Assessing the Impact of Vocabulary Similarity on Multilingual Information Retrieval for Bantu Languages

ABSTRACT

Despite the availability of massive open information and efforts to promote multilingualism on the Web, content in some languages such as Bantu languages remains negligible. Information Retrieval (IR) systems, such as the Google search engine, use algorithms that work well with languages with the most content on the Web. The aim of this study is to investigate indexing strategies for Multilingual Information Retrieval(MLIR) and its effect on the quality of retrieval for related languages. Similarities across related languages such as vocabulary and structural overlap, can potentially be exploited to provide more opportunities for information access for under-resourced and under-documented languages. A multilingual collection of documents in two African languages, in the family of Bantu languages, and English is used. The results in the study shows that when comparing related and unrelated language pairs, MLIR indexing strategies result in comparable or worse quality of retrieved results.

CCS Concepts

•**Information systems** → **Multilingual and cross-lingual retrieval;** *Document collection models; Test collections;*

Keywords

Multilingual Information Retrieval; Test Collection; Information Retrieval Evaluation

1. INTRODUCTION

The current digital revolution has changed how people seek and use information. The unprecedented large volumes of information available on the Web is accessible to everyone at almost no cost. Accordingly, the Web has emerged to become the primary source of information of our time. The success of the Web has been driven by many factors. The Web is a decentralised form of media where anyone can publish and consume information. For example, anyone can become a Web publisher through User Generated Content (UCG) by using different services and systems that are available for free. Although UCG services such as social media have pushed the diversity of the content on the Web, and in spite of the fact that the Web is becoming more multilingual, very little content has been published in many of the languages of the world. English and other widely spoken languages continue to dominate. Consequently, the majority of the content currently available on the Web does not represent the cultural and language diversity of the world. The major challenge is to make such small amounts of content available to users who are interested to read content in these languages.

While search engines such as Google allow users to easily find relevant content on the Web in any language, languages with limited content are disadvantaged due to ranking algorithms that based on statistical methods which result in higher rankings for content that is in dominant languages. This is also true for documents with mixed languages [13]. For example, if a user is interested in '*nyumba ya galasi*' (glass house in Chichewa), Google search engine return results in Kiswahili, Chichewa and English, on galaxy and gala (see Fig. 1). This may be due to term similarity and lack of language identification for the query. Although the query is in Chichewa, the returned document set consists of results mainly in Kiswahili and English. The Kiswahili results are based on the keyword '*nyumba*' which is a shared term with Chichewa, Citumbuka and other related languages. Interestingly, relevant results in the language of the query were found much lower in the list.

This paper assesses the impact of vocabulary similarity in the context of Multilingual Information Retrieval (MLIR) for Bantu Languages. Vocabulary similarity in this context refers to a situation where multiple languages have the same words in their lexicons. The paper investigates the impact of vocabulary similarity on retrieval performance for test queries in two Bantu languages namely, Citumbuka and Chichewa.

Bantu languages is a family of languages spoken by over 200 million people across the Sub-Saharan Africa [18]. Bantu languages are highly inflectional and agglutinative and these features poses challenges in Bantu language IR [5]. There is no widely agreed upon genealogical classification of Bantu languages. However, the Guthrie classification is the most widely used classification of Bantu languages [10]. Guthrie classified Bantu languages into several classes or groups and labelled these languages using an alphanumeric system. For example, Chichewa and Citumbuka fall in zone N and were

Kolkata '2016 Kolkata, India

DOI: 10.1145/1235

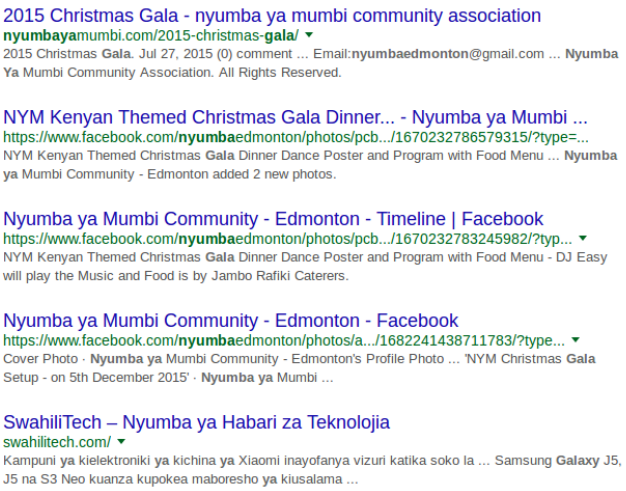

Figure 1: The top 5 results for the query 'nyumba ya galasi' on August 10, 2016 on Google Search engine

