Ranking by Language Similarity for Resource Scarce Southern Bantu Languages

Catherine Chavula
University of Cape Town
Cape Town, South Africa
cchavula@cs.uct.ac.za

Hussein Suleman
University of Cape Town
Cape Town, South Africa
hussein@cs.uct.ac.za

ABSTRACT

Resource Scarce Languages (RSLs) lack sufficient resources to use Cross-Lingual Information Retrieval (CLIR) techniques and tools such as machine translation. Consequentially, searching using RSLs is frustrating and usually ends in unsuccessful struggling search. In such search tasks, search engines return low-quality results; relevant documents are either limited and lowly ranked or non-existent. Previous work has shown that alternative relevant results written in similar languages, including dialects, neighbouring and genetically related languages, can assist multilingual RSLs speakers to complete their search tasks. To improve the quality of search results in this context, we propose the re-ranking of documents based on the similarity between the language of the document and the language of the query. Accordingly, we created a dataset of four Southern Bantu languages that includes documents, topics, topical relevance and intelligibility features, and document utility annotations. To understand the intelligibility dimension of the studied languages, we conducted online intelligibility test experiments and used the data for feature selection and intelligibility prediction. We performed re-ranking of search results using offline evaluation, exploring Learning To Rank (LTR). Our results show that integrating topical relevance and intelligibility in ranking slightly improves retrieval effectiveness. Further, results on intelligibility prediction show that classification of intelligibility is feasible at a fair accuracy.

CCS CONCEPTS

• **Information systems** → **Multilingual and cross-lingual retrieval**; **Learning to rank**.

KEYWORDS

Multilingual Information Retrieval, Retrieval Models and Ranking

ACM Reference Format:
Catherine Chavula and Hussein Suleman. 2021. Ranking by Language Similarity for Resource Scarce Southern Bantu Languages. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3471158.3472251

1 INTRODUCTION

Searching and finding information on the Web plays a key role in operating effectively in modern society. However, there is very little content written in many languages of the world on the Web, and English and other widely spoken languages continue to dominate the Web. Resource Scarce Languages (RSLs) refers to such languages that lack large monolingual or parallel corpora and other linguistic resources sufficient to build Natural Language Processing (NLP) applications [43]. Prior Information Retrieval (IR) research, particularly Cross-Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR), use translation of either queries or documents to allow users to access information in their own languages [47]. However, RSLs lack such resources and tools to enable CLIR and MLIR [54]. Consequently, speakers of RSLs struggle to find information written in their local languages on the Web [13]. Queries in RSLs usually return low quality search results: relevant content is limited and in some cases non-existent [13, 42]. In cases of limited data availability, search engines often rank relevant documents lowly because such queries are rare and lack previous examples to learn from [22], and this usually leads to users being presented with irrelevant results in other Web dominant languages that they may not be able to understand [13, 42].

Matching and presenting users with search results written in related languages has been shown to reduce users' frustration in the context where users speak RSLs, the languages are highly similar and the users are highly multilingual [14]. Unfortunately, there are some challenges for such search systems to overcome. Notably, users may experience difficulty in reading documents written in a language that is unfamiliar to them - and hence the need to maximise the gains between intelligibility and topical relevance through search results re-ranking. Intelligibility in this context refers to the degree to which a speaker of a language understands written text of another closely related language [25].

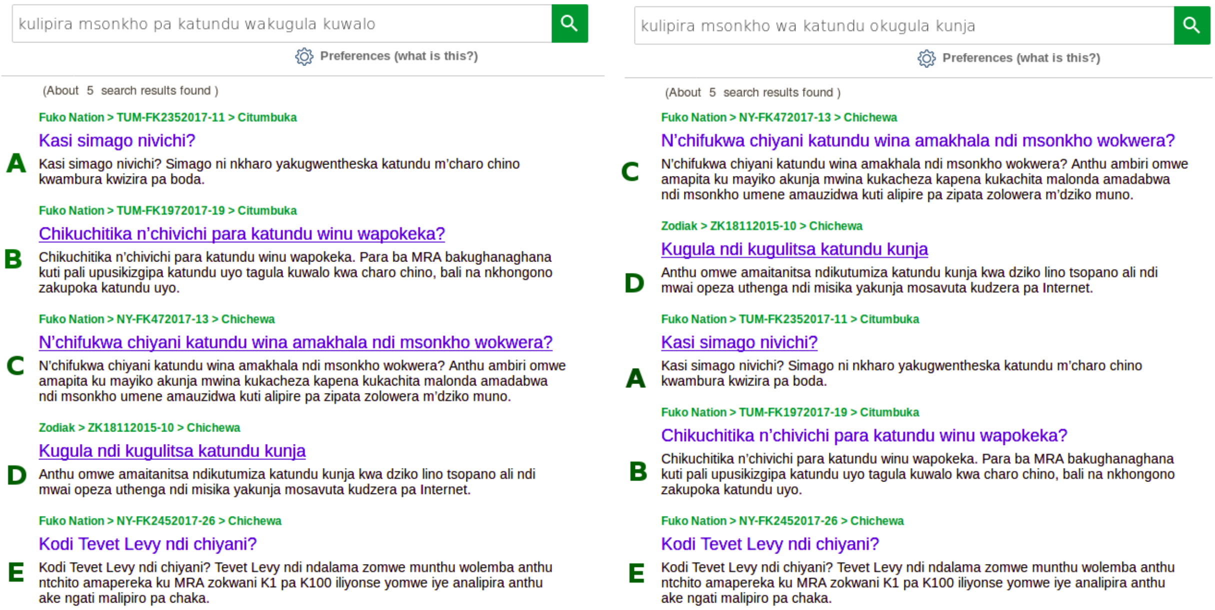
Imagine a user, who is a retail trader looking for information on tax on imported goods. If we assume that the user is a monolingual speaker of Citumbuka, for example, the user's preference would be Citumbuka documents over any other returned by the search engine. The same behaviour would be expected if the user was a monolingual Chichewa speaker. For example, Figure 1 shows two pages with five search results on '*tax on imported goods*', i.e., *A*, *B*, *C*, *D* and *E* ranked differently for Citumbuka or Chichewa monolingual speakers. The left side illustrates the results for Citumbuka speakers and the right Chichewa monolingual speakers. The results are ranked based on topical relevance and intelligibility. All the results are relevant except for document *E*, which cannot change its rank regardless of the user's first language. Although the other four documents are equally relevant, their ranks could change based on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICTIR '21, July11, 2021, Virtual Event, Canada

