox, those that appear to be more relevant to the query. The URLs will be disabled in order to prevent a participant from wasting time trying to follow every URL. All relevant judgements must be made based on the snippets.

f) Appreciation Message.

The guided multilingual search engine's final page will contain a thank you message and a request to repeat if a participant has enough time.

5. **Output**

a) Query and click-through logs.

B.3 EXTRA TABLES

B.3.1 *Grouping of Topics*

Table B.1: 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.

B.3.2 *TL Preferences in Super-topics and Topics*

Table B.2: Testing query language preferences in super-topics, 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, $\alpha = 1 - \text{BINOMDIST}(x - 1, n, P_0, 1)$ = Type I error, $\beta = \text{BINOMDIST}(x - 1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

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

Table B.3: Testing query language preferences in topics, 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 $n \leqslant 30$, 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, $\alpha = 1 - \text{BINOMDIST}(x - 1, n, P_0, 1)$ = Type I error, $\beta = \text{BINOMDIST}(x - 1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

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
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

... Continued on next page

Table B.3 – continued from previous page

SN.	Topic	Sw	En	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
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
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

Table B.4: Testing preferences for language of results in super-topics, 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, $\alpha = 1 - \text{BINOMDIST}(x - 1, n, P_0, 1)$ = Type I error, $\beta = \text{BINOMDIST}(x - 1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

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
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	NP
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

Table B.5: Testing preferences for language of results in topics, x = minimum number of common responses given by $x = (n/2) + z\sqrt{n/4}$, $z = 1.64$ for $n \leqslant 30$, 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, $\alpha = 1 - \text{BINOMDIST}(x - 1, n, P_0, 1)$ = Type I error, $\beta = \text{BINOMDIST}(x - 1, n, P_{max}, 1)$ = Type II error and $1 - \beta$ = power, NP = No Preference, En = English, and Sw = Kiswahili.

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
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

... Continued on next page

Table B.5 – continued from previous page

SN.	Topic	En	Sw	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
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
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

... Continued on next page

Table B.5 – continued from previous page

SN.	Topic	En	Sw	n	x	P_0	P_d	P_{max}	α	β	$1-\beta$	Decision
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

[1] D. I. Adelani, J. Abbott, G. Neubig, *et al.*, "Masakhaner: Named entity recognition for african languages," *Computing Research Repository (CoRR)*, vol. abs/2103.11811, 2021. [Online]. Available: https://arxiv.org/abs/2103.11811 (cit. on pp. 49, 50).

[2] T. E. African, *Botswana schools to teach Kiswahili*, 2020. [Online]. Available: https://www.theeastafrican.co.ke/tea/news/rest-of-africa/botswana-introduce-swahili-schools-2371206 (visited on 05/20/2021) (cit. on pp. 42, 46).

[3] N. K. Agarwal, "Information source and its relationship with the context of information seeking behavior," in *Proceedings of the 2011 IConference*, ser. iConference '11, Seattle, Washington, USA: ACM, 2011, 48–55, ISBN: 9781450301213. DOI: 10.1145/1940761.1940768 (cit. on p. 14).

[4] Ž. Agić and I. Vulić, "JW300: A wide-coverage parallel corpus for low-resource languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 3204–3210. DOI: 10.18653/v1/P19-1310 (cit. on pp. 55, 57).

[5] M. Agosti, F. Crivellari, and G. M. Di Nunzio, "Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 663–696, 2012. DOI: 10.1007/s10618-011-0228-8 (cit. on p. 10).

[6] M. Aliannejadi, M. Harvey, F. Crestani, L. Costa, and M. Pointon, "Understanding mobile search task relevance and user behaviour in context," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '19, Glasgow, Scotland UK: Association for Computing Machinery, 2019, 143–151, ISBN: 9781450360258. DOI: 10.1145/3295750.3298923 (cit. on pp. 28, 30).

[7] E. Amigó, H. Fang, S. Mizzaro, and C. Zhai, "Are we on the right track? an examination of information retrieval methodologies," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18, Ann Arbor, MI, USA: Association for Computing Machinery, 2018, 997–1000, ISBN: 9781450356572. DOI: 10.1145/3209978.3210131 (cit. on p. 9).

[8] P. Arora, D. Shterionov, Y. Moriya, D. Dzendzik, A. Kaushik, and G. Jones, "An investigative study of multi-modal cross-lingual retrieval," in *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech*, ser. CLSSTS'20, Marseille, France: European Language Resources Association, 2020, pp. 58–67, ISBN: 979-10-95546-55-9. [Online]. Available: https://aclanthology.org/2020.clssts-1.10 (cit. on p. 61).

[9] J. A. Aslam, E. Yilmaz, and V. Pavlu, "A geometric interpretation of r-precision and its correlation with average precision," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '05, Salvador, Brazil: Association for Computing Machinery, 2005, 573–574, ISBN: 1595930345. DOI: 10.1145/1076034.1076134 (cit. on p. 24).

[10] A. Aula and M. Kellar, "Multilingual search strategies," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '09, New York, NY, USA: Association for Computing Machinery, 2009, 3865–3870, ISBN: 9781605582474. DOI: 10.1145/1520340.1520585 (cit. on pp. 4, 30–33, 79, 83, 102).

[11] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM press, New York, 1999, vol. 463 (cit. on pp. 13, 15, 33, 46).

[12] P. Bański and B. Wójtowicz, "A repository of free lexical resources for african languages: The project and the method," in *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*, ser. AfLaT 2009, Athens, Greece: Association for Computational Linguistics, 2009, pp. 89–95. [Online]. Available: https://aclanthology.org/W09-0713.pdf (cit. on pp. 57, 58).

[13] H. Batibo, "The implications of the presence of south-eastern bantu features in kiswahili," *Kioo cha Lugha*, vol. 6, no. 1, 2018. [Online]. Available: http://www.journals.udsm.ac.tz/index.php/kcl/article/viewFile/1420/1318 (cit. on p. 44).

[14] S. M. Beitzel, E. C. Jensen, and O. Frieder, "Mean average precision - map," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 1691–1692. DOI: 10.1007/978-0-387-39940-9_492 (cit. on p. 24).

[15] B. Berendt and A. Kralisch, "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*, vol. 12, no. 3, 380–399, 2009, ISSN: 1386-4564. DOI: 10.1007/s10791-008-9086-4 (cit. on p. 31).

[16] R. Bernard, F. Dulle, and H. Ngalapa, "Assessment of information needs of rice farmers in Tanzania; A case study of Kilombero District, Morogoro," *Library Philosophy and Practice (e-journal)*, 2014. [Online]. Available: http://41.73.194.142/handle/123456789/1164 (visited on 09/02/2020) (cit. on p. 29).

[17] E. Boschee, J. Barry, J. Billa, *et al.*, "SARAL: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 19–24. DOI: 10.18653/v1/P19-3004. [Online]. Available: https://aclanthology.org/P19-3004 (cit. on p. 61).

[18] T. Brants, "Natural language processing in information retrieval," *Computational Linguistics in the Netherlands Journal*, CLIN 14, vol. 111, 2003. [Online]. Available: https://clinjournal.org/CLIN_proceedings/XIV/brants.pdf (cit. on pp. 43, 46).

[19] M. Braschler and P. Schäuble, "Using corpus-based approaches in a system for multilingual information retrieval," *Information Retrieval*, vol. 3, no. 3, pp. 273–284, Oct. 2000. DOI: 10.1023/A:1026525127581 (cit. on pp. 36, 39, 40).

[20] V. Braun and V. Clarke, "To saturate or not to saturate? questioning data saturation as a useful concept for thematic analysis and sample-size rationales," *Qualitative research in sport, exercise and health*, vol. 13, no. 2, pp. 201–216, 2021. DOI: 10.1080/2159676X.2019.1704846 (cit. on p. 66).

[21] G. Buscher, R. W. White, S. Dumais, and J. Huang, "Large-scale analysis of individual and task differences in search result page examination strategies," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12, New York, NY, USA: Association for Computing Machinery, 2012, 373–382, ISBN: 9781450307475. DOI: 10.1145/2124295.2124341 (cit. on p. 139).

[22] S. Büttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, 2010 (cit. on p. 25).

[23] A. CONNEAU and G. Lample, "Cross-lingual language model pretraining," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf (cit. on p. 21).

[24] J. P. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '95, Seattle, Washington, USA: Association for Computing Machinery, 1995, 21–28. DOI: 10.1145/215206.215328 (cit. on p. 34).