labelled N30 and N21 respectively. Chichewa is spoken by over 10 million people in Malawi, Zambia and Mozambique. It is widely spoken and understood by most people in Malawi. Citumbuka is a Bantu language spoken by over 2 million people in northern Malawi and eastern Zambia. The Chichewa and Citumbuka languages are not inherently intelligible, i.e., one cannot understand either of the language without learning the languages. In terms of vocabulary similarity, Citumbuka and Chichewa share words and cognates. In the linguistics community, there are many methods for measuring vocabulary similarity. One commonly used method is lexicostatistics, i.e., a quantitative measure of language relatedness in which words sharing a common form and meaning are counted. The percentage of common words is used as a vocabulary or lexical similarity measure. A simple technique that is used is to use a standard list such as a Swadesh list that has 207 terms which are deemed to be universal and culturally independent concepts. Kiso [10] composed a Swadesh list for Citumbuka and Chichewa and found out that 113 out of 203 (56 %) words were similar or cognates (four of the terms were untranslatable).

This paper assesses the effect of language similarity on retrieval performance in a multilingual environment. The process of building the test collection is described together with the characteristics of the collection. Experiments to study the impact of vocabulary similarity on performance were done. Queries were run on collections with documents in related languages and unrelated languages. Specifically, the paper makes the following contributions:

1. Insights into the impact of vocabulary similarity on MLIR
2. A multilingual test collection for two resource scarce languages.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 describes the process of creating a test collection used in the experiments. Section 4 describes the experimental design and Section 5 discusses the results. Final conclusions are drawn in section 6.

2. RELATED WORK

IR in Bantu Languages. A few research studies exist in the area of IR for Bantu Languages. Malumba et al. [12] constructed a custom Web search engine for *Isizulu*, a Bantu language spoken in South Africa. Morphological processing is compared with the traditional affix-based stemming, and better performance is reported for the later. Additionally, statistical modelling is used for automatic language identification to identify documents written in IsiZulu for the crawler. Cosjin et al. [5] investigated Cross Language Information Retrieval (CLIR) between Isizulu and English. Approximate string matching techniques such as n-gram matching were investigated as a method for matching words in queries with index entries. In addition, challenges of IR for Isizulu were identified namely, lack of electronic resources such as machine readable dictionaries, lack of terminology, translation issues associated with paraphrasing and borrowed words, and agglutination.

No translation. Vocabulary similarity has been used in CLIR environments with no query translation but string similarity matching methods were used to match query terms and index terms. For instance, English-French cognates were used together with spelling rules to perform CLIR between English and French [2]. Gey used this no translation method in retrieval from Chinese queries to Japanese and vice versa with the assumption that the Japanese Kanji alphabet was derived from Chinese language [6]. Both studies reported low performance.

Language similarity. Work on the effects of language relatedness on IR in the context of the Latent Semantic Indexing (LSI) model are investigated in [3]. The study investigates script similarity and genetic relatedness focusing on Indo-European and Semitic languages. Training data for the LSI model was manipulated to include text from related languages and unrelated languages. The study concluded that MIR improves as the number of languages for parallel text in training increases and that text from related languages significantly boosts retrieval.

3. BUILDING A TEST BED

A test collection for an IR system evaluation consists of a series of documents, search topics and judgements of the topics [4]. Such a collection does not exist for Chichewa and Citumbuka, the languages which have been used in the study. Therefore, work was carried out to build a test collection consisting of documents in Chichewa, Citumbuka and English. A set of 50 topics [11] was formulated based on the documents in the collection and relevance judgements were done through crowd-sourcing. Figure 2 depicts the process of developing the test collection.