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472251



kulipira msonkho pa katundu wakugula kuwalo

(About 5 search results found)

Fuko Nation > TUM-FK2352017-11 > Citumbuka

A Kasi simago nivichi?

Kasi simago nivichi? Simago ni nkharo yakugwentheska katundu m'charo chino kwambura kwizira pa boda.

Fuko Nation > TUM-FK1972017-19 > Citumbuka

B Chikuchitika n'chivichi para katundu winu wapokeka?

Chikuchitika n'chivichi para katundu winu wapokeka. Para ba MRA bakughanaghana kuti pali upusikizgipa katundu uyo tagula kuwalo kwa charo chino, bali na nkhongono zakupoka katundu uyo.

Fuko Nation > NY-FK472017-13 > Chichewa

C N'chifukwa chiyani katundu wina amakhala ndi msonkho wokwera?

N'chifukwa chiyani katundu wina amakhala ndi msonkho wokwera? Anthu ambiri omwe amapita ku mayiko akunja mwina kukacheza kapena kukachita malonda amadabwa ndi msonkho umene amauzidwa kuti alipire pa zipata zolowera m'dziko muno.

Zodiak > ZK18112015-10 > Chichewa

D Kugula ndi kugulitsa katundu kunja

Anthu omwe amaitanitsa ndikutumiza katundu kunja kwa dziko lino tsopano ali ndi mwai opeza uthenga ndi misika yakunja mosavuta kudzera pa Internet.

Fuko Nation > NY-FK2452017-26 > Chichewa

E Kodi Tevet Levy ndi chiyani?

Kodi Tevet Levy ndi chiyani? Tevet Levy ndi ndalama zomwe munthu wolemba anthu ntchito amapereka ku MRA zokwani K1 pa K100 iliyonse yomwe iye analipira anthu ake ngati malipiro pa chaka.

kulipira msonkho wa katundu okugula kunja

(About 5 search results found)

Fuko Nation > NY-FK472017-13 > Chichewa

C N'chifukwa chiyani katundu wina amakhala ndi msonkho wokwera?

N'chifukwa chiyani katundu wina amakhala ndi msonkho wokwera? Anthu ambiri omwe amapita ku mayiko akunja mwina kukacheza kapena kukachita malonda amadabwa ndi msonkho umene amauzidwa kuti alipire pa zipata zolowera m'dziko muno.

Zodiak > ZK18112015-10 > Chichewa

D Kugula ndi kugulitsa katundu kunja

Anthu omwe amaitanitsa ndikutumiza katundu kunja kwa dziko lino tsopano ali ndi mwai opeza uthenga ndi misika yakunja mosavuta kudzera pa Internet.

Fuko Nation > TUM-FK2352017-11 > Citumbuka

A Kasi simago nivichi?

Kasi simago nivichi? Simago ni nkharo yakugwentheska katundu m'charo chino kwambura kwizira pa boda.

Fuko Nation > TUM-FK1972017-19 > Citumbuka

B Chikuchitika n'chivichi para katundu winu wapokeka?

Chikuchitika n'chivichi para katundu winu wapokeka. Para ba MRA bakughanaghana kuti pali upusikizgipa katundu uyo tagula kuwalo kwa charo chino, bali na nkhongono zakupoka katundu uyo.

Fuko Nation > NY-FK2452017-26 > Chichewa

E Kodi Tevet Levy ndi chiyani?

Kodi Tevet Levy ndi chiyani? Tevet Levy ndi ndalama zomwe munthu wolemba anthu ntchito amapereka ku MRA zokwani K1 pa K100 iliyonse yomwe iye analipira anthu ake ngati malipiro pa chaka.

Figure 1: Ranking by relevance and intelligibility for a Citumbuka speaker on the left and Chichewa speaker on the right for the same search task

the language of the query. The assumption here is that users would assign higher utility to topically relevant documents written in their first language. Therefore, a retrieval system that presents users with documents written in related languages needs to optimise the utility of the returned search results by presenting users with documents that are highly relevant and comprehensible to them.

We propose re-ranking of search results by integrating traditional relevance features to estimate topical relevance and intelligibility features for intelligibility estimation. Our ranking problem is that of constructing a ranking model that finds the best combination of relevance and intelligibility features. We model the problem as a supervised ranking problem, i.e., automatically constructing a ranking model using training data. We also explored unsupervised ranking by using a weighted combination of topical relevance and intelligibility based on relative importance of these factors.