[25] B. Carterette, "Precision and recall," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 2126–2127. DOI: 10.1007/978-0-387-39940-9_5050 (cit. on p. 22).

[26] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview," in *Proceedings of the Learning to Rank Challenge*, O. Chapelle, Y. Chang, and T.-Y. Liu, Eds., ser. Proceedings of Machine Learning Research, vol. 14, Haifa, Israel: PMLR, 2011, pp. 1–24. [Online]. Available: http://proceedings.mlr.press/v14/chapelle11a.html (cit. on p. 102).

[27] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09, Hong Kong, China: Association for Computing Machinery, 2009, 621–630, ISBN: 9781605585123. DOI: 10.1145/1645953.1646033 (cit. on pp. 26, 27).

[28] O. Chapelle and Y. Zhang, "A dynamic bayesian network click model for web search ranking," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09, Madrid, Spain: Association for Computing Machinery, 2009, 1–10. DOI: 10.1145/1526709.1526711 (cit. on p. 103).

[29] A. Chen and F. C. Gey, "Combining query translation and document translation in cross-language retrieval," in *Comparative Evaluation of Multilingual Information Access Systems CLEF 2003. Lecture Notes in Computer Science, vol 3237.*, C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, Eds., Berlin, Heidelberg: Springer, 2004, pp. 108–121, ISBN: 978-3-540-30222-3 (cit. on p. 17).

[30] Z. Cheng and J. Shen, "On effective location-aware music recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 2, 2016, ISSN: 1046-8188. DOI: 10.1145/2846092 (cit. on pp. 29, 30).

[31] G. G. Chowdhury, *Introduction to modern information retrieval*. London, UK: Facet publishing, 2010, ISBN: 978-1-85604-694-7 (cit. on p. 14).

[32] P. Chu and A. Komlodi, "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*, ser. CHI EA '17, Denver, Colorado, USA: Association for Computing Machinery, 2017, 2466–2472. DOI: 10.1145/3027063.3053262 (cit. on pp. 6, 19, 30, 37).

[33] A. Chuklin, I. Markov, and M. d. Rijke, "Click models for web search," *Synthesis lectures on information concepts, retrieval, and services*, vol. 7, no. 3, pp. 1–115, Jul. 2015. DOI: 10.2200/S00654ED1V01Y201507ICR043 (cit. on p. 87).

[34] K. Church and B. Smyth, "Understanding mobile information needs," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '08, Amsterdam, The Netherlands: Association for Computing Machinery, 2008, 493–494, ISBN: 9781595939524. DOI: 10.1145/1409240.1409325 (cit. on pp. 28, 30).

[35] M. A. Clarke, J. L. Belden, R. J. Koopman, L. M. Steege, J. L. Moore, S. M. Canfield, and M. S. Kim, "Information needs and information-seeking behaviour analysis of primary care physicians and nurses: A literature review," *Health Information and Libraries Journal*, vol. 30, no. 3, pp. 178–190, 2013. DOI: 10.1111/hir.12036 (cit. on pp. 14, 28, 30).

[36] P. Clough and I. Eleta, "Investigating Language Skills and Field of Knowledge on Multilingual Information Access in Digital Libraries," *International Journal of Digital Library Systems*, vol. 1, no. 1, pp. 89–103, 2010. DOI: 10.4018/jdls.2010102705 (cit. on pp. 31, 32).

[37] P. Clough and M. Sanderson, "Evaluating the performance of information retrieval systems using test collections," *Information Research*, vol. 8, no. 2, 2013. [Online]. Available: http://informationr.net/ir/18-2/paper582.html#.YFct9WgzaUk (cit. on p. 22).

[38] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747 (cit. on p. 49).

[39] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology*, vol. 13, no. 1, pp. 3–21, 1990. DOI: 10.1007/BF00988593 (cit. on p. 68).

[40] N. Craswell, "Precision at n," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 2127–2128, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_484 (cit. on pp. 23, 24).

[41] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08, Palo Alto, California, USA: Association for Computing Machinery, 2008, 87–94. DOI: 10.1145/1341531.1341545 (cit. on p. 103).

[42] M. Creutz, K. Lagus, K. Lindén, and S. Virpioja, "Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages," in *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia, 2005, pp. 107–112. [Online]. Available: https://helda.helsinki.fi/bitstream/handle/10138/29382/Creutz05hlt.pdf?sequence=2 (cit. on pp. 51, 52).

[43] W. B. Croft, "Language models for information retrieval," in *Proceedings of the 19th International Conference on Data Engineering*, ser. Cat. No. 03 CH37405, IEEE, Mar. 2003, pp. 3–7. DOI: 10.1109/ICDE.2003.1260777 (cit. on p. 46).

[44] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Reading: Addison-Wesley, 2010, vol. 520 (cit. on p. 15).

[45] G. D., *Swahili stopwords collection*, 2016. [Online]. Available: https://github.com/stopwords-iso/stopwords-sw (visited on 05/21/2021) (cit. on p. 47).

[46] B. Daille, C. Fabre, and P. Sébillot, "Applications of computational morphology," *Many morphologies*, pp. 210–234, 2002. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.642&rep=rep1&type=pdf (cit. on p. 49).

[47] G. De Pauw and G.-M. De Schryver, "Improving the computational morphological analysis of a swahili corpus for lexicographic purposes," *Lexikos*, vol. 18, 2008. DOI: 10.5788/18-0-488 (cit. on pp. 49–52).

[48] G. De Pauw, G. M. De Schryver, and P. W. Wagacha, "A Corpus-based Survey of Four Electronic Swahili-English Bilingual Dictionaries," *Lexikos*, vol. 19, pp. 340–352, 2009. DOI: 10.5788/19-0-443 (cit. on pp. 43, 59).

[49] G. De Pauw, T. Laureys, and W. Daelemans, "A comparison of two different approaches to morphological analysis of dutch," in *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, ser. SIGPHON, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 62–69. [Online]. Available: https://aclanthology.org/W04-0108 (cit. on p. 50).

[50] G. De Pauw, G.-M. de Schryver, and P. W. Wagacha, "Data-driven part-of-speech tagging of kiswahili," in *International Conference on Text, Speech and Dialogue*, P. Sojka, I. Kopeček, and K. Pala, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 197–204, ISBN: 978-3-540-39091-6. DOI: 10.1007/11846406_25 (cit. on pp. 44, 49–52).

[51] G. De Pauw, P. W. Wagacha, and G.-M. De Schryver, "Exploring the sawa corpus: Collection and deployment of a parallel corpus english—swahili," *Language resources and evaluation*, vol. 45, no. 3, pp. 331–344, 2011. DOI: 10.1007/s10579-011-9159-7 (cit. on p. 55).

[52] G. De Pauw, P. W. Wagacha, and G.-M. de Schryver, "Bootstrapping machine translation for the language pair english - kiswahili," in *Special Topics in Computing and ICT Research - Strengthening the Role of ICT in Development*, 2008, pp. 30–37. [Online]. Available: https://biblio.ugent.be/publication/430436/file/4148635 (cit. on p. 54).

[53] ——, "The sawa corpus: A parallel corpus english-swahili," in *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*, G. De Pauw, G.-M. de Schryver, and L. Levin, Eds., Athens, Greece: Association for Computational Linguistics, 2009, pp. 9–16, ISBN: 9781932432251. [Online]. Available: http://hdl.handle.net/1854/LU-671211 (cit. on pp. 54, 55, 57).

[54] G.-M. De Schryver, "State-of-the-art software to support intelligent lexicography," in *International Seminar on Kangxi Dictionary & Lexicology*, China Sociale Wetenschappen Publishing House, vol. 2, 2010, pp. 584–599. [Online]. Available: https://biblio.ugent.be/publication/1179991/file/6726679.pdf (cit. on pp. 45, 57, 58).

[55] G. M. De Schryver, S. Wolfer, and R. Lew, "The relationship between dictionary look-up frequency and corpus frequency revisited: A log-file analysis of a decade of user interaction with a Swahili-English dictionary," *GEMA Online® Journal of Language Studies*, vol. 19, no. 4, pp. 1–27, 2019, ISSN: 25502131. DOI: 10.17576/gema-2019-1904-01 (cit. on p. 57).