3.1 Corpus Collection and Preparation

Chichewa and Citumbuka are poor-resourced languages, and documents written in these languages are scarce on the Web, e.g., Wikipedia has 328 articles written in Chichewa[1] and 559 articles in Citumbuka[2]. Documents written in Chichewa and Tumbuka available on the Web are mainly religious writings, health education and agriculture. The documents that contributed to the test collection were collected manually from the Web by one of the investigators. Firstly, litera-

[1]https://ny.wikipedia.org/wiki/Main_Page
[2]https://tum.wikipedia.org/wiki/Main_Page

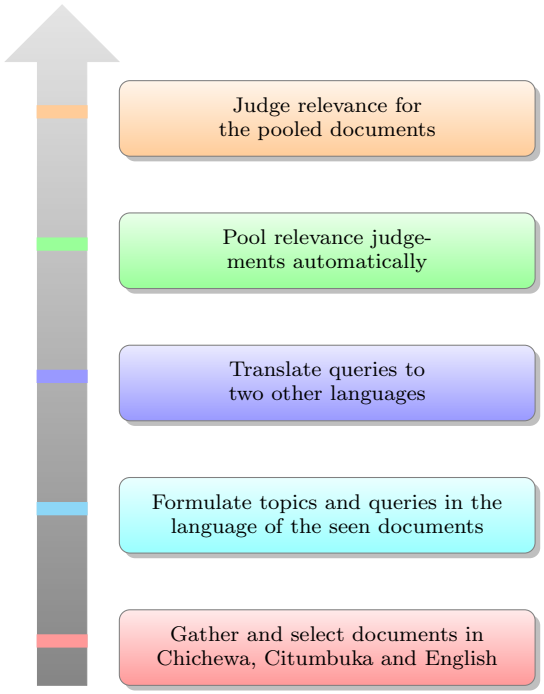

Figure 2: Steps used to build the test collection

ture was investigated to find sources that have been used for these languages. A list of sites that have published documents in these languages is documented in the Crubadan Project [15]. Documents were gathered from these sites and other sites that were discovered through Google search. Examples of sites from which documents were sourced include Wikipedia, Hesperian Health Guides[3], K4Health [4], Indigenous tweets and blogs and religious websites.

Domains with documents in all the three languages were selected namely, documents in the domains of agriculture, health, religion and culture. English is the most prominent language on the Web and only relatively similar documents written in the other two languages were used. Also, in some cases, equivalent documents were found and such documents were incorporated in the document collection, e.g., translated religious writings such as the Bible. Collected documents were analysed for their quality to ensure that the documents are monolingual and do not contain other languages. Documents within the identified domains were selected and documents in PDF format were divided into individual pages to control the sizes of the documents. This generated a collection of monolingual documents in three languages on similar topics. Table 1 provides the summary of the statistics of the collection.

	Chichewa	Citumbuka	English	Total
Number of documents	19,435	11,234	13,785	44,454
Size of collection	3.3GB	2.96GB	1.9GB	8.16GB
Number of words	5,088,633	3,916,654	6,477,446	15,482,733
Number of distinct words	183,196	153,411	64,612	401,219

Table 1: Document Collection Summary Statistics

[3]http://hesperian.org/
[4]https://www.k4health.org/

3.1.1 Evaluating the Collection

To understand the frequency distribution of the words in the collected corpora, [20]. Zipf's law was used to identify any imbalances in the dataset and to test if they conform to the distribution of human language utterances. Distinct words in each of the three languages were ranked according to their frequencies. Their normalised rank was plotted against their frequency. Figure.3.1.1 depicts the Zipf plots for the corpora in Chichewa, Tumbuka and English. The plots show that the frequency distributions fit into the Zipf curve.

3.2 Recruiting Assessors

Multilingual participants who can read, write and speak Chichewa, Citumbuka and English were recruited to generate topics, queries, translations of the queries and to provide relevance judgements. Messages were sent by e-mail to students at universities in Malawi asking them to participate in the task. Participants' language proficiency in the three languages was not formally tested but a pre-task questionnaire was used for participants in the topic generation task. English is used as a language of instruction in Malawi, whereas Chichewa is a lingua franca in Malawi. Citumbuka is the predominantly spoken language in the Northern Region and participants had to declare that they speak the language, including completing a pre-task questionnaire. A pre-task questionnaire was administered to obtain demographic data, participant language skills and attitudes in searching for information on the Web as well as their experience and attitudes towards searching using the investigated languages. All participants had either Citumbuka or Chichewa as their mother tongue, and one of the other as a second language. In addition, due to issues with political language planning and urbanisation, many people who speak Citumbuka can also speak Chichewa.