Previous work on related languages retrieval focused on understanding user behaviour [14], and matching of queries and documents across languages without translation [7, 13, 16, 21, 29]. Such studies have provided insights on how users may interact with such results as well as evidence that using similarities in languages is feasible in the search context of resource constrained languages. However, these studies do not provide the means of ranking and presenting search results with variable document intelligibility due to language variation.

One of the challenges of incorporating intelligibility in ranking is choosing a set of features to be used by the ranking model: known intelligibility determinants have mostly been studied in isolation, and mainly focusing on Indo-European languages. In our current investigation, we are interested in a representative set of intelligibility determinants that would give the best prediction accuracy for a model that classifies how a speaker of L_1 understands written text of languages closely related to their L_1. We focus on a few languages in the cluster of Southern Bantu languages namely: Chichewa [38], Citumbuka [15], Cinyanja as spoken in Zambia [31] and Citonga [40] as spoken in Malawi. These languages have limited digital content available on the Web and lack tools such as bilingual dictionaries and machine translation. The main contributions of our work are as follows:

- We develop a dataset that has features estimating both topical relevance and intelligibility between the language of the query and document.
- We conduct feature selection on several intelligibility features and propose a set of features that are strong predictors of intelligibility. We also perform intelligibility prediction as a multi-class classification task.
- We develop ranking models that integrate relevance and intelligibility features. We show that we can improve the quality of search results written in related languages when intelligibility is considered in ranking.

The rest of the paper is structured as follows. We first discuss related work in Section 2. We then describe the dataset used in our experimentation in Section 3. We provide the results for feature selection and intelligibility prediction in Section 4. This is followed by a presentation of our ranking experimental setting and results in Section 5. Section 6 provides a discussion of the results, including implications and limitations of our approach. Finally, Section 7 concludes and provides future direction of the work presented in the paper.

2 RELATED WORK

Re-ranking search results based on the similarity between the language of the query and language of the document uses different techniques from IR and linguistics. Relevant topics to our work include: (i) retrieval for related languages, (ii) user perspectives in retrieval beyond topical relevance, and (iii) intelligibility of related languages.

2.1 Related Languages Retrieval

Research on retrieval for related languages has so far mainly focused on using language similarities, such as vocabulary similarity, to retrieve documents written in related languages without the translation step typically used in CLIR and MLIR systems [7, 13, 16, 21, 29]. Generally, untranslated queries [13, 16, 21], together with fuzzy string similarity matching methods [7, 29], have been used to match index and query terms for closely related languages. This is done on the premise that matching is possible due to cognates across languages and avoids the costs associated with translation systems. Results obtained using this approach are comparable or worse relative to classic approaches such as using a bilingual dictionary for query translation. While these studies only explored the matching of queries and documents for similar languages, they are an appropriate starting point to investigate any opportunities when similar languages are involved. In this paper, we focus on ranking models that combine topical relevance and intelligibility as ranking criteria for such search results.

2.2 User Perspectives Beyond Topical Relevance

Notions of Relevance: Several studies have demonstrated that relevance has multiple dimensions for users in different contexts, including topicality, novelty, reliability and understandability [18, 39, 51, 61]. For example, Xu and Chen [61] found that understandability and reliability were secondary relevance criteria while topicality and novelty were primary. Likewise, Chavula and Suleman [14] asked participants to rank search results written in related languages and found that relevance was used as a primary criteria and intelligibility was a secondary criteria. Similar to our task of intelligibility prediction, Steichen et al. [52] conducted a study to predict the search result list language preferences for multilingual users based on their L_1, user's subjective features, and topic features, and obtained a fair accuracy. These results provide insights on user preferences in terms of the notions of relevance – demonstrates the primacy of topical relevance – there is no question to the critical importance of topical relevance. Further, the results provide insights on how users in different contexts may interact with search results, and how relevance in its entirety may shift.

Understandability: Our work is similar in spirit with the task of ranking documents based on relevance and understandability [45] or readability [17] of documents. Palotti et al. [45] ranked health Web pages for topical relevance and understandability to improve readability of documents for non-expert health information users. Their relevance features focused on the query and document similarity scores and the readability features of documents captured surface level properties of text. Similarly, with the goal of personalising search results based on user reading level, Collins-Thompson et al. [17] re-ranked documents based on estimated reading level of the user and reading difficulty of documents. These studies are similar to our ranking approach, with the difference that the studies focused on understandability of documents [45] and reading level of a user [17] of text of the same language. Our work uses text based similarity features to estimate intelligibility of related languages to re-rank documents.

Evaluation metrics: Evaluation metrics that account for the different dimensions of relevance have been proposed, including understandability. Zuccon [64] proposed and used [65] understandability as an evaluation criteria integrated with topicality, i.e., understandability biased evaluation, based on the Gain Discount Framework proposed by Carterette [10]. This family of measures 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 our work since the goal of re-ranking is to provide users with highly relevant and intelligible results early in the search results list.

