Intercomprehension in Retrieval: User Perspectives

ABSTRACT

The majority of web content is published in languages not accessible to many potential users who may only be able to read and understand their local languages. Prior research has focused on using translation to provide users with information written in other languages, yet there are still many languages with little or no such resources. In this paper, we propose the use of intercomprehension – a form of communication in which speakers of two different languages communicate using their own languages, mainly due to similarities between the languages. Accordingly, we conducted a user study to explore user interaction behaviour in a retrieval environment where intercomprehension is expected; to investigate the usefulness of search results, which assumes intelligibility and relevance; and investigate affective episodes associated with intercomprehension in retrieval through retrospection. Although intercomprehension may come with a cost to understand unfamiliar languages, user preference of ranking of results in related languages incorporates intelligibility, which assumes intercomprehension. Our findings also suggest that intercomprehension is useful in retrieval for related languages – users are able to identify relevant documents as well as complete search tasks by applying intercomprehension. However, the negative emotions or frustration associated with intercomprehension suggest that this type of interaction should be used in extreme cases where there are no relevant or few documents available associated with the query.

ACM Reference format:

. 2019. **Intercomprehension in Retrieval: User Perspectives**. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 11 pages.

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital content such as information on the web provide opportunities for information access. Unfortunately, the majority of available content is published in languages not accessible to many potential users who may only be able to read and understand their local languages. Previous Information Retrieval(IR) approaches, particularly Cross-Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR), have focused on translation of either queries or documents in the retrieval pipeline to accommodate such needs [33]. However, many of the languages of this nature are Scarce Resources Languages (SRLs) and lack the necessary resources such as machine translation tools to make the available content accessible[39]. In this setting, the main challenge is how to make the available content to users with resource constraints. We propose Intercomprehension – a form of communication in which speakers of two different languages communicate using their own languages, mainly due to similarities between the languages[32] – to be used in IR with the assumption that users submitting queries in a particular language can understand content in other closely related languages or dialects.

Users applying intercomprehension in a retrieval scenario are expected to perform reading intercomprehension [38], i.e., understanding text written in a language different from the language of the reader or query, to find the relevant content from the retrieved documents. Certainly, users may face difficulty to understand content in another language [16] as understandability may depend on how intelligible the user language is with respect to the language of the retrieved document. Accordingly, it is necessary to investigate whether presenting users with results that require intercomprehension is useful to the user, and more specifically, how results with varying levels of relevance and intelligibility should be presented to the user. In addition, reading intercomprehension may require more effort from the user and may lead to frustration and anxiety, and therefore we study the type of emotions associated with different levels of comprehension in our context.

In this paper, we present the first user study on intercomprehension in retrieval. We explore the user interaction behaviour in retrieval environment in which intercomprehension is expected. Specifically, we study the usefulness of search results, which assumes intelligibility and relevance, i.e., topicality dimension. Our user study also studies the ranking preference of users in this context. We further investigate affective episodes associated with intercomprehension in retrieval through retrospection. The study answers the following questions:

- RQ1 How should search results retrieved based on topicality and intelligibility be ranked? Does intelligibility matter in the rank preference of such results?
- RQ2 Are search results in closely related languages useful to the user?
- RQ3 What type of emotions do users experience when interacting with search results that require intercomprehension?

Our user study focused on six Bantu languages with varying levels of intelligibility namely Cisena, Chichewa, Citonga, Cinyanja, Citumbuka and Luganda. We hypothesised that users would prefer documents that are relevant and comprehensible to be ranked highly. Our results show that user preference of ranking search results that assumes intercomprehension incorporate intelligibility for relevant documents only, and this observation matched our proposed theory. Interestingly, users find relevant search results written in related languages useful and are able to complete a search task using a document written in a language that they have not seen before or casually familiar with. Ironically, we observed that users experience mixed emotions in response to intercomprehension stimulus – some of the participants were surprised that they

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

https://doi.org/10.1145/nnnnnnn.nnnnnnn

would be able to understand the contents of documents written in an unfamiliar language and had a positive search experience while others were frustrated and experienced negative emotions, which affected their search experience.

The rest of the paper is organised as follows. Section 2 describes related work on topics including retrieval for related languages, emotions in retrieval and document understandability in retrieval. Section 3 introduces the concept of intelligibility in the manner explored in the study. Section 4 presents the procedure and assumptions made for the study. Section 5 outlines the results obtained through the study. Section 6 discusses our findings and their implications and we conclude in section 7.

2 RELATED WORK

Research work presented in this paper is related to studies in the following areas: (i) retrieval for similar languages, (ii) comprehension and readability in retrieval, and (iii) emotions in retrieval. We proceed to discuss each area.

2.1 Retrieval for Related Languages

Intercomprehension has not been studied in the context of retrieval for similar languages. However, retrieval techniques that could use intercomprehension have been studied with regard to retrieval performance in system oriented studies. To improve retrieval effectiveness, language similarity in terms of vocabulary similarity has been used to retrieve documents in related languages [3, 6, 12]. Vocabulary similarity has been used in CLIR – untranslated queries together with string similarity matching methods are used to match index terms of a closely related language. For instance, Buckley et al.[3] used English-French cognates with spelling rules to perform CLIR between English and French. CLIR with no query translation involving related languages has also been used on non-alphabetic languages; Gey [12] used Chinese queries on Japanese text and vice versa with the assumption that the Japanese Kanji alphabet was derived from Chinese language. Both studies reported lower performance than retrieval involving query translation. Although these studies are system oriented, the absence of translation assumes that users can understand the retrieved documents in the related language either through intercomprehension or bilingualism. The other possible scenario is that understandability of the documents is completely ignored.

Similarly, the effects of language relatedness on retrieval effectiveness has been investigated in system oriented studies [5, 6] . Chew at al. [6] investigated script similarity and genetic relatedness for Indo-European and Semitic languages. The authors use Latent Semantic Indexing (LSI) model and manipulated the training data for the LSI model to include text from related languages and unrelated languages. The study concluded that retrieval improves as the number of languages for parallel text in training increases and that text from related languages significantly boosts retrieval. Related to this work, Chavula and Suleman [5] investigated the interplay between language similarity and different indexing strategies using two Bantu languages spoken in Africa and English, and found no differences in retrieval effectiveness when using different indexing strategies and languages with different intelligibility levels. Both studies involved related languages but did not investigate the aspect of understandability or comprehension.