[56] S. Demergazzi, L. Pastore, G. Bassani, M. Arosio, and C. Lonati, "Information needs and information-seeking behavior of italian neurologists: Exploratory mixed methods study," *Journal of Medical Internet Research*, vol. 22, no. 4, 2020, ISSN: 1438-8871. DOI: 10.2196/14979 (cit. on pp. 28, 30).

[57] S. Deng, A. Zhao, S. Fu, Y. Liu, W. Fan, and Y. Jiang, "Music-search behaviour on a social q&a site: A cross-gender comparison," *Journal of Information Science*, vol. 46, no. 4, pp. 560–574, Jul. 2020. DOI: 10.1177/0165551519861605 (cit. on pp. 29, 30).

[58] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Computing Research Repository (CoRR)*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805v2 (cit. on pp. 21, 46, 49).

[59] G. Domingues and C. T. Lopes, "Characterizing and comparing portuguese and english wikipedia medicine-related articles," in *Companion Proceedings of the 2019 World Wide Web Conference*, ser. WWW '19, San Francisco, USA: Association for Computing Machinery, May 2019, 1203–1207. DOI: 10.1145/3308560.3316758 (cit. on p. 102).

[60] M. Dryer and M. Haspelmath, *World Atlas of Language Structures: Language Swahili*, 2014. [Online]. Available: https://wals.info/languoid/lect/wals_code_swa (visited on 05/24/2021) (cit. on p. 45).

[61] G. E. Dupret and B. Piwowarski, "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*, ser. SIGIR '08, New York, NY, USA: Association for Computing Machinery, 2008, 331–338. DOI: 10.1145/1390334.1390392 (cit. on p. 103).

[62] T. Elly and E. E. Silayo, "Agricultural information needs and sources of the rural farmers in Tanzania A case of Iringa rural district," *Library Review*, vol. 62, no. 8/9, pp. 547–566, 2013. DOI: 10.1108/LR-01-2013-0009 (cit. on p. 29).

[63] S. Engine, *Corpus types*, 2021. [Online]. Available: https://www.sketchengine.eu/corpora-and-languages/corpus-types/ (visited on 05/24/2021) (cit. on p. 54).

[64] Ethnologue, *Ethnologue: Languages of the World*, 2021. [Online]. Available: https://www.ethnologue.com/language/swh (visited on 02/24/2021) (cit. on pp. 5, 42, 44, 65).

[65] C. Fluhr, R. E. Frederking, D. Oard, A. Okumura, K. Ishikawa, and K. Satoh, "Multilingual (or cross-lingual) information retrieval," *Proceedings of the Multilingual Information Management: Current Levels and Future Abilities*, vol. Volume XIV-XV, 1999. [Online]. Available: http://www.cs.cmu.edu/~ref/mlim/chapter2.html (cit. on pp. 3, 16, 17).

[66] P. J. L. Frankl, "The indifference to gender in swahili and other bantu languages: Part 2 in consultation with yahya ali omar," *South African Journal of African Languages*, vol. 13, no. 3, pp. 85–89, 1993. DOI: 10.1080/02572117.1993.10586970 (cit. on p. 44).

[67] A. J. Fugard and H. W. Potts, "Supporting thinking on sample sizes for thematic analyses: A quantitative tool," *International Journal of Social Research Methodology*, vol. 18, no. 6, pp. 669–684, 2015. DOI: 10.1080/13645579.2015.1005453 (cit. on p. 66).

[68] W. Gao, C. Niu, M. Zhou, and K.-F. Wong, "Joint ranking for multilingual web search," in *Advances in Information Retrieval*, C. Berrut, M. Boughanem, J. Mothe, and C. Soule-Dupuy, Eds., ser. ECIR 2009. Lecture Notes in Computer Science, Springer, vol. 5478, Berlin, Heidelberg, 2009, pp. 114–125, ISBN: 978-3-642-00958-7. DOI: 10.1007/978-3-642-00958-7_13 (cit. on pp. 37, 40).

[69] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of swahili resources for an automatic speech recognition system," in *Spoken Language Technologies for Under-Resourced Languages*, ser. SLTU-2012, 2012, pp. 94–101. [Online]. Available: https://www.isca-speech.org/archive/sltu_2012/papers/su12_094.pdf (cit. on pp. 44, 53, 55).

[70] M. R. Ghorab, D. Zhou, S. Lawless, and V. Wade, "Multilingual user modeling for personalized re-ranking of multilingual web search results," in *CEUR Workshop Proceedings – 20th Conference on User Modeling, Adaptation, and Personalization, UMAP 2012*, vol. 872, Montreal QC, 2012. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.552&rep=rep1&type=pdf (cit. on p. 4).

[71] M. R. Ghorab, D. Zhou, B. Steichen, and V. Wade, "Towards multilingual user models for personalized multilingual information retrieval," in *Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval*, ser. PMHR '11, Eindhoven, Netherlands: Association for Computing Machinery, 2011, 42–49. DOI: 10.1145/2047403.2047411 (cit. on p. 7).

[72] I. Gialampoukidis, A. Moumtzidou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Retrieval of multimedia objects by fusing multiple modalities," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16, New York, New York, USA: Association for Computing Machinery, 2016, 359–362, ISBN: 9781450343596. DOI: 10.1145/2911996.2912068 (cit. on pp. 5, 38).

[73] A. Goyal, M. Kumar, and V. Gupta, "Named entity recognition: Applications, approaches and challenges," *International Journal of Advance Research in Science and Engineering*, vol. 6, no. 10, pp. 1902–1916, 2017. [Online]. Available: http://ijarse.com/images/fullpdf/1508996434_GNC849ijarse.pdf (cit. on p. 48).

[74] A. Grotov, A. Chuklin, I. Markov, L. Stout, F. Xumara, and M. de Rijke, "A comparative study of click models for web search," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, Eds., ser. CLEF 2015. Lecture Notes in Computer Science, vol. 9283, Cham: Springer International Publishing, 2015, pp. 78–90. DOI: 10.1007/978-3-319-24027-5_7 (cit. on p. 88).

[75] M. Groves and K. Mundt, "Friend or foe? google translate in language for academic purposes," *English for Specific Purposes*, vol. 37, pp. 112–121, 2015. DOI: 10.1016/j.esp.2014.09.001 (cit. on p. 54).

[76] V. N. Gudivada and K. Arbabifard, "Chapter 3 - open-source libraries, application frameworks, and workflow systems for nlp," in *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, ser. Handbook of Statistics, V. N. Gudivada and C. Rao, Eds., vol. 38, B.V.: Elsevier, 2018, pp. 31–50. DOI: 10.1016/bs.host.2018.07.007 (cit. on p. 48).

[77] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos, "Click chain model in web search," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09, Madrid, Spain: Association for Computing Machinery, 2009, 11–20. DOI: 10.1145/1526709.1526712 (cit. on p. 103).

[78] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais, "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*, ser. SIGIR '11, Beijing, China: Association for Computing Machinery, 2011, 335–344. DOI: 10.1145/2009916.2009964 (cit. on p. 30).

[79] W. Guo, H. Gao, J. Shi, B. Long, L. Zhang, B.-C. Chen, and D. Agarwal, "Deep natural language processing for search and recommender systems," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, Anchorage, AK, USA: Association for Computing Machinery, 2019, 3199–3200. DOI: 10.1145/3292500.3332290 (cit. on p. 46).

[80] D. Hiemstra, "Information retrieval models," in *Information Retrieval*, A. Goker and J. Davies, Eds., John Wiley & Sons, Ltd, 2009, ch. 1, pp. 1–19, ISBN: 9780470033647. DOI: 10.1002/9780470033647.ch1 (cit. on p. 15).

[81] D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld, "Translation resources, merging strategies, and relevance feedback for cross-language information retrieval," in *Cross-Language Information Retrieval and Evaluation*, C. Peters, Ed., ser. CLEF 2000. Lecture Notes in Computer Science, vol. 2069, Berlin, Heidelberg: Springer, 2001, pp. 102–115, ISBN: 978-3-540-44645-3. DOI: 10.1007/3-540-44645-1_10 (cit. on p. 34).

[82] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119, PMLR, 2020, pp. 4411–4421. [Online]. Available: http://proceedings.mlr.press/v119/hu20b.html (cit. on p. 48).

[83] A. Hurskainen, "A two-level computer formalism for the analysis of bantu morphology. an application to swahili," *Nordic journal of African studies*, vol. 1, no. 1, 1992. [Online]. Available: https://njas.fi/njas/article/view/61 (cit. on p. 43).