2.3 Intelligibility of Related Languages

Languages are diverse and are always changing. Similarities among languages may stem from close genetic relations such as dialects or through contact – for example, through lexical borrowing among neighbouring languages in the same geographical area [9, 59]. In the linguistics community, research on intelligibility has focused on identifying factors that determine intelligibility mostly for Indo-European languages [24]. These factors have been divided into two categories [57], namely: linguistic and extra-linguistic factors. Linguistic features are based on inherent language similarities at different levels of linguistic description, including lexical [28], phonological [28], orthographic [53] and syntactical [26]. Extra-linguistic features are subjective features that are dependent on an individual's prior language experiences, attitudes and personality traits [25]. Recent research on estimating intelligibility from the perspective of Information Theory has focused on modelling the cognitive processes in reading text in related languages [23], such as conditional entropy [53] and surprisal [23, 27]. Attributes that exploit the statistical language properties of text have also been proposed in literature to provide statistical evidence for intelligibility through language modelling [19, 20].

Similar to our task of feature selection, Kürschner et al. [34] used regression on intelligibility scores from Danish speakers presented with Swedish words and found that Levenshtein distance had higher importance in predicting intelligibility. Gooskens and Swarte [26] investigated the relative importance of linguistic and extra-linguistic predictors of mutual intelligibility using regression analysis for five Germanic languages (Danish, Dutch, English, German and Swedish), and found that extra-linguistic factors were strong predictors but attitude had less effect. Linguistic distances such as lexical and phonetic distances were found to be stronger predictors than syntactic factors. Our work uses a different approach to investigate intelligibility of languages that have not been widely explored in this context.

2.4 Contributions Over Previous Work

The paper makes several contributions over previous work. First, we focus on the task of intelligibility prediction, a task that is still unexplored in both linguistics and machine learning. Although, we draw features from several studies in linguistics, our work investigates languages that have not been studied widely in this context. Our investigation includes feature selection exploring different types of features and we model intelligibility scores as classes. This is different from previous studies where intelligibility was estimated as a continuous relative measure [26, 34]. Second, we introduce two metrics that have not been used in intelligibility studies. These features can be explored further as determinants of intelligibility and used in linguistic studies. Finally, we investigate re-ranking of search results using topical relevance and intelligibility features for RSLs. Studies involving RSLs are very rare in IR. The problem of lack of enormous resources of RSLs means that the traditional approaches employed for multilingual retrieval is infeasible due to the need for translation. Our approach offers multilingual search results by using language similarities among languages that are highly similar [36] for users who are highly multilingual [44] by incorporating intelligibility features in ranking. The approach uses minimal data and the features can be used individually with topical relevance features. Offering users alternative results in closely related languages increases the interaction involving RSLs and can provide more legitimate training examples for retrieval models.

3 EXPERIMENTAL DATA

In this section, we describe the process of preparing the data used in the feature selection and ranking experiments. Our experimental data consists of topics in three languages, documents in five languages (only four are used in the experiments), utility labels, linguistic intelligibility features, and extra-linguistic intelligibility features.

3.1 Languages

Our study involved languages that belong to the Bantu family spoken in Southern and South–eastern Africa. Specifically, we included Citonga, Citumbuka, Cinyanja and Chichewa. Cisena was used in the preliminary stages of collecting information but was not used in the evaluations due to lack of participants in some of the studies. Citonga, Chichewa and Citumbuka are neighbouring languages that have borrowed words from each other, and have been classified to belong to the same language family cluster [44]. Chichewa is widely spoken in Malawi and is taught as a subject in most schools, and as such, many of Citumbuka and Citonga speakers are familiar with Chichewa language. However, many Chichewa L_1 speakers are not familiar with Citumbuka or Citonga as these languages are mostly spoken in specific areas. Cinyanja is a dialect of Chichewa spoken in Zambia and has borrowed from other local languages in Zambia. Malawian Citonga is spoken only in Malawi by the Tonga people in the lake region of northern Malawi. Citumbuka is a language spoken by the Tumbuka people in Malawi and Zambia.

3.2 Documents

Due to limited availability of information in digital format, and due to copyright constraints, we obtained information from two media houses in the form of newspaper articles and news bulletins. Topics in these documents include current and development news, as well as health and religious articles. The radio news bulletins are written in Chichewa and English, while the newspaper articles are written in Chichewa and Citumbuka. The rest of the documents that form part of the collection are written in Citumbuka, Citonga, Cinyanja and Cisena, and were obtained from the Web. Documents were converted from such formats as PDF and HTML (Web pages) into text files. The text file documents were cleaned by removing irrelevant information, such as header or footer text. Metadata, including title of the document, language of the content, document identifier and source, were extracted from the original documents. Missing data fields were added to documents that lacked such information. The text files and metadata were used to create XML documents following the TREC [60]. In total, there were 13,627 documents. Table 1 shows the statistics of the corpus, including number of words and documents.

Table 1: Corpus Statistics. Chichewa corpus has the most number of documents.

Source	Number of Documents	Distinct Words	Word Total	Average Document Length
Chichewa	9,380	114,369	1,092,518	116
Citumbuka	2,258	58,390	459,789	203
Citonga	1,367	48,124	297,793	217
Cisena	449	19,161	159,660	355
Cinyanja	173	12,796	59,935	346
Summary	**13,627**	**252,840**	**2,069,695**	**151**

Table 2: Topics Statistics. Most of the translations used equivalent terms and the queries have similar properties.