2.2 Document Understandability

Relevance has been argued to be multi-dimensional with notions such as topicality, reliability, scope, novelty and understandability [41]. However, the evaluation of retrieval systems with respect to relevance has been shown to be limited to topicality [7, 27, 29, 41]. Zuccon [43] proposed understandability as an evaluation criteria integrated with topicality, i.e., understandability biased evaluation, based on Gain Discount Framework proposed by Carterette in [4]. This family of measure is based on an assumption that a relevant document is not useful if the searcher cannot understand the contents of the document. This assumption is important to the line of research presented in this paper, i.e., it is necessary to know the threshold of intelligibility a user is able to handle to have successful intercomprehension. The evaluation measure has since been proposed for evaluating consumer health search engines [44].

With the same objective of providing users with relevant and understandable documents, features that capture the understandability of documents have been used to train ranking models for retrieval systems . Palotti et al. [31] used readability features as well as medical vocabulary features to train ranking model for consumer Health corpora, and improved retrieval effectiveness was observed readability features to be in ranking models [42].

Recent years have seen more user centered research in retrieval to model user search behaviour to improve user search experience in different search contexts and information needs [22]. Correspondingly, understandability has been studied in a user controlled experiment. Dodson et al. studied the effect of highlighting in digital text on comprehension and found that relevant highlighting has no effect on comprehension while negative highlighting has negative effect [8].

2.3 Emotions

People experience emotions in all their interactions, and therefore, it is unsurprising that several models for Interactive IR (IIR) have proposed affect or emotion as one of the factors affecting IR interactions [10, 36, 37]. Moreover, previous studies have shown that emotions affect how people search and use information; the research community has focused on what emotions are experienced in search tasks and causes of the triggered emotions, the role of the experienced emotions on search behaviour [25], and how prior emotional state of a searcher or emotiveness of information objects such as music and images can influence his/her choice [35].

Varying emotions are triggered in different search tasks – the emotional polarities experienced in search sessions have been found to correlate with positive attributes of the interaction including successful search completion, easiness of tasks and readability of the document while negative emotions were associated with negative attributes such as frustration and difficulty to find the answer [25]. The objective of our work is similar to the following studies: first, Arapakis at al. [1] studied emotions associated with search tasks of varying difficulty and found that emotional polarity moved from the positive to the negative side of the emotion spectrum when task difficulty changed from low to high, and second, Lopatovska and

Mokros [26] found that simplicity of writing style caused positive emotions on participants who were asked to rate retrieved documents. In the linguistics community, the research objectives have been to identify linguistics and non-linguistics factors or features and statistical metrics that contribute to successful intercomprehension or predict intelligibility of languages, but no studies exist investigating emotions in relation to intercomprehension. In the light of these findings and setting, it is vital to know what emotions are associated with users interacting with search results in which intercomprehension is expected – our assumption was that users would respond differently to this type of search scenario with regard to emotions due to the unnaturalness of reading in an unfamiliar language.

The main objective of the work presented in this paper is to understand user experiences and behaviour in retrieval scenarios where intercomprehension is expected to be applied by the searcher to meet his/her information needs. Our contribution is towards improving user experience of retrieval systems in resource constrained environment. Specifically, this paper extends previous work in the following ways: (1) we examine the issue of relevance and ranking of results written in related languages, (ii) through analysis of search behaviour and user explicit feedback, we explore the issue of intelligibility and user propensity to accept search results that require intercomprehension and the usefulness of such results, finally (iii) we investigate the affective nature of intercomprehension in a retrieval environment.

3 INTELLIGIBILITY

Our work spans across different disciplines and involves topics that have not been widely discussed in the field of information retrieval. Therefore, we introduce intelligibility from linguistics in the light of the studied languages.

Intelligibility is the degree to which a speaker of a language understands the speaker of another closely related language [15]. Intelligibility is known to be affected by linguistic factors including vocabulary, phonetics, morpho-syntax and extra-linguistics factors such as previous language knowledge or exposure and attitude [16]. In linguistics, intelligibility is broadly measured by two methods namely: (i) opinion testing in which L1 speakers of a particular language rate themselves using a scale on how they understand another unfamiliar language under study, and (ii) function testing – participants complete tasks such as translation of a word list or answer multiple choice questions from a text given in the task[14]. Intelligibility is also expressed as linguistic distance, the smaller the distance the more related the languages are [16]. Linguistic distance is estimated using language features such as vocabulary, syntax and morphology. For example, the percentage of the number of cognates (i.e., words with approximate similarity with respect to sound or (orthographic) form and equivalent meaning) across a vocabulary list of two languages is expressed as a lexical distance [19]. Computational approaches based on information theory and statistics use metrics such as entropy[21], surprisal [13, 18, 38] and perplexity distance to estimate intelligibility [9, 11]. These methods also use language features such as vocabulary or corpus as determinants of intelligibility .

Languages investigated in the study belong to the family of Bantu languages in group N. Bantu languages are spoken by over 240 million people found in about twenty eight countries in the Sub-Saharan Africa [30]. Bantu languages are uniquely identified by a character code system of three to four letters proposed by Guthrie (1967 – 71) [17]. The first character in the code, an uppercase letter, indicates the regional zone (A to S) and is followed by two digits in which tens indicate the language group and the ones indicate the individual language. The code sometimes ends with a lowercase letter, which corresponds to a dialect. Languages in group N are spoken in Southern and South–east of Africa in countries including Zimbabwe, Malawi, Zambia, Mozambique and Zambia. Specifically, our languages of focus are identified as Citumbuka (N20), Chichewa(N31a and Zambian dialect Cinyanja(N31b), Cisena(N40) and Citonga(N15). Group N languages are known to have major similarities based on syntax, vocabulary and morphology, and be truly genetically related [30]. Kiso [23] reported that Cisena, Citumbuka and Chichewa are not intelligible, based on information obtained from informants. Chichewa is widely spoken in Malawi, i.e.,taught in most schools, and many of Cisena, Citumbuka and Citonga speakers are familiar with the language. However, many Chichewa speakers are not familiar with Citumbuka, Cisena or Citonga as these languages are only spoken in specific areas. Cinyanja is variant of Chichewa spoken in Zambia and has borrowed from other local languages. Malawi Citonga is only available in Malawi and is spoken by Tonga people on the northern part of Malawi in the lake region. However, not any form of analysis has been done to understand the intelligibility of these languages. For the purposes of our study, we implemented both opinion and function testing to understand the intelligibility of the investigated languages. We added documents written in Luganda (JE15) for the search task to have a language that is new to all participants. Luganda is a Bantu language widely spoken in Uganda by the Buganda people. Section 5 reports intelligibility scores for participants on opinion testing and functional testing task of reading text in other languages.

4 CONTROLLED USER STUDY

To investigate user behaviour and usefulness of intercomprehension in retrieval, a controlled user controlled study was conducted with four tasks: ranking of search results, completing search tasks, emotional reflection and text comprehension.