[84] ——, "Computational testing of five swahili dictionaries," in *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki, Finland, 2004. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.9034&rep=rep1&type=pdf (cit. on p. 43).

[85] ——, "Helsinki corpus of swahili," *Technical Report, Compiler: Institute for Asian and African Studies (University of Helsinki) and CSC*, 2004 (cit. on pp. 50, 54, 56).

[86] ——, "Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications," *Nordic Journal of African Studies*, vol. 13, no. 3, pp. 363–397, 2004. [Online]. Available: https://www.njas.fi/njas/article/view/293 (cit. on pp. 44, 50–52, 58).

[87] ——, *Helsinki corpus of swahili 2.0 (hcs 2.0) annotated version*, 2016. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2016011301 (visited on 03/22/2021) (cit. on pp. 55, 56).

[88] ——, "Challenges of language technology of kiswahili," *Kioo cha Lugha*, vol. 9, no. 1, 2018. [Online]. Available: http://www.journals.udsm.ac.tz/index.php/kcl/article/viewFile/1513/1411 (cit. on pp. 46, 52).

[89] ISO, *ISO 5964:1985 Documentation – Guidelines for the establishment and development of multilingual thesauri*, 1985. [Online]. Available: https://www.iso.org/standard/12159.html (visited on 02/02/2021) (cit. on p. 18).

[90] ——, *ISO 25964-2:2013 Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies*, 2013. [Online]. Available: https://www.iso.org/standard/53658.html (visited on 02/02/2021) (cit. on p. 18).

[91] R. Ikoja-Odongo and D. N. Ocholla, "Information seeking behavior of the informal sector entrepreneurs: The uganda experience," *Libri*, vol. 54, no. 1, pp. 54–66, 2004. DOI: 10.1515/LIBR.2004.54 (cit. on p. 29).

[92] M. S. Islam and S. M. Zabed Ahmed, "The information needs and information-seeking behaviour of rural dwellers: A review of research," *IFLA Journal*, vol. 38, no. 2, pp. 137–147, 2012. DOI: 10.1177/0340035212444513 (cit. on pp. 28, 30).

[93] B. J. Jansen and A. Spink, "How are we searching the world wide web? a comparison of nine search engine transaction logs," *Information processing & management*, vol. 42, no. 1, pp. 248–263, 2006. DOI: 10.1016/j.ipm.2004.10.007 (cit. on p. 85).

[94] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, 422–446, Oct. 2002, ISSN: 1046-8188. DOI: 10.1145/582415.582418 (cit. on pp. 25, 114, 126).

[95] R. S. Jhangiani, I.-C. A. Chiang, C. Cuttler, and D. C. Leighton, *Research methods in psychology*, 4th ed. Kwantlen Polytechnic University, 2019 (cit. on p. 10).

[96] J. Jimmy, G. Zuccon, B. Koopman, and G. Demartini, "Health cards for consumer health search," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19, Paris, France: Association for Computing Machinery, 2019, 35–44, ISBN: 9781450361729. DOI: 10.1145/3331184.3331194 (cit. on p. 87).

[97] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," *SIGIR Forum*, vol. 51, no. 1, 4–11, 2017, ISSN: 0163-5840. DOI: 10.1145/3130332.3130334 (cit. on p. 88).

[98] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, ser. WSDM '17, Cambridge, United Kingdom: Association for Computing Machinery, 2017, 781–789. DOI: 10.1145/3018661.3018699 (cit. on p. 88).

[99] D. Joffe, M. MacLeod, and G.-M. De Schryver, "Software demonstration: The tshwanelex electronic dictionary system," in *13th EURALEX International Congress*, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, 2008, pp. 421–424. [Online]. Available: https://biblio.ugent.be/publication/430431/file/6810491 (cit. on pp. 57, 58).

[100] F. Kalimi, *Twitter now speaks Swahili. Poa sana!* 2018. [Online]. Available: https://edition.cnn.com/2018/05/09/africa/twitter-swahili-african/index.html (visited on 03/21/2021) (cit. on p. 46).

[101] T. Kaloki, *Facebook launches in Swahili*, 2009. [Online]. Available: https://www.bizcommunity.com/Article/414/23/37112.html (visited on 03/21/2021) (cit. on p. 46).

[102] S. G. Kanakaraddi and S. S. Nandyal, "Survey on parts of speech tagger techniques," in *2018 International Conference on Current Trends towards Converging Technologies*, ser. ICCTCT, Coimbatore, India: IEEE, 2018, pp. 1–6. DOI: 10.1109/ICCTCT.2018.8550884 (cit. on p. 49).

[103] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," *Machine Translation*, vol. 32, no. 1, pp. 167–189, 2018. DOI: 10.1007/s10590-017-9203-5 (cit. on p. 84).

[104] D. Kassab and X. Yuan, "Understanding Information Needs and Search Behaviors of Mobile Users," *Information Research*, vol. 17, no. 4, pp. 1–10, Dec. 2012. [Online]. Available: http://InformationR.net/ir/17-4/paper551.html (cit. on pp. 28, 30).

[105] E. S. Kayi, V. Anand, and S. Muresan, "MultiSeg: Parallel data and subword information for learning bilingual embeddings in low resource scenarios," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, Marseille, France, 2020, pp. 97–105, ISBN: 979-10-95546-35-1. [Online]. Available: https://aclanthology.org/2020.sltu-1.13 (cit. on p. 53).

[106] T. Khumalo, *South African Schools to Teach Kiswahili in 2020*, 2018. [Online]. Available: https://www.voanews.com/africa/south-african-schools-teach-kiswahili-2020 (visited on 05/20/2021) (cit. on pp. 42, 46).

[107] M. Kisilowska, "Informational priorities in health information systems," in *Health Information Systems: Concepts, Methodologies, Tools, and Applications*, A. N. Dwivedi, Ed., IGI Global, 2009, 763–781. DOI: 10.4018/978-1-60566-988-5.ch030 (cit. on p. 13).

[108] A. K. Kozorovitsky and O. Kurland, "Cluster-based fusion of retrieved lists," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11, Beijing, China: Association for Computing Machinery, 2011, 893–902. DOI: 10.1145/2009916.2010035 (cit. on p. 38).

[109] A. Kralisch and T. Mandl, "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*, ser. HICSS'06, vol. 3, Kauai, HI, USA: IEEE, 2006, 54b–54b. DOI: 10.1109/HICSS.2006.71 (cit. on pp. 4, 32, 33, 102).

[110] H. T. Lawless and H. Heymann, *Sensory evaluation of food: principles and practices*. Springer Science & Business Media, 2013 (cit. on p. 90).

[111] A. Le Calvé and J. Savoy, "Database merging strategy based on logistic regression," *Information Processing & Management*, vol. 36, no. 3, pp. 341–359, 2000. DOI: 10.1016/S0306-4573(99)00036-9 (cit. on pp. 5, 37, 39, 40, 106).

[112] J. H. Lee and J. S. Downie, "Survey of Music Information Needs, Uses, and Seeking Behaviours: Preliminary Findings," in *Proceedings of the 5th International Conference on Music Information Retrieval*, ser. ISMIR 2004, Barcelona, Spain, Oct. 2004, pp. 441–446. [Online]. Available: https://archives.ismir.net/ismir2004/paper/000232.pdf (cit. on pp. 29, 30).

[113] H. Li, "Learning to rank for information retrieval and natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 7, no. 3, pp. 1–121, 2014. DOI: 10.2200/S00607ED2V01Y201410HLT026 (cit. on pp. 15, 33).

[114] S. Liang, I. Markov, Z. Ren, and M. de Rijke, "Manifold learning for rank aggregation," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, 1735–1744, ISBN: 9781450356398. DOI: 10.1145/3178876.3186085 (cit. on p. 38).

[115] S. Liang and M. de Rijke, "Burst-aware data fusion for microblog search," *Information Processing & Management*, vol. 51, no. 2, pp. 89–113, 2015, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.008 (cit. on p. 38).

[116] W.-C. Lin and H.-H. Chen, "Merging mechanisms in multilingual information retrieval," in *Advances in Cross-Language Information Retrieval*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., ser. CLEF 2002. Lecture Notes in Computer Science, vol. 2785, Berlin, Heidelberg: Springer,

2003, pp. 175–186, ISBN: 978-3-540-45237-9. DOI: 10.1007/978-3-540-45237-9_14 (cit. on pp. 4, 19, 20, 34, 35, 39, 106).