Attribute	Citumbuka	Cinyanja	Chichewa
Number of topics	129	129	129
Maximum length	14	15	18
Minimum length	1	1	1
Average length	5	5	5
Average number of relevant documents	17	15	19

3.3 Topics

We recruited five assessors who were L_1 speakers of the investigated languages to formulate topics. Assessors came up with topics after browsing the collection using a Web based retrieval system that run Solr and used BM25 scoring on the back-end. Each assessor reviewed the top one hundred documents for topical relevance. A topic with at least five seen relevant documents was admitted to the list of topics. One hundred and twenty nine topics were formulated and were translated to Chichewa, Citumbuka and Cinyanja. In total, 387 topics were realised. Each topic was formatted in XML and had title, description, narrative and identifier fields [60]. Table 2 shows the properties of the queries and the average number of relevant documents judged per topic of each language.

3.4 Utility Annotations

Document utility assessments were done using a Web interface on top of Solr. One hundred (100) documents were assessed for each topic. Four different Solr scoring functions were used to increase the diversity of documents retrieved, namely: (i) BM25 on space delimited tokens, (ii) BM25 on 3 and 4 character n-grams, (iii) probabilistic model using Divergence from Randomness (DFR) [1] and (iv) language modelling based retrieval model using Bayesian smoothing with Dirichlet priors [63] with $\mu = 2000$. BM25 used standard parameter values – $k_1 = 1.2$ and $b = 0.75$. Monolingual speakers assessed the documents by providing graded utility labels between 0 and 3: : 0 Not relevant, 1 Marginally relevant, 2 Fairly relevant and 3 Highly relevant. The assessors provided utility labels by estimating the topical relevance of the document and the effort required to understand the document given their L_1 [62]. The assumption was that intelligibility would be implicitly incorporated in the user judgements. The assessors were asked to evaluate how each document would be of use given a query. Each topic was assessed twice and disagreements between assessments were resolved by involving a third assessor. In total, six assessors were recruited in the task of providing utility labels of the documents relative to the topics.

3.5 Features

Topical Relevance Features: For query–document features, topical relevance similarity scores between the document and query were used. BM25 was used as a feature for representing the matching degree of topical relevance between query terms and document terms [48]. We used standard values for BM25 parameters, i.e., $k_1 = 1.2$ and $b = 0.75$. We also included TF, IDF and TF-IDF scores as features. Additionally, we used language modelling with Jelinek-Mercer Smoothing scores. We calculated the features using untranslated queries based on tri-gram tokens and space delimited tokens from the document – words as they appear in the corpus.

Linguistic Intelligibility Features: Our linguistic intelligibility features were drawn from previous studies in linguistics [4, 26, 53] except for Kullback-Leibler Divergence and Jensen-Shannon Divergence. We proposed these two features as measures of similarity between the language model of the user's L_1, i.e., language of the query, and L_2 language model, i.e., any language that the search results may be written in. We assumed that the measure of how the distribution of one language is different/similar from the distribution of another related language would provide a linguistic distance estimate of the two languages. Our formulation focuses on lexical similarity, with the understanding that the lower the two distances the more similar the language models, and therefore, the more similar the two languages. Our language models were based on character tri-grams and were developed using Maximum Likelihood Estimation (MLE) with Laplace smoothing.

The linguistic intelligibility features were derived from wordlists and a small parallel corpora that were prepared for the experiments. The used word list was based on the Swadesh list [56]. The Swadesh list is a collection of about 200 concepts, which are deemed to be universal and culturally independent. Initially, we obtained the English version of the Swadesh list and asked two L_1 speakers of each of the five languages to translate the list to their L_1. The translations were done cooperatively to ensure that the translators agreed on the translations. These word lists were used to compute lexical distance [28], Levenshtein distance [28], surprisal [53] and conditional entropy [30]. Incompy[1] [41] was used to calculate Conditional Entropy, Surprisal and Levenshtein between pairs of wordlists. The parallel corpora consisted of text of about 800 to 1000 words. We used this data to calculate the following features: perplexity distance [20], Kullback-Leibler Divergence [32], Jensen-Shannon Divergence [35] and cosine similarity. We aligned five paragraphs of text from newspaper extracts to investigate positional correspondences of grammatical features in the investigated language pairs. This text was used to calculate indel, movement measure and tri-gram correlation.

Extra-linguistic Intelligibility Features: We conducted intelligibility experiments to obtain extra-linguistic features such as language contact and learning and frequency of use, and the intelligibility score of the participant for a given language. The experiments involved five languages namely: Cinyanja, Citumbuka, Chichewa, Citonga and Cisena. Announcements were sent using social media to recruit participants, and participants in this study did not take part in the document assessment task. One hundred and four (104) participants completed the test. The participants had different languages as their L_1, namely: Chichewa (50%), Citumbuka (20.6%), Cinyanja (9.8%) and Citonga (19.6%).

The experiments were done online and were divided into two phases. In the first phase of the experiment, participants provided their demographic information such as age and highest academic qualification, as well as their self-reported proficiency on the five languages, areas they have lived in their first ten years and the last ten years and attitudes towards the languages [24]. Finally, participants translated text from each of the languages to English. The translations were mapped to scores for each of the languages for each user. The translations were categorised into five intelligibility classes (0 to 4) depending on the quality of the translation, with 0 as not understanding anything in the document, 1 as marginally comprehensible – recognising a few words, 2 as fairly comprehensible – understanding some sections of the document, 3 comprehensible – understanding everything except a few words, and 4 being able to understand everything and providing a perfect translation.