3.3 Creating Topics and Queries

15 participants assessed 50 documents in total on an online system. Each participant was given a set of five documents in three languages to identify the topic of the document and to create queries that might result into the given documents. A topic with three queries was created for each of the documents. Topics were written in English. In total, 50 topics were generated and each topic contained three queries in three languages (150 queries in total). The submitted information was transformed to Text REtrieval Conference (TREC)-like topics, i.e., adding topic number and other formatting, such as putting the topics in separate topic file.

3.4 Creating Relevance Judgements

Creating relevance judgements is one of the laborious tasks in building a test collection [16]. Pooling was employed to reduce the number of documents to be judged per topic and query [9]. 150 queries were run on three separate collections, each in a different language. Solr 5 was used as an IR platform, selecting the top 35 documents for each query. 25 assessors were recruited, each one to judge 35 documents on each query. The assessors were recruited using the same procedure as for creating the queries. Some of the participants who created the queries also performed the topic relevance judgements. Recruited assessors first completed an online tutorial about grading relevance judgements. The tutorial included examples of judgements as well as infor-

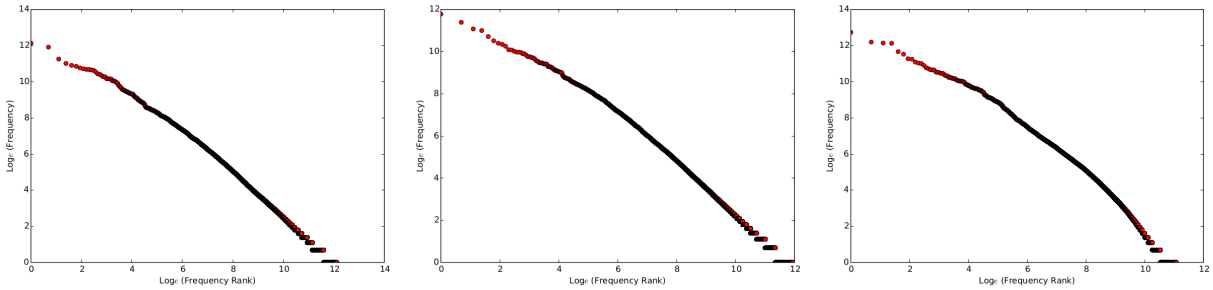

Figure 3: Zipfian plots for the Chichewa, Citumbuka and English Corpora respectively

Description	Value
Average number of words per query	4.35
Number of equivalent queries	6
Number of queries with one term difference	24
Average number of similar words per query	1.43
Number of non-English queries with terms in English	4
Number of queries sharing at least one term	68
Number of queries per language	50
Total number of queries	150

Table 2: Summary statistics of queries

mative text about judgements. Judgements were done on a four point scale as follows [17] : 0 for irrelevant documents, 1 for marginally relevant documents, 2 for fairly relevant documents and 3 for highly relevant documents.

4. EXPERIMENTS AND RESULTS

The focus of the study was to investigate the extent to which vocabulary similarity affects quality of retrieval results. A test collection made up of documents in Chichewa, Citumbuka and English were used in the study. Solr 5, an open source IR platform, was used in the experiments. No linguistic analysis was performed on the text apart from tokenisation and lowercasing. Retrieval was restricted to content only and no metadata were used in the querying process. Solr uses a Boolean model to match documents and the query. Vector space model is used to come up with similarity measure scores that are used for ranking [7]. The experiments used Solr's default scoring algorithm, which uses TF-IDF as a principal component [14].

4.1 Experimental Design

Four multilingual indexing strategies were investigated. For each strategy, three different language combinations were used namely, Chichewa and Citumbuka, Chichewa and English and Citumbuka and English.

4.1.1 Multilingual Indexing Strategies

The design of the experiments was based on four indexing MLIR strategies [7]. High level diagrams of these strategies are given in Fig. 4.