4.1 Experimental Design

Our strategy to answer the stated research questions holistically was to use the same tasks for all participants regardless of their prior language knowledge. Using a within-subject design, each participant performed four search tasks. Each participant completed the tasks in a single session and the average completion time was 60 minutes. We used Graeco-Latin square to rotate tasks to avoid task sequence interference and to minimise fatigue effect. The study had a single independent variable namely, intelligibility and three dependent variables based on the subtask: (i) emotional experience, (ii) rank, (iii) search task completion and document usefulness, and (iv) comprehension. Pre-defined sets of documents were presented to the participants regardless of their search queries. The languages of

relevant documents were varied and rotated around four languages namely Citumbuka, Chichewa, Cinyanja, Citonga. Additionally, two documents in Luganda and Cisena were included in the retrieved documents but none of these documents were relevant.

4.2 Participants

An invitation to participate in the study was distributed via email to all registered students and through social media outlets for Zambia and Malawi student societies.All participants were living in South Africa at the time of the study. We were particularly interested in the participants language competencies and we used their self-reported mother tongue competency on the registration form to assign them a language for the study. Twenty four respondents(13 male and 11 female) were enrolled into the study. No competency tests were performed for participants to qualify for the study. Participants signed a consent form before taking part in the study.

4.3 Apparatus

4.3.1 Questionnaires. Participants first completed an entry questionnaire, which consisted of demographic questions, language competency questions on five languages, i.e., Citumbuka, Citonga, Cisena, Chichewa and Cinyanja and questions on their search experience using their L_1. A participant's L_1 was used to assign participants to a language in which the study was conducted in. A post-task questionnaire was administered after completing each search task to ascertain the emotional episodes associated with each search task. The questionnaire had four questions taken from Geneva Appraisal Questionnaire (GAQ) (questions 4, 5, 8 and 33). We also adapted GAQ's question 34 for use with Plutchik's wheel [34] to provide more choices of emotions than those listed in GAQ. Plutchik's wheel lists several emotions with varying intensity and is used in studies in which emotional intensity is important [35]. A post-session interview was used to obtain qualitative data to understand more about their general attitude and behaviour towards intercomprehension in retrieval.

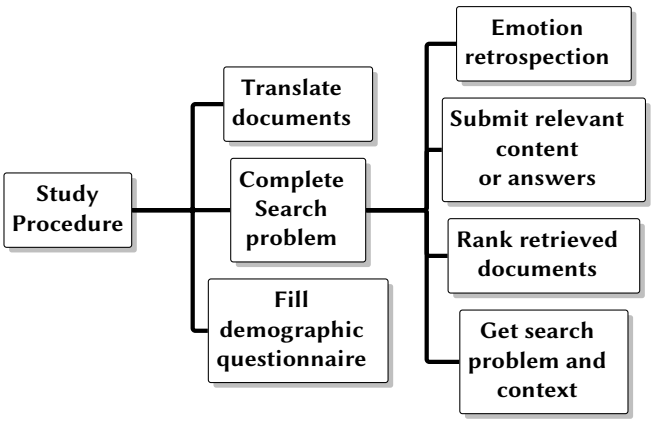

Figure 1: Procedure Used to Conduct the Study

Task	Title	Description
Task 1	Benefits of Drinking water	What are the benefits of drinking water? What is the minimum quantity of water an adult should drink in a day?
Task 2	Prevention of diseases	How can people protect themselves from sicknesses?
Task 3	Life after death	What theories do different communities teach about the place where those who have died go?
Task 4	Origin of life	What theories have people formulated about the beginning of life?

Table 1: Description of tasks completed in the search task

4.3.2 Tasks and Topics. The study consisted of four sub-tasks namely: ranking of search results, completing search tasks, retrospection of emotion episodes and testing of text comprehension. Four search problems were used in the study and all participants completed the same search problems. These search problems were formulated by three assessors selected from the respondents interested to take part in the study.

The assessors translated the tasks to three languages namely, Chichewa, Citumbuka and Cinyanja. Thereafter, the assessors judged the documents for relevance. The assessors used graded relevance to assess the documents based on the following scale: 0 for not relevant, 1 for marginally relevant, 2 for fairly relevant and 3 for highly relevant. In total, 24 documents were presented to participants in the search task and five documents in the text comprehension task. Only six documents were returned in each search task. All participants were presented with the same documents regardless of the language used in the study. However, the search problems were presented to participants in their L_1.

4.4 Procedure

The study was divided into three major tasks:(i) participants first filled a demographic questionnaire, (ii) performed four search tasks and (iii) translated four documents written in other languages as well as submitting a score on how they understood the contents of the document.

4.4.1 Getting Started. Participants participated individually. Each participant was welcomed and was led to the researcher's laboratory where experiments were being conducted. The researcher explained the purpose of the study and tasks to be completed. Thereafter, the participant was asked if he or she was willing to proceed with the experiment and the participant signed a consent form. The researcher explained the procedure for completing the experiment and the participant was given login details for the web application custom built for the study. After a successful login, a tutorial page

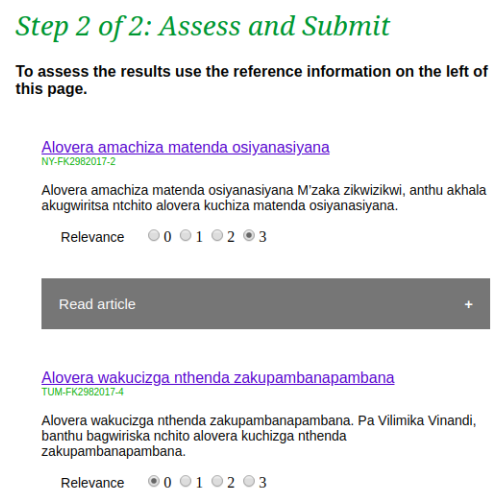

Figure 2: Assessing the documents for relevance page

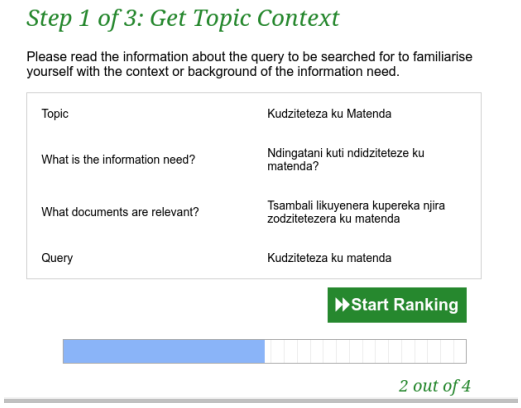

Figure 3: Example of topic presented to participants

about the tasks to be done was loaded and participants proceeded to fill a demographic questionnaire after reading the tutorial.