The data obtained from the online tests was combined with linguistic intelligibility data to create data used in the feature selection phase, and the intelligibility test scores were used as a target variable. For our ranking task, only linguistic features were used as intelligibility features: including extra-linguistic features would require participants in the online intelligibility tests to assess the documents as well.

4 EXPLORING INTELLIGIBILITY

Several intelligibility features were extracted, including new features that have not been used in intelligibility studies before (see Table 3). The dataset consisted of intelligibility features both linguistic and extra-linguistic features and intelligibility test scores. This section provides a description of feature selection of intelligibility features and intelligibility prediction.

4.1 Feature Selection

We conducted feature selection to choose the optimal subset of features that would give the best intelligibility prediction accuracy for the four languages in our study. Cisena was not included in this task due to no L_1 participants in the intelligibility experiments. We combined the dataset of linguistic features with the dataset obtained from user Web intelligibility tests. Therefore, each user had four entries, one for each of the four languages: personal information such as gender, age and qualification; other extra-linguistic

[1] https://github.com/uds-lsv/incompy/blob/master/utils.py

Table 3: List of intelligibility and topical relevance features. The features marked with * are used as intelligibility variables for the first time.

Feature	**Description**
	Topical Relevance
TF	Term frequency in body and title.
IDF	Inverse Document Frequency in body and title.
BM25	BM25 in body and title.
Normalised BM25	Normalised BM25 value using Maximum and average values for the collection.
Normalised TF–IDF	TF–IDF value using Maximum and average values for the collection.
TF–IDF	TF–IDF in body and title.
LM	Query likelihood language model scores with Jelinek-Mercer Smoothing.
$\|D\|$	Length of body and title.
	Intelligibility Features
Levenshtein Distance (word)	Average distance measuring the number of operations required to transform a cognate in one language to a word in another language .
Levenshtein Distance (stem)	Average distance measuring the number of operations required to transform a stem cognate in one language to a stem in another language.
Conditional Entropy	The uncertainty or difficulty of mapping a word in a non-native language to a word in a native language.
Perplexity Distance	Measures how well a probability distribution of n-grams from a corpus predicts a model.
Cosine Similarity	Measures the similarity between two vectors of an inner product space of tri-grams of two documents.
Surprisal	Measure of uncertainty in a word being transformed to a cognate.
Kullback-Leibler Divergence *	Measure of how the distribution of tokens (words or n-grams) in one language is different from the distribution in another language – asymmetric.
Jensen–Shannon Divergence *	Measure of similarity between the distribution of tokens (words or n-grams) between two languages – symmetric.
Lexical Distance	The percentage of the number of words that are not cognates for any two given languages.
Movement Measure	Number of words that are moved when translating a sentence from one language to another closely related language.
Indel	Number of inserted or deleted words when translating text from one language to another language.
Tri-gram	Correlation of the number of frequencies of word tri-grams in a corpus of two languages.
	Extra-linguistic
Age	Age of participant. Values were transformed into four classes, i.e., 1 to 4.
Gender	Gender of participant. The data was transformed to binary values, i.e., 1 for male and 0 for female.
Qualification	Highest level of academic qualification. The values were transformed to integers, i.e., 1 to 4.
Contact	Representing whether the participant had contact with the language. Binary values represented whether the participant had previous contact with the language or not.
Attitude	Participant perception of the beauty of the language. Scores were on a scale of 1 to 5.
Familiarity	Represented the contact frequency of the language. Scores were on a scale of 0 to 4.
Learning	Represented whether participant had learnt the language before participating in the study.

features such as language contact information, frequency of use of the language and information indicating whether the participant learnt the language; and linguistic features such as entropy and perplexity.

Due to some of the features being highly correlated, we used a Random Forest (RF) trees based technique for feature selection, namely: permutation importance [6, 55]. Figure 2 shows the plot of permutation importance for our intelligibility dataset. The approach ranked *age*, *gender* and *qualification* lowly. We also used Boruta algorithm [33] to identify relevant features and the three features were marked as irrelevant features. Therefore, we excluded the three features from our further experimentation. Our newly introduced features, Jensen-Shannon Divergence and Kullback-Leibler, were rankly highly with the extra-linguistic features, i.e., Jensen-Shannon Divergence is the highest ranking linguistic feature. The feature ranking also shows that contact frequency as well as learning a language have more importance in predicting intelligibility. Overall, for the linguistic features, features based on the surface form of the words, were ranked more highly than syntactical features.

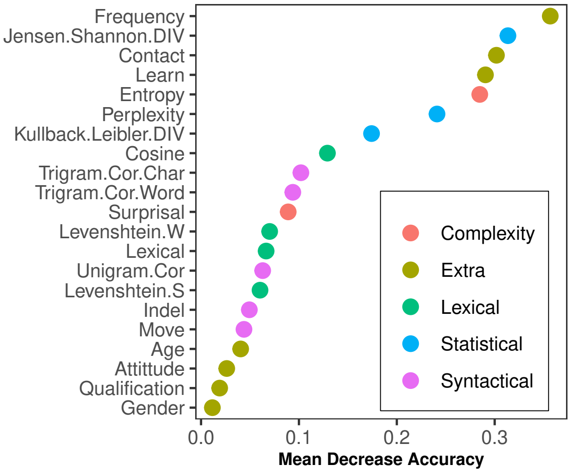


Figure 2: Plot of conditional importance and permutation importance for intelligibility features. Colours of the dots represent the type of the feature.