Single multilingual index. A single multilingual index set-up in which text content from documents regardless of language are put in a single content field. All queries were routed to this single field.

Single multilingual index with a content field per language. In this approach, a single multilingual index has a language specific field. Content in a particular language is mapped to a specific field. Similarly, queries are run on language specific fields.

Single multilingual index with language tagged terms. A single multilingual index in which text content from documents regardless of language are mapped to a single content field. However, each term is tagged with it's own language. Similarly, query terms are tagged with language codes. ISO 639-1 codes, short forms for language names were used as tags, i.e., en, ny and tum.

Index per language. In this set-up, each language has its own index. Queries are routed to language specific index.

4.1.2 Language Relatedness based Document sets

Three different document collection set-ups were used for the experiments, namely, a collection containing documents of genetically related languages that share some vocabulary, i.e., Chichewa and Citumbuka, and collections containing two genetically unrelated languages, i.e., Chichewa and English; and Citumbuka and English. Queries in all three languages were run on each type of collection and indexing set-up.

4.2 Experiments

All 150 queries,i.e., 50 queries in each language, were run for each of the four indexing strategies. Each indexing set-up consisted of a combination of two languages. Queries in all three languages were then routed to the different indexes or fields that contained content in the languages being investigated. Queries were also run on monolingual collections, i.e., within language retrieval, and were used as baseline performance for the experiments.

4.3 Performance Evaluation

Mean Average Precision (MAP) was used to evaluate the overall performance of the experiments as it provides a single metric that is used to compare different set-ups. Top 100 results (for queries which had an answer set of over 100 documents) were used in the evaluation. MAP scores for all the experiments are shown in Table 3. Table 4 provides MAP scores for monolingual collection experiments.

It is also important to investigate where relevant documents are retrieved, i.e., whether relevant documents appear at the top or at the end of the ranking. The test collection adopted graded relevance assessments and allowed graded relevance evaluation to be used. Normalised Discounted Cumulated Gain (NDCG) [8] for top ten results was calculated

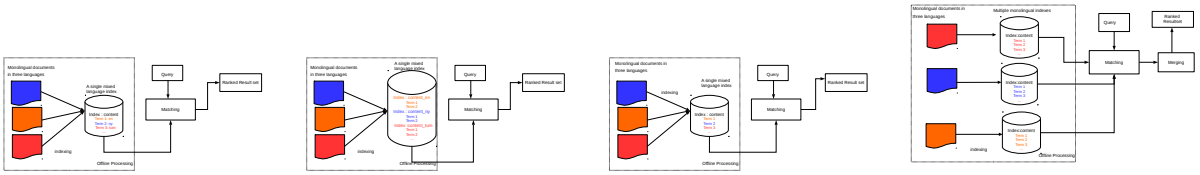

Figure 4: High level diagrams MLIR indexing strategies

Doc Language	ny–tum			ny–en			tum-en		
Query Language	**ny**	**tum**	**en**	**ny**	**tum**	**en**	**ny**	**tum**	**en**
Centralised	0.0825	0.0494	0.0096	0.0829	0.0261	0.0957	0.0276	0.0569	0.0902
Language Tagged	0.0712	0.0281	0	0.0709	0	0.0775	0	0.0279	0.0776
Field per language	0.0758	0.0426	0.0188	0.0590	0.0213	0.0862	0.0240	0.0418	0.1105
Multiple index	0.0801	0.0437	0.0086	0.0718	0.0242	0.0241	0.0256	0.0460	0.0998

Table 3: MAP for different indexing strategies and document set languages

for all the experiments and are given in Table . 5. Table . 5 provides NDCG for the monolingual experiments.

The language tagged terms index had lower MAP and NDCG values for different language combination collections and query languages. For example, querying with a language different from the documents in collection returned nothing for all queries. A centralised language naive index gave a MAP value closer to the within-language baseline.

Averaging precision over many queries may not provide important insights on the phenomena under study [1]. Investigations were also done to observe the behaviour of individual queries with different characteristics. Table. 7 provides summary statistics for ten queries which were used in the investigation, i.e., five Chichewa and five Citumbuka queries which shared the same characteristics). For example, number of words in the query for the same topic may differ because of translation. Figure. 4.3 provides ranking of relevant documents for the ten queries for within-language retrieval (ny1,tm1) and multilingual collection (ny2 and tm2) of documents in Citumbuka and Nyanja. A single index and single field set-up was used because it had relatively consistent metrics for MAP and NDCG@10.