4.4.2 Search Problem. After completing the questionnaire, participants proceeded to run the study tasks. Firstly, participants were presented with a page explaining an information need. The information need description had the following sections following TREC topic style: title, description and an explanation of what a relevant document should contain. This information was given in the language assigned for the study for each participant. The information provided at this stage was used to complete the following task:

(1) **Ranking** – Participants proceeded with a search task and six documents were retrieved pre-selected for the search problem. The documents were presented to the participant unranked. Participants ranked the retrieved documents by dragging documents to preferred positions – participants were asked to order documents the way they would have preferred a search engine to rank them.
(2) **Search Task completion**– After submitting the preferred order, participants received the same set of documents but ordered using their rank preferences. In this page, the participant was instructed to find the relevant content that meets the search problem. Once the participant was convinced of the answer, the participant clicked on a button to submit: i) the relevant information obtained in the documents written in his/her language used in the study and (ii) the title of documents were the answer or relevant content was found.
(3) **Emotion Introspection** After completing each search task, participants answered questions about their emotions' experience when they were finding the answers to the information need. The questions aimed to understand whether the participant was frustrated or enjoyed as a consequence of reading documents in certain languages to complete the tasks.

After completing each search problem and all the three sub-tasks associated with it, a new search problem was loaded. After completing four search problems, participants were given a page describing the translation task.

4.4.3 Text comprehension. Four documents each written in a different language were used in the text comprehension task depending on L_1 of the participant. Each participant translated the title and first two sentences of a paragraph written in a language different from his/her L_1. Also, the participant gave a scale on how they understood each of the documents using the following reference: '0 for I understand nothing', '1 for I recognize a few words', '2 for I understand a few sentences or some sections', '3 for I understand everything except a few words' and '4 for I understand everything'.

The study was completed after four translation tasks were finished. The participant notified the researcher that the study is complete. The researcher asked the participant a question about the tasks. Finally, the participant signed a payment form and was given a compensation of R100.

5 RESULTS AND ANALYSIS

5.1 Participants Characteristics

5.1.1 Participants Demographics. Participants were of diverse educational background as follows: Science (7), Law (3), Humanities (5), Commerce (2), Engineering (3) and Health Sciences (4). Participants were also studying at different levels namely PhD(7), MSc(6), and Undergraduate(11). They ranged in age from 18 to 50: 18 to 25 (11), 26 - 35 (7) and 36 - 50 (6). Participants were also asked if they had used their mother tongue prior to the experiment to search the web and 16 out of 24 (67%) participants claimed to have used their L_1 to search or read information on the web on topics such as current affairs, music, poems, plays and videos, religious material and translation of words. Participants who had not searched using their mother tongue indicated that they use English to search and believed that they would not find relevant content using their mother tongue.

5.1.2 Language Competencies. Three languages were used by participants(L_1) in the study and the following are the languages: Citumbuka(7), Cinyanja(6) and Chichewa (11). We obtained language competency scores through self reported method before and after the study. Competency scores reported after the study was

done after completing a text comprehension test for a particular language. Figure 4 shows a bar chart for self reported competency levels registered before the study. All the participants had good

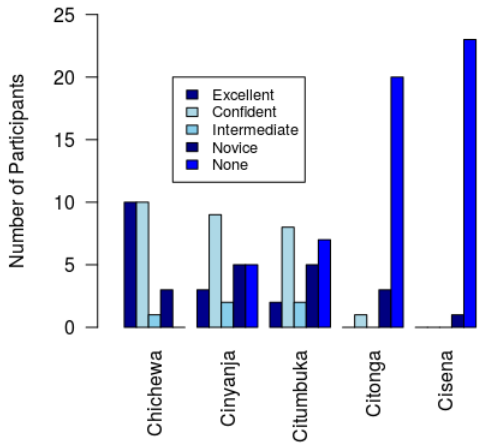

Figure 4: Self-reported competency scores for participants before completing any task in the study

knowledge of Chichewa. Most of the participants had some knowledge of Cinyanja and Citumbuka. Most of the participants had no knowledge of Citonga and Cisena.

Participants were given documents to read and score themselves on how they understood the document on a scale of 0 to 4 towards the end of the study. Figure 4 shows a bar chart of text comprehension scores during the study. The scores reported before and during

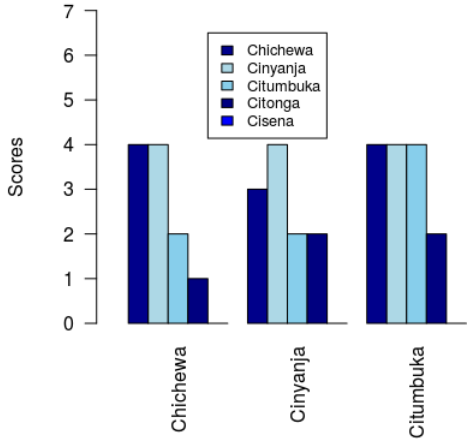

Figure 5: Average reading comprehension scores by L_1

the study using a text comprehension task shows that the self reported scores were lower than those obtained during the study, which may have been due to participants being unfamiliar with the language, i.e., it is hard to give opinion on a language that you are unfamiliar with or encountered. To measure the variation of competency scores measured after the text comprehension task and self-reported or opinion scores, we calculated Intraclass Correlation Coefficients (ICC). ICC estimates and their 95% confidence intervals were calculated based on single measure, absolute-agreement and 2-way mixed-effects model. The ICC value was 0.641 and its 95% confidence interval was between 0.257 and 0.811, which means there is 95% probability that the true ICC value can be at any point between 0.257 and 0.811. We conclude that there is poor to strong agreement between the two used methods. Similar to our results, previous work has found some variation in scores reported using the two methods. Also, previous work has found that functional testing corresponds more to true intelligibility than opinion testing [16]. Accordingly, we proceeded to use the scores for text comprehension in our analysis of results on ranking and emotions.

5.2 Ranking

We were interested to know ranking preferences of users for search results written in related languages and ultimately wanted to answer the following question(s):

RQ1 What are the ranking preferences of users for search results in related languages? Does intelligibility matter in the rank preference of such results?

Participants conducted a ranking task in which they provided their rank preference of six documents for the four tasks. We transformed the position of each document to a rank – the first document in the rank to the value of 1 and proceeded in this manner up to 6 for the last document appearing in the list.

5.2.1 Ranking Agreement. We first calculated correlation coefficients for each ranking provided by each participant against every participant ranking to investigate whether the ranking of participants were similar regardless of their L_1. Figure 6 shows the plot of correlation coefficients of the rankings. There are some correlation for the rankings regardless of L1, which may be due to the ranking of relevant documents. However, two participants ranked documents very differently from the other participants : low to negative correlation coefficients were reported .