4.2 Predicting Intelligibility

We formulated the prediction of intelligibility as a multi-class classification problem – a prediction model uses our intelligibility features to predict intelligibility classes; each class as the intelligibility score in the text comprehension task (0,1,2,3,4). Unlike previous studies where the problem of predicting intelligibility was formulated as a regression problem, we categorised the intelligibility scores of our participants into four classes. These classes correspond to how much the participants would understand text written in another language and therefore, providing some expectation on how understandable the text in the different languages were to different L_1 speakers. The classification tasks were done using several algorithms but we report only on SVM and RF as they provided stable results. Figure 4 shows intelligibility prediction results produced by the two classifiers. The results show that using the features selected as relevant features in the dataset produced better results than using any of the other subsets such as using linguistic features only and extra linguistic features only. Overall, the performance is reasonable given that the dataset size was small. To examine the effect of the dataset size on performance, we investigated SVM and RF classifiers with different sizes of the dataset. Figure 3 shows the accuracy of the models using different sizes of the dataset. The graph shows that performance of the classifiers increased with the increased size of the dataset. This analysis provides insights on what features have higher predictive power for intelligibility and are therefore useful to improve ranking quality using topical relevance and intelligibility features.

5 RANKING FOR INTELLIGIBILITY

Ranking documents retrieved on the basis of topical relevance and intelligibility needs to optimize both relevance and intelligibility. Our ranking problem is that of constructing a ranking model that finds the best combination of relevance and intelligibility features that matches with user ranking preferences. In this section, we describe the experimental design and results of the evaluation of the proposed ranking models.

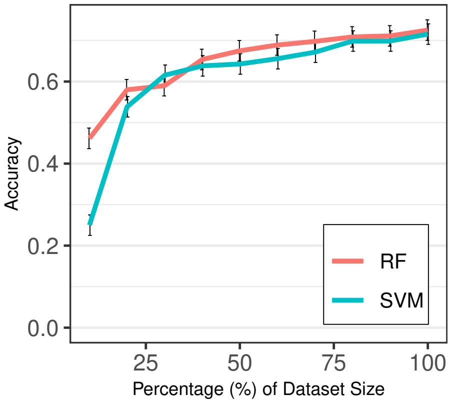


Figure 3: Accuracy of two classifiers, Support Vector Machine (SVM) and Random Forest (RF) at different dataset sizes.

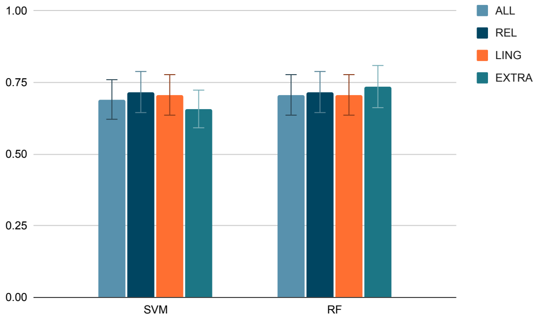


Figure 4: Accuracy plot for SVM and RF classifiers on different subsets of the data: all the features, relevant features only, extra-linguistic features only and linguistic features only.

5.1 Experimental Setting

Our study investigated how intelligibility can be incorporated in matching and ranking documents written in several languages to improve the quality of results. Learning To Rank (LTR) was used to train models that rank documents of related languages. We used LambdaMART for our supervised experiments because it has shown excellent performance in previous studies [8, 46, 58]. Our experimental set-up used LTR as follows: i) LTR with relevance features only, ii) LTR with all relevance features and a single intelligibility feature, and iii) LTR with all proposed features for relevance and intelligibility. We also explored unsupervised methods.

Weighted Sum: Previous studies with users on ranking preferences showed that relevance was used as a primary feature and intelligibility was used as a secondary criterion [14]. We explored using an unsupervised method for combining multiple objectives using weighted linear combination – aggregating normalised BM25 and cosine similarity scores between the language of the query and language of the document. The two features were first multiplied with selected weights, i.e., $f(x) = w_1 x_1 + w_2 x_2$. ROC is used when the rank order of the true weights is the only known information about the weights [2, 3]. We calculated the weights as follows [3]:

$$w_j(ROC) = \frac{1}{n}\sum_{k=j}^{n}\frac{1}{r_k}, \quad (1)$$

where n is the number of weights, r is the rank and j is the weight being calculated for the position j $j = 1, \ldots, n$.

Additional Baselines: Previous studies have proposed strategies for aggregating multilingual search results such as using raw and normalised relevance similarity scores [37]. We used min-max BM25 normalised scores as a baseline for the evaluation of the unsupervised ranking experiments.

5.2 Experimental Results and Analysis

We used LTR to learn a ranking function of documents as preferred by a monolingual user looking at documents in related languages, and hence, linguistic intelligibility features and topical relevance features described in Section 3 were used to train and test the models. We present results at model level and overall model performance based on five-fold cross validation.

Model Level Analysis: We provide an evaluation based on hold-out method through nDCG@10 [12, 49, 50]. The nDCG@10 scores and their comparisons are shown in Table 4. The results show that the weighted sum unsupervised model performed better than the other unsupervised models. However, the supervised model using topical relevance features performed better than the unsupervised methods. This is not surprising as supervised methods generally perform better than unsupervised methods. The models using one intelligibility feature and topical relevance features had mixed performance. The models using intelligibility features based on language models performed better than models using syntactic and lexis based features. The model using all the features – both topical relevance and intelligibility features – had slightly better performance.