4.4 Discussion

Indexing strategies that separate documents in terms of language gave relatively lower quality results. The related languages collection with documents in Chichewa and Citumbuka results are in the first three columns(with results) of Table 3. The performance for the Chichewa queries are slightly lower but similar to monolingual Chichewa collection (see Table 4). The performance for Citumbuka queries is similarly lower than for a monolingual Citumbuka collection. English has a non-zero result because some Chichewa/Citumbuka documents are still relevant to the translation of the English query even though there are no English documents.

The second 3 columns are the non-related languages of Chichewa and English. Performance for Chichewa queries is slightly lower but similar to that for a monolingual Chichewa collection. Performance for English queries is similarly lower than for a monolingual English collection. Citumbuka has a non-zero result because some Chichewa/English documents are still relevant to the translation of the Citumbuka query even though there are no Citumbuka documents. Given that these languages are different, this performance was expected. The last 3 columns show similar results for the Citumbuka-English collection. This pattern is also observed in the NDCG metrics in 5.

Ranks for queries described in 7 are given in 4.3 for q1, q2, q3, q4 and q5 respectively. q4 is one query and provides relatively the same results for all the queries. Queries with more than one word difference had no relevant results in the top 100 documents returned for Citumbuka queries, i.e., all the relevant documents are in Chichewa and these results may have been ranked towards the end of the results. q2 and q3 Citumbuka queries had better results after adding Chichewa documents (tm). This shows that languages with less relevant results can leverage similarities within languages to get better quality results, i.e., the more related the query the better the results.

The experiment has demonstrated that when comparing similar and dissimilar language pairs, all combinations result in comparable or worse performance. The language pairs that are similar do not result in better performance, so the standard indexing techniques, without some explicit algorithmic changes, are not able to exploit language similarity to improve results. For example, completely labelling index terms with its language tag, allows only within language retrieval. If a search engine exploited language similarity, then the results for Citumbuka and Chichewa would have more relevant documents for a multilingual collection, but this is not the case.

5. CONCLUSION

The paper investigated the impact of vocabulary similarity of two related under-resourced Bantu languages, Chichewa and Citumbuka, on retrieval quality in a MLIR environment. To this end, an IR test collection was built and experiments exploring different indexing strategies for MLIR were conducted. A language agnostic indexing gave results similar to within-language retrieval. Language tagged term indexing had the poorest results when querying using languages not in the collection. However, the more similar the queries in an indexing environment where documents are not separated based on language, the better the results. This is an opportunity for languages with less content as cross-lingual vocabulary links can be leveraged to find content in other related languages. Future work will focus on techniques that can exploit language relatedness to improve MLIR quality

Doc Language	ny			tum			en		
Query Language	**ny**	**tum**	**en**	**ny**	**tum**	**en**	**ny**	**tum**	**en**
Monolingual	0.0839	0.0264	0.0084	0.0320	0.0566	0.0182	0.0009	0.0025	0.1038

Table 4: MAP for monolingual collection experiments

Doc Language	ny–tum			ny–en			tum-en		
Query Language	**ny**	**tum**	**en**	**ny**	**tum**	**en**	**ny**	**tum**	**en**
Centralised	0.4749	0.2424	0.0356	0.3945	0.1121	0.4756	0.0479	0.1709	0.503
Language Tagged	0.4358	0.1262	0	0.3606	0	0.4008	0	0.1155	0.45
Field per language	0.4608	0.1778	0.0289	0.34530	0.092	0.377	0.02667	0.09705	0.45
Multiple index	0.4131	0.1718	0.0284	0.2456	0.0785	0.3172	0.01765	0.0905	0.4399

Table 5: NDCG for different indexing strategies and document set languages

Doc Language	ny			tum			en		
Query Language	**ny**	**tum**	**en**	**ny**	**tum**	**en**	**ny**	**tum**	**en**
Monolingual	0.4849	0.124	0.028	0.0588	0.1562	0.0146	0.010	0.01	0.53