5.2.2 Rank Preference by L_1. We next wanted to find out how participants of a particular L_1 ranked documents in each task to observe if there are any differences. We plotted box plots for rankings for each task and grouped the plots by L_1. There were three relevant documents for task 1 written in Chichewa, Citumbuka and Cinyanja documents. Citumbuka and Chichewa documents were highly relevant while the Cinyanja document was fairly relevant. Cinyanja and Citumbuka speakers ranked the Chichewa document highly but most of Citumbuka speakers ranked Citumbuka document highly. This shows some evidence of users preferring documents in their own languages or more closely related languages when multiple relevant documents exist. Almost all Cinyanja and Chichewa participants are not familiar with Citumbuka, and relied on intercomprehension to decide on the ranking, i.e., participants mostly ranked Cinyanja and Chichewa documents on either position 2 or 3. Interestingly, the Luganda document was ranked highly than

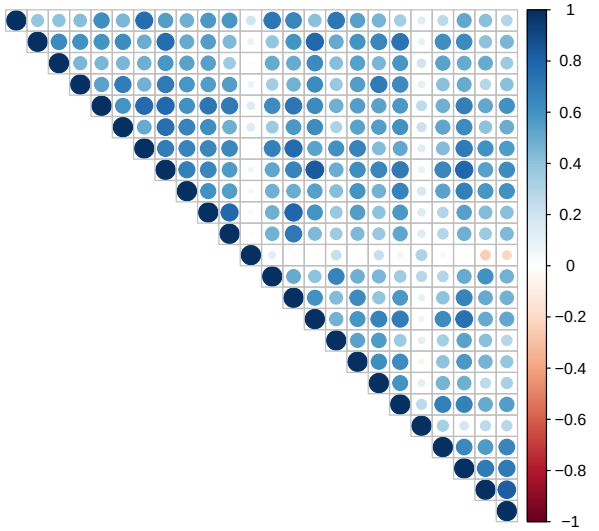

Figure 6: Plot of correlation coefficient of rankings of each participant against every participant.Rankings from two participants reported very low to negative correlation.

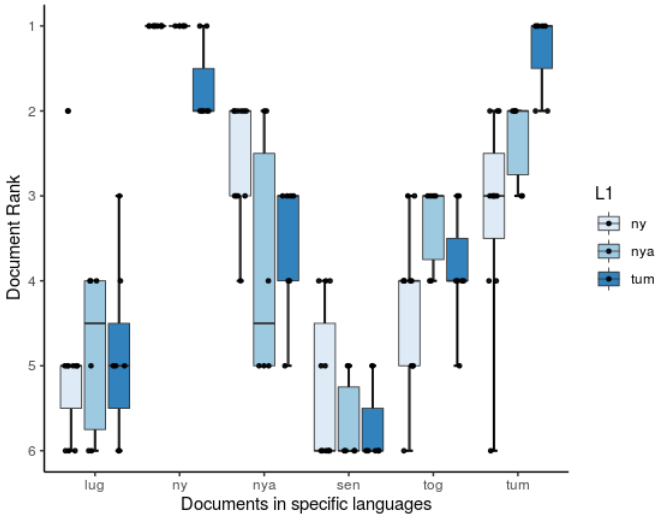

Figure 7: Rank preferences for Task 1 grouped by L_1. The relevant documents for this task was Cinyanja, Citumbuka and Chichewa with relevance scores of 2, 4 and 4 respectively.

Cisena document although both documents were irrelevant and Cisena is in the same family of languages with the rest of the languages. This might be due to participants not interested in ranking documents they know are not relevant but also incomprehensible to them.

Task 2 had only one relevant document written in Citumbuka. Almost all participants ranked the Citumbuka document on first position except one Chichewa participant who ranked it at 6. The ranking of Citonga document was also consistent on position two

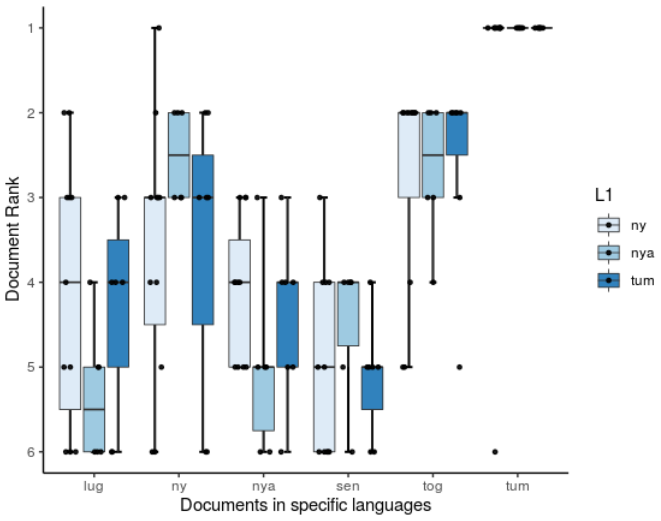

Figure 8: Rank preferences for Task 2 grouped by L_1. The Citumbuka document was the only relevant document in the task.

with a few participants ranking it at different positions(see Figure 8).

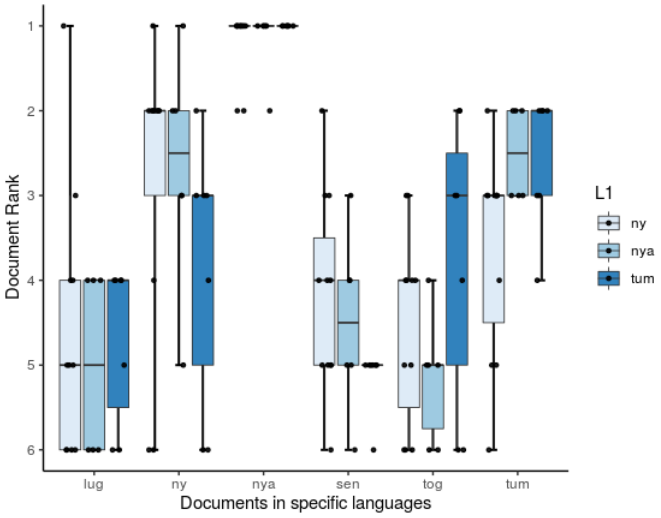

Figure 9: Rank preferences for Task 3 grouped by L_1 and Cinyanja document was the only relevant document. Chichewa , Citumbuka and Citonga documents were on topics in the domain of the task topic.

The relevant document for task 3 was written in Cinyanja. Three documents written in Chichewa , Citumbuka and Citonga discussed related content but their topics were far to that of the task. Most of the participants ranked the Cinyanja document highly (see 9). Citumbuka and Chichewa documents were ranked relatively higher by most participants unlike the Citonga documents, which may

have been due to the low relevance as well as intelligibility, i.e., Citonga is not widely used and almost all the participants were not familiar with the language.