[117] W. Lin and H. Chen, "Merging multilingual information retrieval results based on prediction of retrieval effectiveness," in *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, ser. NTCIR-4, National Center of Sciences, Tokyo, Japan: National Institute of Informatics (NII), 2004. [Online]. Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/CLIR/NTCIR4-CLIR-LinWC.pdf (cit. on pp. 36, 39, 40).

[118] K. Lindén, "A probabilistic model for guessing base forms of new words by analogy," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed., ser. CICLing 2008. Lecture Notes in Computer Science, Springer, vol. 4919, Berlin, Heidelberg, 2008, pp. 106–116, ISBN: 978-3-540-78135-6. DOI: 10.1007/978-3-540-78135-6_10 (cit. on p. 51).

[119] C. Ling, B. Steichen, and A. G. Choulos, "A comparative user study of interactive multilingual search interfaces," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, ser. CHIIR '18, New Brunswick, NJ, USA: Association for Computing Machinery, 2018, 211–220, ISBN: 9781450349253. DOI: 10.1145/3176349.3176383 (cit. on pp. 30, 31, 37, 79, 87, 139).

[120] C. Ling, B. Steichen, and S. Figueira, "Multilingual news – an investigation of consumption, querying, and search result selection behaviors," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 516–535, 2020. DOI: 10.1080/10447318.2019.1662636 (cit. on pp. 4, 28, 31, 32, 83, 84).

[121] P. Littell, K. Price, and L. S. Levin, "Morphological parsing of Swahili using crowdsourced lexical resources," in *The 9th Edition of the Language Resources and Evaluation Conference*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3333–3339, ISBN: 978-2-9517408-8-4. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/896_Paper.pdf (cit. on p. 51).

[122] K.-L. Liu, W. Meng, J. Qiu, C. Yu, V. Raghavan, Z. Wu, Y. Lu, H. He, and H. Zhao, "Allinonenews: Development and evaluation of a large-scale news metasearch engine," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '07, Beijing, China: Association for Computing Machinery, 2007, 1017–1028. DOI: 10.1145/1247480.1247601 (cit. on pp. 5, 38).

[123] T.-Y. Liu, *Learning to rank for information retrieval*. Springer Science & Business Media, 2011 (cit. on pp. 33, 37, 88, 103, 106).

[124] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *Computing Research Repository (CoRR)*, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692 (cit. on p. 46).

[125] Y. Liu, J. Miao, M. Zhang, S. Ma, and L. Ru, "How do users describe their information need: Query recommendation based on snippet click model," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13 847–13 856, 2011, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2011.04.188 (cit. on p. 139).

[126] R. T.-W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," in *Proceedings of the 5th Dutch-Belgian Information Retrieval Workshop*, R. v. Zwol, Ed., ser. DIR'05, vol. 5, Utrecht, the Netherlands: Center for Content and Knowledge Engineering, 2005, pp. 17–24, ISBN: 90-393-0031-3. [Online]. Available: http://www.dcs.gla.ac.uk/~rachel/publications/rtlo_DIR05.pdf (cit. on p. 47).

[127] B. Long, J. Ye, Z. Li, H. Gao, and S. K. Jha, "Deep natural language processing for search and recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20, Virtual Event, China: Association for Computing Machinery, 2020, 2461–2463. DOI: 10.1145/3397271.3401465 (cit. on p. 46).

[128] C. T. Lopes and C. Ribeiro, "Measuring the value of health query translation: An analysis by user language proficiency," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 5, pp. 951–963, 2013. DOI: 10.1002/asi.22812 (cit. on p. 102).

[129] R. Lowe and B. Steichen, "Multilingual Search User Behaviors – Exploring Multilingual Querying and Result Selection Through Crowdsourcing," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '17, Bratislava, Slovakia: Association for Computing Machinery, 2017, pp. 303–307. DOI: 10.1145/3079628.3079702 (cit. on pp. 6, 31–33, 84, 102).

[130] E. T. Lwoga, P. Ngulube, and C. Stilwell, "Information needs and information seeking behaviour of small-scale farmers in Tanzania," *Innovation: Journal of appropriate librarianship and information work in Southern Africa*, vol. 40, pp. 80–103, 2010. [Online]. Available: https://hdl.handle.net/10520/EJC46504 (cit. on p. 29).

[131] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1064–1074. DOI: 10.18653/v1/P16-1101 (cit. on p. 49).

[132] C. D. Manning, H. Schütze, and P. Raghavan, *Introduction to information retrieval*. Cambridge university press, 2008 (cit. on p. 22).

[133] J. Marlow, P. Clough, J. C. Recuero, and J. Artiles, "Exploring the effects of language skills on multilingual web search," in *Advances in Information Retrieval*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, Eds., ser. ECIR 2008, vol. 4956, Berlin, Heidelberg: Springer, 2008, pp. 126–137, ISBN: 978-3-540-78646-7. DOI: 10.1007/978-3-540-78646-7_14 (cit. on p. 31).

[134] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named entity recognition: Fallacies, challenges and opportunities," *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482–489, 2013. DOI: 10.1016/j.csi.2012.09.004 (cit. on p. 48).

[135] G. L. Martin, M. E. Mswahili, and Y.-S. Jeong, "Sentiment classification in swahili language using multilingual bert," *Computing Research Repository (CoRR)*, 2021. [Online]. Available: https://arxiv.org/abs/2104.09006 (cit. on p. 54).

[136] F. Martínez-Santiago, A. Urena-López, and M. Martín-Valdivia, "A merging strategy proposal: The 2-step retrieval status value method," *Information Retrieval*, vol. 9, no. 1, pp. 71–93, 2006. DOI: 10.1007/s10791-005-5722-4 (cit. on pp. 36, 39, 40).

[137] B. Masua and N. Masasi, "Enhancing text pre-processing for swahili language: Datasets for common swahili stop-words, slangs and typos with equivalent proper words," *Data in Brief*, vol. 33, p. 106 517, 2020. DOI: 10.1016/j.dib.2020.106517 (cit. on pp. 47, 55, 56).

[138] M. C. Meilgaard, G. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques*, 5th ed. CRC Press, 2016, ISBN: 13:978-1-4822-1691-2 (cit. on pp. 90, 91).

[139] M. A. Mohamed, *Modern Swahili Grammar*. East African Publishers, 2001 (cit. on p. 44).

[140] J. Monti, M. Monteleone, M. P. di Buono, and F. Marano, "Cross-lingual information retrieval and semantic interoperability for cultural heritage repositories," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2013, pp. 483–490. [Online]. Available: https://aclanthology.org/R13-1063 (cit. on p. 61).

[141] W. P. Mtega, "Access to and Usage of Information among Rural Communities: a Case Study of Kilosa District Morogoro Region in Tanzania," *Partnership - The Canadian Journal of Library and Information Practice and Research*, vol. 7, no. 1, pp. 1–13, 2012. DOI: 10.21083/partnership.v7i1.1646 (cit. on p. 29).

[142] D. Mueller, N. Andrews, and M. Dredze, "Sources of transfer in multilingual named entity recognition," *Computing Research Repository (CoRR)*, vol. abs/2005.00847, 2020. [Online]. Available: https://arxiv.org/abs/2005.00847 (cit. on pp. 49, 50).

[143] W. Mueller, T. H. Silva, J. M. Almeida, and A. A. Loureiro, "Gender matters! analyzing global cultural gender preferences for venues using social sensing," *EPJ Data Science*, vol. 6, no. 1, p. 5, 2017. DOI: 10.1140/epjds/s13688-017-0101-0 (cit. on p. 33).

[144] MustGo, *About World Languages*, 2021. [Online]. Available: https://www.mustgo.com/worldlanguages/swahili/ (visited on 02/26/2021) (cit. on pp. 44, 65).

[145] K. Mwendera, *Karibu! SADC Adopts KiSwahili As An Official Language*, 2019. [Online]. Available: https://www.forbesafrica.com/current-affairs/2019/08/26/karibu-sadc-adopts-kiswahili-as-an-official-language/ (visited on 05/24/2021) (cit. on p. 46).

[146] NPM, *Stopwords for multiple languages*, 2015. [Online]. Available: https://www.npmjs.com/package/multi-stopwords (visited on 05/24/2021) (cit. on p. 47).