Table 6: NDCG for monolingual collection experiments

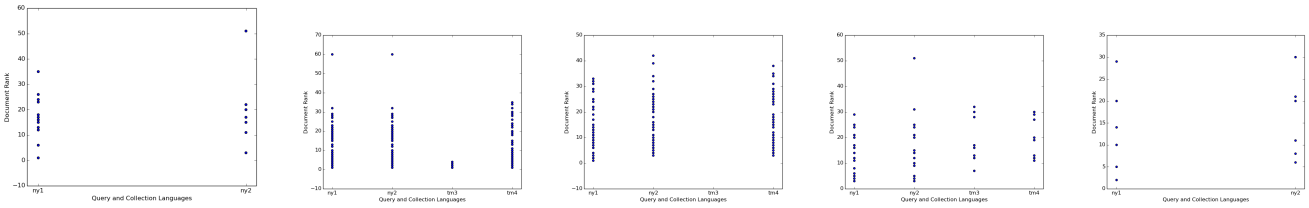

Figure 5: Ranks of documents for 5 queries

Id	similar terms	different terms	Total	Relevant items
q1	4	2	6	24
q2	1	1	2	22
q3	2	1	3	27
q4	1	0	1	33
q5	1	3	4	6

Table 7: Characteristics of selected queries

of results for Bantu Languages.

6. REFERENCES

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[2] C. Buckley, M. Mitra, J. Walz, and C. Cardie. Using clustering and superconcepts within smart: Trec 6. *Inf. Process. Manage.*, 36(1):109–131, Jan. 2000.

[3] P. A. Chew and A. Abdelali. The effects of language relatedness on multilingual information retrieval: A case study with indo-european and semitic languages. In *IJCNLP*, pages 1–9, 2008.

[4] P. Clough and M. Sanderson. Evaluating the performance of information retrieval systems using test collections. *information research journal*, 18(2), June 2013.

[5] E. Cosijn, A. Pirkola, T. Bothma, and K. Järvelin. Information access in indigenous languages: a case study in zulu. In *In: Proceedings of the 4th International Conference on Conceptions of Library and Information Science. COLIS*, pages 221–238, 2002.

[6] F. Gey. Search between chinese and japanese text collections. In *Proceedings of NTCIR-6 Workshop Meeting*, UC Data Archive and Technical Assistance University of California, Berkeley, May 2007.

[7] T. Grainger and T. Potter. *Solr in Action*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2014.

[8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[9] S. Jones and C. Van Rijshergen. Report on the need for and provision of an ideal information retrieval test collection, british library research and development report 5266, 1975.

[10] A. Kiso. Tense and aspect in chichewa, citumbuka and cisena: a description and comparison of the tense-aspect systems in three southeastern bantu languages. 2012.

[11] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a large-scale test collection: An analysis of the search results from the first ntcir workshop. *Inf. Retr.*, 5(1):41–59, 2002.

[12] N. Malumba, K. Moukangwe, and H. Suleman. *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings*, chapter AfriWeb: A Web Search Engine for a Marginalized Language, pages 180–189.

Springer International Publishing, Cham, 2015.

[13] M. Mustafa and H. Suleman. Multilingual querying. In *Proceedings of the Arabic Language Technology International Conference (ALTIC), Alexandria, Egypt*, 2011.

[14] C. Peters, M. Braschler, and P. D. Clough. *Multilingual Information Retrieval - From Research To Practice*. Springer, 2012.

[15] K. P. Scannell. The crúbadán project: Corpus building for under-resourced languages. 2007.

[16] P. Sheridan, J. P. Ballerini, and P. Schäuble. *Building a Large Multilingual Test Collection from Comparable News Documents*, pages 137–150. Springer US, Boston, MA, 1998.

[17] E. Sormunen. Liberal relevance criteria of trec -: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 324–330, New York, NY, USA, 2002. ACM.

[18] M. Van de Velde, D. Nurse, K. Bostoen, and G. Philippson. *The Bantu Languages*. Routledge Language Family Series. Taylor & Francis, 2006.

[19] H. young Rieh and S. Y. Rieh. Web searching across languages: Preference and behavior of bilingual academic users in korea. *Library and Information Science Research*, 27(2):249 – 263, 2005.

[20] G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.