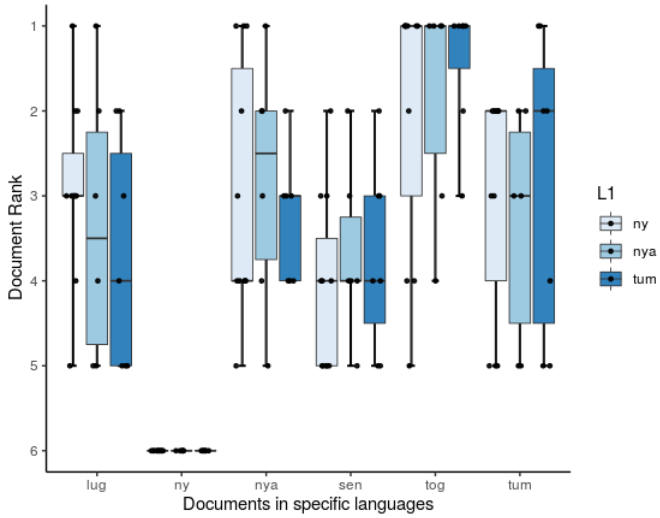

Figure 10: Rank preferences for Task 4 grouped by L_1. The Citonga document was the only relevant document for the task. Luganda and Citumbuka documents discussed broad other unrelated topics in the domain of the task.

Relevant document for the fourth task was written in Citonga, one of the languages most of the participants were unfamiliar with. Ranking this document highly required intercomprehension. 16 out of 24 participants ranked the Citonga document highly on first position. Additionally, there were some documents in other languages that had broad topics related to the search task topic, i.e., Luganda and Citumbuka, and were also ranked highly(see 10).

We observed that preference in ranking was given to more closely related languages when multiple relevant documents existed. Documents with very low relevance (similarity in aboutness of topics) and higher intelligibility were ranked lowly. When documents were not relevant most of the participants did not care about ranking even if the documents were fully comprehensible to them (see 10 for ranking of Chichewa documents (This may have been the last document in the list of the unordered documents when participants first got the search results). The average competency score for Chichewa was four or excellent for Citumbuka and Chichewa speakers and three or confident for Cinyanja speakers).

5.2.3 Ranking by Relevance and Intelligibility. We initially hypothesised that users would want documents to be ranked based on relevance and intelligibility if intercomprehension is assumed as follows: i) Relevant and comprehensible documents should be ranked highly, ii) relevant documents but less comprehensible should follow, iii) if relevance is the same, priority in ranking should be given to more comprehensible documents to the participant. Essentially, ranking should be based on relevance first and intelligibility should be used as a secondary attribute. Our proposed ranking principles are expressed in an algorithm SimRank in 1. Using the proposed approach, we created a ranking for each task for each L_1 – our input being the average scores for reading intercomprehension task for participants of a specific L_1 and relevance judgements provided by monolingual assessors. What follow, is our investigation on the similarity or distance between the hypothetical rankings and rankings provided by participants. We proceeded with two types of tests as proposed in [28] for comparing rankings in retrieval using Kendall Tau τ and Kolmogorov-Smirnov Test.

Data Transformation. We merged the rankings of participants by L_1 and task using Borda Count voting model. Borda Count is an election method in which voters rank candidates by preference and the winner is chosen based on the points accumulated from beating other candidate. Borda Count has previously been applied in retrieval in the context of aggregating meta-search results, and has produced competitive results relative to more advanced techniques using supervised learning [2, 24, 40].

Given a set of rankings for task i (i = 4), $R_i = R_{i1}, R_{i2}, \ldots, R_{im}$ (where m is the number of participants using $L1$ k) of a set of documents $D_i = d_{i1}, d_{i2}, \ldots, d_{in}$ where $n = 6$. For each ranking R_i , assign to document d_{ij} points equal to the number of documents ranked lower than itself, i.e, a document ranked on first position gets n or 6, second position gets $n - 1$ or 5, third position 4 and last position gets 1. The total count for document d_j is the number of points it accumulates from all its rankings seen in the sample for participants with this L_1 on this task. The accumulated points are used to rank documents in descending order for task i using $L1$ k.

Ranking Correlation using Kendall Tau. Kendall Tau measures the strength of association between two sets of ranks given to a same set of objects. We test the null hypothesis that Kendall Tau = 0, i.e., the two sets of ranks are not similar. The alternative hypothesis is that the ranks are correlated, i.e., Kendall Tau is non-zero. The

L_1	Kendall Correlation	p-value
Citumbuka	.6096	0.0001835
Chichewa	0.625	0.0001232
Cinyanja	0.6333	9.993e-05

Table 2: Kendall Correlation by L_1

Kendall Tau statistic values indicate that there is a strong correlation between the observed user ranking and the expected rankings from the hypothetical ranking algorithm. The p-values are very small ($p < 0.05$) and we reject the null hypothesis(Tau = 0). Therefore, we conclude that the ranking provided by the user is similar to the hypothetical rankings.

Goodness of Fit Test. We proceeded with the Kolmogorov–Smirnov(K-S) Test to see if the empirical ranks from the sample come from the same distribution or follows the distribution of our hypothetical ranking. The K-S statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. K-S Test is suitable for ordered categorical data [20] with a large sample size and to off-set that we use bootstrap method to estimate the best p-value for D.

We test the null hypothesis that the two sets of ranks come from the same distribution. The obtained D values indicate that the two samples come from the same distribution, i.e. D is close or equal to zero (D gives the maximum distance between the Cumulative Distribution Functions(CDFs) of the two samples). We accept the null hypothesis that the distributions are the same, i.e., our p-value is 1(for significance level p > 0.05) The results of Kendall Tau and

L_1	Kolmogorov–Smirnov D	p-value
Citumbuka	0.041667	1
Chichewa	0	1
Cinyanja	0	1

Table 3: Kolmogorov–Smirnov Statistic and test by L_1

K-S statistics and metrics indicate that the two rankings are similar.

5.3 Emotions and Search Completion

Participants were asked to complete a search task by submitting the title(s) of relevant document and topic answers that they found from the documents. The first task had three relevant documents written in Cinyanja, Chichewa and Citumbuka. The topic was an informational task and information required were facts. All the participants were able to complete the task.

The second and third topics were also fact topics with one relevant document each written in Citumbuka and Cinyanja respectively. Two Chichewa and one Cinyanja participants did not complete the second task, i.e., were not able to submit answers for the task. Three Chichewa speakers were not able to complete the third task. Analysing the data further showed that the same two participants did not complete tasks in both tasks and their rankings were out of agreement with those provided by other participants sharing the same $L1$.