[147] W. Nekoto, V. Marivate, T. Matsila, *et al.*, "Participatory research for low-resourced machine translation: A case study in african languages," *Computing Research Repository (CoRR)*, vol. abs/2010.02353, 2020. [Online]. Available: https://arxiv.org/abs/2010.02353 (cit. on pp. 56, 61).

[148] D. Ngonyani, "Language shift and national identity in tanzania," *Ufahamu: A Journal of African Studies*, vol. 23, no. 2, 1995. [Online]. Available: https://escholarship.org/content/qt8072719q/qt8072719q.pdf (cit. on pp. 5, 80).

[149] J.-Y. Nie and F. Jin, "A multilingual approach to multilingual information retrieval," in *A multilingual approach to multilingual information retrieval*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., ser. CLEF 2002. Lecture Notes in Computer Science, Springer, vol. 2785, Heidelberg, Berlin, 2002, pp. 101–110. DOI: 10.1007/978-3-540-45237-9_8 (cit. on pp. 4, 16, 17, 19–21, 36, 40, 54, 60).

[150] S. Niu, J. Guo, Y. Lan, and X. Cheng, "Top-k learning to rank: Labeling, ranking and evaluation," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '12, Portland, Oregon, USA: Association for Computing Machinery, 2012, 751–760. DOI: 10.1145/2348283.2348384 (cit. on pp. 23, 25, 26, 114, 126).

[151] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, 2013. DOI: 10.1016/j.artint.2012.03.006 (cit. on p. 48).

[152] P. Nzomo, I. Ajiferuke, L. Vaughan, and P. McKenzie, "Multilingual Information Retrieval & Use: Perceptions and Practices Amongst Bi/Multilingual Academic Users," *The Journal of Academic Librarianship*, vol. 42, no. 5, pp. 495–502, 2016. DOI: 10.1016/j.acalib.2016.06.012 (cit. on pp. 4, 17, 31, 32).

[153] P. Nzomo, L. Vaughan, I. Ajiferuke, and P. McKenzie, "Multilingual information access (mlia) tools on google and worldcat: Bi/multilingual university students' experience and perceptions," *Journal of Library Administration*, vol. 59, no. 8, pp. 831–853, 2019. DOI: 10.1080/01930826.2019.1661750 (cit. on pp. 6, 19).

[154] D. Ochieng, "The revival of the status of english in tanzania: What future does the status of the english language have in tanzania?" *English Today*, vol. 31, no. 2, pp. 25–31, 2015. DOI: 10.1017/S0266078415000073 (cit. on p. 5).

[155] A. M. Oirere, R. R. Deshmukh, P. P. Shrishrimal, and V. B. Waghmare, "Swahili Text and Speech Corpus: A Review," *Asian Journal of Computer Science and Information Technology*, vol. 2, no. 11, pp. 286–290, 2012. [Online]. Available: http://repository.mut.ac.ke:8080/xmlui/handle/123456789/3002 (cit. on p. 43).

[156] N. Pakenham-Walsh and F. Bukachi, "Information needs of health care workers in developing countries: A literature review with a focus on Africa," *Human Resources for Health*, vol. 7, no. 30, 2009. DOI: 10.1186/1478-4491-7-30 (cit. on p. 29).

[157] G. Paltoglou, M. Salampasis, and M. Satratzemi, "Hybrid results merging," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ser. CIKM '07, Lisbon, Portugal: Association for Computing Machinery, 2007, 321–330, ISBN: 9781595938039. DOI: 10.1145/1321440.1321487 (cit. on pp. 34–36, 39, 40).

[158] C. Peters, M. Braschler, and P. Clough, *Multilingual Information Retrieval - From Research to Practice*. Springer Science & Business Media, 2012 (cit. on pp. 3, 4, 15–19, 34, 43, 54, 60, 84, 106).

[159] D. Petrelli, S. Levin, M. Beaulieu, and M. Sanderson, "Which user interaction for cross-language information retrieval? design issues and reflections," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 5, pp. 709–722, 2006. DOI: 10.1002/asi.20332 (cit. on pp. 31, 33).

[160] A. Phiri, G. T. Chipeta, and W. D. Chawinga, "Information needs and barriers of rural smallholder farmers in developing countries: A case study of rural smallholder farmers in malawi," *Information Development*, vol. 35, no. 3, pp. 421–434, 2019. DOI: 10.1177/0266666918755222 (cit. on pp. 29, 30).

[161] E. C. Polomé, *Swahili language handbook.* Education Resources Information Center (ERIC), 1967 (cit. on p. 44).

[162] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles, "The impact of database selection on distributed searching," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '00, Athens, Greece: Association for Computing Machinery, 2000, 232–239, ISBN: 1581132263. DOI: 10.1145/345508.345584 (cit. on pp. 34, 39).

[163] T. Qin, X. Geng, and T.-Y. Liu, "A new probabilistic model for rank aggregation," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., ser. NIPS'10, Vancouver, British Columbia, Canada: Curran Associates Inc., Dec. 2010, 1948–1956.

[Online]. Available: https://dl.acm.org/doi/10.5555/2997046.2997113 (cit. on p. 38).

[164] T. Qin and T.-Y. Liu, "Introducing letor 4.0 datasets," *arXiv preprint arXiv:1306.2597*, 2013 (cit. on p. 5).

[165] E. Rabinovich, O. Rom, and O. Kurland, "Utilizing relevance feedback in fusion-based retrieval," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '14, Gold Coast, Queensland, Australia: Association for Computing Machinery, 2014, 313–322. DOI: 10.1145/2600428.2609573 (cit. on p. 38).

[166] R. Rahimi, A. Shakery, and I. King, "Multilingual information retrieval in the language modeling framework," *Information Retrieval Journal*, vol. 18, no. 3, pp. 246–281, May 2015. DOI: 10.1007/s10791-015-9255-1 (cit. on pp. 3, 4, 16, 19, 21, 39, 114).

[167] E. Rasmussen, "Stoplists," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., New York, NY: Springer New York, 2018, pp. 3721–3722, ISBN: 978-1-4614-8265-9. DOI: 10.1007/978-1-4614-8265-9_955 (cit. on p. 47).

[168] Y. Rasolofo, F. Abbaci, and J. Savoy, "Approaches to collection selection and results merging for distributed information retrieval," in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ser. CIKM '01, Atlanta, Georgia, USA: Association for Computing Machinery, 2001, 191–198, ISBN: 1581134363. DOI: 10.1145/502585.502618 (cit. on pp. 35, 39, 40).

[169] M. K. Rather and A. G. Shabir, "Information needs of users in the tech savvy environment and the influencing factors," in *Encyclopedia of Information Science and Technology*, M. Khosrow-Pour, Ed., vol. 4, IGI Global, 2018, pp. 2264–2279. DOI: 10.4018/978-1-5225-2255-3.ch197 (cit. on p. 13).

[170] H.-Y. Rieh and S. Y. Rieh, "Web searching across languages: Preference and behavior of bilingual academic users in korea," *Library & information science research*, vol. 27, no. 2, pp. 249–263, 2005, ISSN: 0740-8188. DOI: 10.1016/j.lisr.2005.01.006 (cit. on pp. 31, 33).

[171] T. Russell-Rose and M. Stevenson, "The role of natural language processing in information retrieval: Searching for meaning and structure," in *Information Retrieval: Searching in the 21st Century*, A. Goker and J. Davies, Eds., Wiley Online Library, 2009, pp. 215–231, ISBN: 978-0-470-03363-0 (cit. on p. 46).

[172] T. E. Rutalemwa and D. P. Theodorus, "The prospects of kiswahili as a medium of instruction in the tanzanian education and training policy," *Journal of Language and Education*, vol. 4, no. 3, pp. 88–98, 2018. DOI: 10.17323/2411-7390-2018-4-3-88-98 (cit. on pp. 5, 65, 75, 79, 80, 100).

[173] T. Sakai, "Expected reciprocal rank," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., New York, NY: Springer, 2017, pp. 1–2, ISBN: 978-1-4899-7993-3. DOI: 10.1007/978-1-4899-7993-3_80617-1 (cit. on p. 26).

[174] G. Salton, *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968, ISBN: 0070544859 (cit. on pp. 13, 15, 46).

[175] G. Salton, "Automatic processing of foreign language documents," *Journal of the American Society for Information Science*, vol. 21, no. 3, pp. 187–194, 1970. DOI: 10.1002/asi.4630210305 (cit. on p. 17).