The relevant document for the fourth task was written in Citonga, and was written in such a way that the reader needed to understand the presented content to submit an answer. Moreover, the document was written in a language participants were not familiar with. Five participants were not able to complete the task.

5.3.1 Emotions. Emotions affect behaviour and may cause users to approach or avoid a system. Therefore, the user study is designed to explore the affective aspects of intercomprehension in a retrieval scenario. Our aim is to understand the level of enjoyment and frustration a user could experience in applying intercomprehension when reading search results. The effort required to understand a document written in an unfamiliar language may lead to frustration and stop a user from completing their search task completely.

RQ3 What type of emotions, i.e., negative and positive, do users experience when interacting with search results that require intercomprehension?

After each search episode participants provided their emotion experience using a Geneva Appraisal Questionnaire (GAQ) and PlutchikâĂŹs wheel. The PlutchikâĂŹs wheel allowed participants to explicitly specify the type of emotion that have just experienced in the task. We classified the emotions provided into two classes

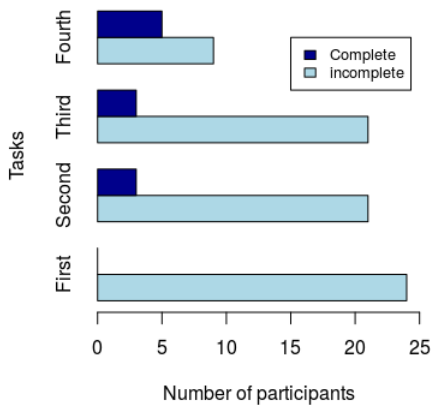

Figure 11: Task completion status by task

namely negative and positive emotions for each participant for each task [34]. Figure 12

Participants were also asked to indicate whether they struggled to complete a task. We wanted to know if there is any association between the type of emotion and whether a participant struggled or not. We conducted Fisher's exact test of independence (p > 0.05) with null hypothesis that emotion type and struggling status are independent, i.e., the probability of experiencing positive or negative emotions is the same whether someone struggling with a task or not. The odds ratio shows how strong the association between struggling and emotions is. The results for Fisher exact test provide

Task	p-value	odds	CI
Task 1	0.001976	0	0.0000000 to 0.3225179
Task 2	0.01087	0	0.0000000 to 0.7203821
Task 3	0.5212	0.4882077	0.02013983 to 34.63301267
Task 4	0.357	0	0.00000 to 23.40011

Table 4: Fisher's exact test by task

mixed results. For task 1 the p-value is very small as well as the value of the odds ratio. our null hypothesis for Fisher's exact test is that struggling and emotions are independent. Therefore, the null hypothesis is rejected (p<0.05 for 95% significance level). A very small odds ratio says that the differences are big. The confidence interval for the odds ratio is small, i.e., the odds ratio has been precisely estimated. This is the same for Task 2. Surprisingly, for task 3 the p-value is big, the null hypothesis is accepted (p>0.05 for 95% significance level). A very small odds ratio says that the differences are big. The confidence interval for the odds ratio is large, i.e., low precision for the odds ratio. This is the same for Task 4. Ignoring, the odds ratio value, we can conclude that the null hypothesis is rejected for Task 1 and Task 2 and accepted for Task 3 and Task 4. Using the odds ratio only indicates that struggling affects emotions.

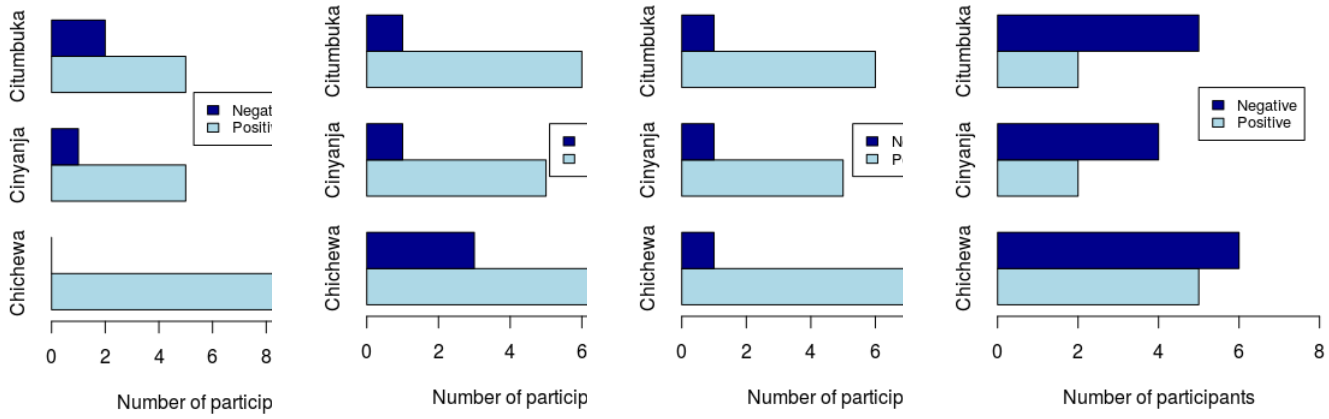

Figure 12: Classification of emotions

6 DISCUSSION

We have shown evidence that search results in closely related languages are useful to the user through our controlled user study.The study provided insights on how users may interact with search results were intercomprehension is expected. Our analysis has shown that users can easily identify relevant documents in related language and can understand the contents based on how intelligible their languages are. This is very important in resou;rce constrained environment were digital content written in local languages is limited.

The study has shown that users ranked highly relevant documents in their mother tongue in compared to documents in other related languages. However, users wanted to see irrelevant documents in their mother tongue ranked lower than fairly relevant documents written in related languages. In terms of completing a search task, our results indicate that users struggled with unfamiliar languages,i.e., their first encounter with the language was in the study. Surprisingly, some participants with similar language profiles enjoyed and completed the tasks with less struggle - such participants reported positive emotions at the end of the task. This may be due to the subjective factors such as personality traits that may affect how people are willing to accept or explore new things [16]. This dimension was not explored in this current study. In cases where users are not able to understand content, users may become frustrated and experience negative emotions. Negative emotions in retrieval have been shown to impact users negatively [25], and may negatively affect the experience of users and make intercomprehension undesirable.Therefore, it is necessary to study how it can be incorporated in retrieval scenarios through methods such as personalisation.

It is possible that a similar study may have different outcomes due to the size of sample in terms of number of languages being explored and their attributes and number of people in the study and their language competencies languages with different attributes. One of the limiting factors in our study was to find people with no knowledge of other languages. This was particularly difficult for speakers of Citumbuka as Chichewa is used widely as a lingua franca as well as taught in schools in Malawi. Additionally, true intelligibility is relative and depends on different factors, and cannot be replicated across individuals.

7 CONCLUSION

We have explored user interaction behaviour in a retrieval scenario in which intercomprehension is expected. In particular, through our user study tasks and analysis, we have shown that ranking results based on intelligibility and relevance is useful in cases of limited relevant search results. We also investigated the type of emotions that users performing intercomprehension in retrieve scenario may experience and have found that users experience both positive and negative emotions: with a combination of positive and negative emotions in cases were intercomprehension is expected. Our future work will investigate how to rank search results with assumed intercomprehension using learning to ranking methods.

REFERENCES

[1] Ioannis Arapakis, Joemon M. Jose, and Philip D. Gray. Affective feedback: An investigation into the role of emotions in the information seeking process. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 395–402, New York, NY, USA, 2008. ACM.

[2] Javed A. Aslam and Mark Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 276–284, New York, NY, USA, 2001. ACM.

[3] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. Using clustering and superconcepts within smart: Trec 6. *Information Processing & Management*, 36(1):109–131, 2000.

[4] Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 903–912, New York, NY, USA, 2011. ACM.

[5] Catherine Chavula and Hussein Suleman. Assessing the impact of vocabulary similarity on multilingual information retrieval for bantu languages. In *Proceedings of the 8th Annual Meeting of the Forum on Information Retrieval Evaluation*, FIRE '16, pages 16–23, New York, NY, USA, 2016. ACM.

[6] Peter A. Chew and Ahmed 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.

[7] Erica Cosijn and Peter Ingwersen. Dimensions of relevance. *Inf. Process. Manage.*, 36(4):533–550, July 2000.

[8] Samuel Dodson, Luanne Freund, and Rick Kopak. Do highlights affect comprehension?: Lessons from a user study. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 381–384, New York, NY, USA, 2017. ACM.

[9] Andrea K. Fischer, Jilles Vreeken, and Dietrich Klakow. Beyond pairwise similarity: Quantifying and characterizing linguistic similarity between groups of languages by MDL. *Computación y Sistemas*, 21(4), 2017.

[10] Nigel Ford. *Frontmatter*, pages i–iv. Facet, 2015.

[11] Pablo Gamallo, JosÃĺ Ramom Pichel, and IÃśaki Alegria. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152 – 162, 2017.

[12] Fredric 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.

[13] Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018, Salt Lake City, Utah, USA, January 7, 2018*, pages 10–18, 2018.

[14] Charlotte Gooskens. *Methods for measuring intelligibility of closely related language varieties*, pages 195–213. Oxford University Press, 2013.

[15] Charlotte Gooskens. *Dialect Intelligibility*, chapter 11, pages 204–218. John Wiley Sons, Ltd, 2018.

[16] Charlotte Gooskens and Femke Swarte. Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics*, 40(2):123âĂŞ147, 2017.

[17] Malcolm Guthrie. *Comparative Bantu : an introduction to the comparative linguistics and prehistory of the Bantu languages.* Farnborough : Gregg, 1967.

[18] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[19] Wilbert Heeringa, Jelena Golubovic, Charlotte Gooskens, Anja Schüppert, Femke Swarte, and Stefanie Voigt. *Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance*, pages 99–137. P.I.E. - Peter Lang, 2013.

[20] Ben Jann. Multinomial goodness-of-fit: large sample tests with survey design correction and exact tests for small samples. ETH Zurich Sociology Working Papers 2, ETH Zurich, Chair of Sociology, January 2008.

[21] Moberg Jens, Charlotte Gooskens, John Nerbonne, and Nathan Vaillette. Conditional entropy measures intelligibility among related languages. 7:51–66, 2007.

[22] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1—2):1–224, January 2009.

[23] Andrea Kiso. *Tense and aspect in Chichewa, Citumbuka and Cisena: A description and comparison of the tense-aspect systems in three southeastern Bantu languages.* PhD thesis, Department of Linguistics, Stockholm University, 2012.

[24] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. Supervised rank aggregation. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 481–490, New York, NY, USA, 2007. ACM.

[25] Irene Lopatovska and Ioannis Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Inf. Process. Manage.*, 47(4):575–592, July 2011.

[26] Irene Lopatovska and Hartmut B. Mokros. Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Inf. Process. Manage.*, 44(1):92–104, January 2008.

[27] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472, New York, NY, USA, 2016. ACM.

[28] Massimo Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, June 2007.

[29] Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[30] D. Nurse and G. Philippson. *The Bantu Languages.* Routledge Language Family Series. Taylor & Francis, 2006.

[31] João R. M. Palotti, Lorraine Goeuriot, Guido Zuccon, and Allan Hanbury. Ranking health web pages with relevance and understandability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 965–968, 2016.

[32] doye Peter. Intercomprehension : Reference study, 2005.

[33] Carol Peters, Martin Braschler, and Paul D. Clough. *Multilingual Information Retrieval - From Research To Practice.* Springer, 2012.

[34] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.

[35] Lev Poretski, Joel Lanir, and Ofer Arazy. Feel the image: The role of emotions in the image-seeking process. *HumanâĂŞComputer Interaction*, 34(3):240–277, 2019.

[36] Tefko Saracevic. The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the ASIST Annual Meeting*, 34:313, 1997.

[37] Reijo Savolainen. Emotions as motivators for information seeking: A conceptual analysis. *Library Information Science Research*, 36(1):59 – 65, 2014.

[38] Irina Stenger, Klara Jagrova, Andrea Fischer, Tania Avgustinova, Dietrich Klakow, and Roland Marti. Modeling the impact of orthographic coding on Czech–Polish and Bulgarian–Russian reading intercomprehension. *Nordic Journal of Linguistics*, 40(2):175âĂŞ199, 2017.

[39] Andreas von Holy, Alon Bresler, Osher Shuman, Catherine Chavula, and Hussein Suleman. Bantuweb: A digital library for resource scarce south african languages. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '17, pages 36:1–36:10, New York, NY, USA, 2017. ACM.

[40] Shengli Wu. *Data Fusion in Information Retrieval.* Springer Publishing Company, Incorporated, 2012.

[41] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, May 2006.

[42] Xin Yan, Dawei Song, and Xue Li. Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 540–549, New York, NY, USA, 2006. ACM.

[43] Guido Zuccon. Understandability biased evaluation for information retrieval. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 280–292, 2016.

[44] Guido Zuccon and Bevan Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Proceedings of the Medical Information Retrieval Workshop at SIGIR co-located with the 37th annual international ACM SIGIR conference (ACM SIGIR 2014), Gold Coast, Australia, July 11, 2014.*, pages 32–35, 2014.