[176] F. Sánchez-Martínez, V. M. Sánchez-Cartagena, J. A. Pérez-Ortiz, M. L. Forcada, M. Esplà-Gomis, A. Secker, S. Coleman, and J. Wall, "An English-Swahili parallel corpus and its use for neural machine translation in the news domain," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 299–308. [Online]. Available: https://aclanthology.org/2020.eamt-1.32 (cit. on pp. 55–57, 61).

[177] M. Sanderson, "Ambiguous queries: Test collections need more sense," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08, Singapore, Singapore: Association for Computing Machinery, 2008, 499–506, ISBN: 9781605581644. DOI: 10.1145/1390334.1390420 (cit. on p. 85).

[178] M. Sanderson and W. B. Croft, "The history of information retrieval research," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, 2012. DOI: 10.1109/JPROC.2012.2189916 (cit. on p. 13).

[179] J. Savoy, "Report on clef-2001 experiments: Effective combined query-translation approach," in *Evaluation of Cross-Language Information Retrieval Systems*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., ser. CLEF 2001. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2002, pp. 27–43, ISBN: 978-3-540-45691-9. DOI: 10.1007/3-540-45691-0_3 (cit. on pp. 34, 39).

[180] ——, "Combining multiple strategies for effective monolingual and cross-language retrieval," *Information Retrieval*, vol. 7, no. 1-2, pp. 121–148, 2004. DOI: 10.1023/B:INRT.0000009443.51912.e7 (cit. on pp. 35, 37, 39, 40).

[181] J. Savoy, A. Le Calvé, and D. Vrajitoru, "Report on the trec-5 experiment: Data fusion and collection fusion," *NIST Special Publication*, pp. 489–502, 1997. [Online]. Available: https://cs.iusb.edu/~danav/papers/trec5.pdf (visited on 04/17/2021) (cit. on pp. 4, 5, 34, 39, 106).

[182] T. C. Schadeberg, "Loanwords in swahili," in *Loanwords in the world's languages*, Berlin, New York: De Gruyter Mouton, 2009, pp. 76–102. DOI: 10.1515/9783110218442.76 (cit. on pp. 44–46).

[183] H. Schmid, "Part-of-speech tagging with neural networks," *Computing Research Repository (CoRR)*, vol. abs/cmp-lg/9410018, 1994. [Online]. Available: http://arxiv.org/abs/cmp-lg/9410018 (cit. on p. 49).

[184] C. D. Schultz, "Do web site visitors vary in their search and surf behavior?" In *Encyclopedia of E-Commerce Development, Implementation, and Management*, I. Lee, Ed., IGI Global, 2016, pp. 1582–1592. DOI: 10.4018/978-1-4666-9787-4.ch112 (cit. on p. 14).

[185] R. Shah, B. Lin, A. Gershman, and R. Frederking, "Synergy: A named entity recognition system for resource-scarce languages such as swahili using online machine translation," in *Proceedings of the Second Workshop on African Language Technology*, G. De Pauw, H. Groenewald, and G.-M. de Schryver, Eds., ser. AfLaT 2010, Valletta, Malta, May 2010, pp. 21–26, ISBN: 2-9517408-6-7. [Online]. Available: https://biblio.ugent.be/publication/1851705/file/6736544#page=33 (cit. on pp. 48–50, 55, 56).

[186] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell, "Lambdamerge: Merging the results of query reformulations," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, Hong Kong, China: Association for Computing Machinery, 2011, 795–804. DOI: 10.1145/1935826.1935930 (cit. on p. 38).

[187] C. S. Shikali and R. Mokhosi, "Enhancing African low - resource languages: Swahili data for language modelling," *Data in Brief*, vol. 31, 2020. DOI: 10.1016/j.dib.2020.105951 (cit. on pp. 55, 56).

[188] C. S. Shikali, R. Mokhosi, Z. Shijie, and L. Qihe, "Learning syllables using conv-lstm model for swahili word representation and part-of-speech tagging," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 4, pp. 1–25, 2021, ISSN: 2375-4699. DOI: 10.1145/3445975 (cit. on pp. 51, 52).

[189] C. S. Shikali, Z. Sijie, L. Qihe, and R. Mokhosi, "Better word representation vectors using syllabic alphabet: A case study of swahili," *Applied Sciences*, vol. 9, no. 18, 2019. DOI: 10.3390/app9183648 (cit. on pp. 51–53).

[190] L. Si, Q. Pan, and X. Zhuang, "An empirical analysis of user behaviour on multilingual information retrieval," *The Electronic Library*, vol. 35, no. 3, pp. 410–426, 2017, ISSN: 0264-0473. DOI: 10.1108/EL-01-2016-0004 (cit. on pp. 4, 32).

[191] L. Si and J. Callan, "Clef 2005: Multilingual retrieval by combining multiple multilingual ranked lists," in *Accessing Multilingual Information Repositories*, C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, Eds., ser. CLEF 2005. Lecture Notes in Computer Science, vol. 4022, Vienna, Austria: Springer Berlin, Heidelberg, 2006, pp. 121–130, ISBN: 978-3-540-45700-8. DOI: 10.1007/11878773_13 (cit. on pp. 37, 39, 40).

[192] T. N. B. of Statistics, *The National Sample Census of Agriculture, Livestock and Fisheries for Agricultural Year 2019/20*, 2020. [Online]. Available: https://www.nbs.go.tz/index.php/en/census-surveys/agriculture-statistics/579-press-release-key-findings-on-the-national-sample-census-of-agriculture-livestock-and-fisheries-for-agricultural-year-2019-20 (visited on 02/16/2021) (cit. on p. 80).

[193] I. W. Stats, *World Internet Usage and Population Statistics*, 2021. [Online]. Available: https://www.internetworldstats.com/stats.htm (visited on 05/26/2021) (cit. on p. 17).

[194] ——, *World Internet Usage and Population Statistics*, 2021. [Online]. Available: https://www.internetworldstats.com/stats7.htm (visited on 05/26/2021) (cit. on p. 18).

[195] B. Steichen and L. Freund, "Supporting the modern polyglot: A comparison of multilingual search interfaces," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15, Seol, Korea: Association for Computing Machinery, 2015, 3483–3492, ISBN: 9781450331456. DOI: 10.1145/2702123.2702541 (cit. on pp. 7, 31, 32).

[196] B. Steichen, M. R. Ghorab, A. O'Connor, S. Lawless, and V. Wade, "Towards personalized multilingual information access-exploring the browsing and search behavior of multilingual users," in *International Conference on User Modeling, Adaptation, and Personalization*, V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, and G.-J. Houben, Eds., Cham: Springer International Publishing, 2014, pp. 435–446, ISBN: 978-3-319-08786-3. DOI: 10.1007/978-3-319-08786-3_39 (cit. on pp. 4, 6, 28, 31–33).

[197] B. Steichen and R. Lowe, "How do multilingual users search? an investigation of query and result list language choices," *Journal of the Association for Information Science and Technology*, vol. 72, no. 6, pp. 759–776, 2021. DOI: 10.1002/asi.24443 (cit. on pp. 4, 31–33, 79).

[198] R. Steinberger, S. Ombuya, M. Kabadjov, B. Pouliquen, L. Della Rocca, J. Belyaeva, M. de Paola, C. Ignat, and E. Van der Goot, "Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili," *Language Resources and Evaluation*, vol. 45, no. 3, pp. 311–330, 2011. DOI: 10.1007/s10579-011-9155-y (cit. on pp. 48–50).

[199] S. M. Strassel, A. Bies, and J. Tracey, "Situational awareness for low resource languages: The lorelei situation frame annotation task," in *SMERP Data Challenge Track - ECIR 2017*, Aberdeen, Scotland, 2017, pp. 32–41. [Online]. Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/smerp2017.pdf (cit. on p. 49).

[200] A. Strauss and J. M. Corbin, *Grounded theory in practice*. Sage Publications, Inc., 1997 (cit. on p. 68).

[201] R. Tatman, *Stopword Lists & Frequency Information for 9 African Languages*, 2017. [Online]. Available: https://www.kaggle.com/rtatman/stopword-lists-for-african-languages (visited on 05/21/2021) (cit. on pp. 47, 48).

[202] J. P. Telemala and H. Suleman, "Exploring information needs and search behaviour of swahili speakers in tanzania," in *Maturity and Innovation in Digital Libraries*, M. Dobreva, A. Hinze, and M. Žumer, Eds., ser. ICADL 2018. Lecture Notes in Computer Science, vol. 11279, Hamilton, New Zealand: Springer International Publishing, Cham, 2018, pp. 185–190, ISBN: 978-3-030-04257-8. DOI: 10.1007/978-3-030-04257-8_18 (cit. on p. 32).

[203] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., ser. LREC'12, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218, ISBN: 978-2-9517408-7-7. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/index.html (cit. on p. 56).

[204] ——, "Opus–parallel corpora for everyone," in *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia, 2016. [Online]. Available: https://aclanthology.org/2016.eamt-2.8 (cit. on pp. 56, 57).

[205] F. Topan, "Tanzania: The development of swahili as a national and official," in *Language and National Identity in Africa*, A. Simpson, Ed., Oxford University Press, 2008, pp. 252–266 (cit. on pp. 5, 44, 65, 75, 79, 80, 100).

[206] M.-F. Tsai, H.-H. Chen, and Y.-T. Wang, "Learning a merge model for multilingual information retrieval," *Information Processing & Management*, vol. 47, no. 5, pp. 635–646, 2011. DOI: 10.1016/j.ipm.2009.12.002 (cit. on pp. 5, 37, 39, 40, 106, 140).

[207] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, "Frank: A ranking method with fidelity loss," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands*, ser. SIGIR '07, New York, NY, USA: Association for Computing Machinery, 2007, 383–390. DOI: 10.1145/1277741.1277808 (cit. on pp. 37, 40).

[208] M.-F. Tsai, Y.-T. Wang, and H.-H. Chen, "A study of learning a merge model for multilingual information retrieval," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08, Singapore, Singapore: Association for Computing Machinery, 2008, 195–202. DOI: 10.1145/1390334.1390370 (cit. on pp. 21, 37, 39, 40, 140).

[209] N. Usunier, M.-R. Amini, and C. Goutte, "Multiview semi-supervised learning for ranking multilingual documents," in *Machine Learning and Knowledge Discovery in Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds., ser. ECML PKDD 2011. Lecture Notes in Computer Science, vol. 6913, Berlin, Heidelberg: Springer, 2011, pp. 443–458, ISBN: 978-3-642-23808-6. DOI: 10.1007/978-3-642-23808-6_29 (cit. on pp. 5, 37, 39, 40, 140).

[210] E. Vassilakaki, E. Garoufallou, F. Johnson, and R. J. Hartley, "An exploration of users' needs for multilingual information retrieval and access," in *Metadata and Semantics Research*, ser. MTSR 2015. Communications in Computer and Information Science, E. Garoufallou, R. J. Hartley, and P. Gaitanou, Eds., vol. 544, Cham: Springer International Publishing, 2015, pp. 249–258, ISBN: 978-3-319-24129-6. DOI: 10.1007/978-3-319-24129-6_22 (cit. on pp. 31–33, 102).

[211] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., ser. NeurIPS 2017, vol. 30, Long Beach, CA, USA: Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 61).

[212] M. Vigo, N. Matentzoglu, C. Jay, and R. Stevens, "Comparing ontology authoring workflows with protégé: In the laboratory, in the tutorial and in the 'wild'," *Journal of Web Semantics*, vol. 57, p. 100 473, 2019, ISSN: 1570-8268. DOI: 10.1016/j.websem.2018.09.004 (cit. on p. 33).

[213] E. M. Voorhees, "Natural language processing and information retrieval," in *Information Extraction*, M. T. Pazienza, Ed., ser. SCIE 1999. Lecture Notes in Computer Science, vol. 1714, Springer Berlin Heidelberg, 1999, pp. 32–48. DOI: 10.1007/3-540-48089-7_3 (cit. on pp. 43, 46).

[214] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '95, Seattle, Washington, USA: Association for Computing Machinery, 1995, 172–179, ISBN: 0897917146. DOI: 10.1145/215206.215357 (cit. on pp. 5, 34, 39, 106).

[215] E. M. Voorhees *et al.*, "The trec-8 question answering track report," in *Trec*, Citeseer, vol. 99, 1999, pp. 77–82. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151495 (visited on 07/07/2021) (cit. on p. 24).

[216] J. Wang and A. Komlodi, "Switching languages in online searching: A qualitative study of web users' code-switching search behaviors," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, ser. CHIIR '18, New Brunswick, NJ, USA: Association for Computing Machinery, 2018, 201–210. DOI: 10.1145/3176349.3176396 (cit. on pp. 4, 6, 31, 32, 102).

[217] J. Wang, A. Komlodi, and O. Ka, "Understanding multilingual web users' code-switching behaviors in online searching," *Proceedings of the Association for Information Science and Technology*, vol. 55, no. 1, pp. 534–543, 2018. DOI: 10.1002/pra2.2018.14505501058 (cit. on pp. 30–33, 83).

[218] M. H. Weik, "Machine translation," in *Computer Science and Communications Dictionary*. Boston, MA: Springer US, 2001, pp. 952–952, ISBN: 978-1-4020-0613-5. DOI: 10.1007/1-4020-0613-6_10823. [Online]. Available: https://doi.org/10.1007/1-4020-0613-6_10823 (cit. on p. 46).

[219] W. Wentland, C. Silberer, and M. Hartung, "Building a multilingual lexical resource for named entity disambiguation, translation and transliteration," in *Proceedings of the Sixth International Language Resources and Evaluation*, C. N., C. K., M. B., M. J., O. J., P. S., and T. D., Eds., ser. LREC'08, Marrakech, Morocco: ELRA, Paris, 2008, pp. 3230–3237. [Online]. Available: http://hdl.handle.net/10230/41874 (cit. on pp. 48–50).

[220] Wikipedia, *List of Wikipedias*, 2020. [Online]. Available: https://meta.wikimedia.org/wiki/List_of_Wikipedias (visited on 05/26/2021) (cit. on pp. 17, 20, 55, 56, 99).

[221] ——, *Swahili language*, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Swahili_language (visited on 05/21/2021) (cit. on pp. 5, 42, 45).

[222] T. D. Wilson, "On user studies and information needs," *Journal of documentation*, vol. 37, no. 1, pp. 3–15, 1981, ISSN: 0022-0418. DOI: https://doi.org/10.1108/eb026702 (cit. on p. 13).

[223] B. Wójtowicz, "Survey of swahili dictionaries: The macrostructure," *Studies in African Languages and Cultures*, no. 50, pp. 5–39, 2016. [Online]. Available: https://www.ceeol.com/search/article-detail?id=843601 (cit. on p. 43).

[224] WorldBrain, *A simple repository to remove 'irrelevant for search' words, support for 51 languages*, 2017. [Online]. Available: https://github.com/WorldBrain/remove-stopwords (visited on 05/21/2021) (cit. on p. 47).

[225] Y. Yamamoto and T. Yamamoto, "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*, ser. JCDL '20, Virtual Event, China: Association for Computing Machinery, 2020, 37–46. DOI: 10.1145/3383583.3398519 (cit. on p. 84).

[226] M. Yarmohammadi, X. Ma, S. Hisamoto, M. Rahman, Y. Wang, H. Xu, D. Povey, P. Koehn, and K. Duh, "Robust document representations for cross-lingual information retrieval in low-resource settings," in *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Dublin, Ireland: European Association for Machine Translation, 2019, pp. 12–20. [Online]. Available: https://aclanthology.org/W19-6602 (visited on 02/23/2021) (cit. on pp. 4, 43, 54, 61).

[227] P. Younger, "Internet-based information seeking-behaviour among doctors and nurses: A short review of the literature," *Health Information and Libraries Journal*, vol. 27, no. 1, pp. 2–10, 2010. DOI: 10.1111/j.1471-1842.2010.00883.x (cit. on pp. 28, 30).

[228] R. Zbib, L. Zhao, D. Karakos, *et al.*, "Neural-network lexical translation for crosslingual ir from text and speech," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19, Paris, France: Association for Computing Machinery, 2019, 645–654, ISBN: 9781450361729. DOI: 10.1145/3331184.3331222 (cit. on pp. 4, 56, 61).

[229] E. Zhang and Y. Zhang, "Precision," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 2126–2126. DOI: 10.1007/978-0-387-39940-9_480 (cit. on pp. 22, 23).

[230] R. Zhang, C. Westerfield, S. Shim, G. Bingham, N. Verma, A. Fabbri, W. Hu, and D. Radev, "Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 3173–3179. DOI: 10.18653/v1/P19-1306 (cit. on pp. 4, 56, 61).

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of January 26, 2022 (`classicthesis`).