We performed an ANOVA test to evaluate the omnibus null hypothesis that the 19 models are the same or equivalent based on nDCG@10. This was accepted with (($F(18, 1382) = 0.044$ $p < 0.05$). The system effect size for the ANOVA analysis was $Fhat^2 = 0.0007$ and the analysis achieved a power of 0.0675, indicating a very small effect size. The analysis shows that 1498 queries will be required to achieve a power of 0.8. These results are promising, even though not significant due to the 19 models being evaluated together – multiple testing is known to be conservative, especially for small data sets [5] – some of the results are significant for pairwise t-test at $p < 0.1$ (see Table 4.)).

We examined the differences between the performance of the model using topical relevance features only and the model using all the additional intelligibility features at topic level to understand the interplay between topical relevance and intelligibility in our dataset. We found that from the test topics, the performance of 18 topics remained the same, 26 topics improved, and 19 were worsened. Figure 5 shows the distribution of topics in terms of their performance differences. Further investigation of the improved topics showed that improvements due to the consideration of intelligibility features were achieved when a more distant language had higher topical relevance score (e.g. BM25) but irrelevant and if a more closely related language had a lower topical relevance score. This shows that weighting topical relevance and intelligibility in this case improves the quality of results.

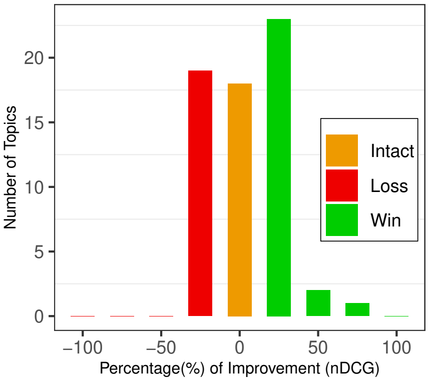


Figure 5: Plot of the number of queries that had their nDCG improved, hurt or remained the same after adding intelligibility features to the ranking models.

Overall Performance: We used a five-fold cross validation method for training and testing. Table 4 shows average nDCG results at different ranks for models using topical relevance only and models using topical relevance and intelligibility features. The reported values are averages of models based on the five-fold cross validation training data and test data. The trend in performance shows that using relevance and intelligibility features generally had a positive impact on nDCG at different ranks. The performance of the model using weighted sum is better than the model using normalised BM25 scores with an average difference in scores of 0.05. The supervised models follow the same trend. The model using topical relevance features and all intelligibility features performed better with an average difference of 0.01 across the ranks. The differences tend to decrease as the rank number increases, indicating that performance of the models converges as rank increases. The bigger differences in performance early in the ranks can improve the search experience of users using RSLs as they may find useful information early in the search session, therefore reducing their frustration.

6 DISCUSSION

Predicting intelligibility is a challenging problem – several factors that may determine intelligibility have been proposed in the literature, namely: linguistic and extra-linguistic features. We measured feature importance of our features using four Random Forest (RF) based feature selection algorithms. We have found that age, qualification, and gender were irrelevant features for intelligibility in our feature set. Our results show that extra-linguistic features such as learning the language, contact with the language, and frequency of use have high predictive power for intelligibility. Features based on measuring character distribution in words or corpus were ranked higher than any of the other classes of features. Lexical features performed fairly well. However, syntactical features had the worst performance. These results are similar to previous studies for intelligibility prediction for Indo-European languages using regression

		Overall nDCG @cutoff					@nDCG10 at Model Level	
Type	**Model**	**1**	**3**	**5**	**10**	**50**	**10**	**% increment**
Unsupervised	BM25	0.3434	0.3659	0.3937	0.4618	0.6101	0.7315 †	▽ 15.74%
	Normalised BM25	0.376	0.412	0.438	0.487	0.6317	0.6938†	▽ 22.2%
	Weighted Sum	0.453	0.4651	0.486	0.5359	0.6535	0.7999 †	▽ 5.83%
Baseline	Relevance Only	0.5511	0.5362	0.5601	0.6011	0.7755	0.8466	
Lexical	Cosine	0.5371	0.5591	0.5753	0.6095	0.7818	0.8342	▽ 1.47%
	Lexical Distance	0.5174	0.529	0.5725	0.6098	0.7817	0.8537	△ 0.84%
	Levenshtein(s)	0.5481	0.5286	0.5763	0.6129	0.776	0.8629	△ 1.93%
	Levenshtein(w)	0.5263	0.5362	**0.5838**	0.6071	0.7799	0.8485	△ 0.23%
Complexity	Entropy	0.561	0.5435	0.5679	0.6159	0.7794	0.8566	△ 1.19%
	KL Divergence	0.5611	0.5585	0.5755	0.6013	0.7837	0.8482	△ 0.19%
	Perplexity	0.5525	0.5498	0.5649	0.6096	0.7825	0.8673†	△ 2.45%
	SL Divergence	**0.5991**	0.5549	0.5817	0.6016	0.7833	0.8507	△ 0.49%
	Surprisal	0.5116	0.5442	0.5746	0.5996	0.7862	0.857	△ 1.24%
Syntactic	Indel	0.5136	0.5478	0.5638	0.617	0.7924	0.8505	△ 0.46%
	Move	0.5506	**0.5599**	0.5781	0.6058	0.7849	0.8326	▽ 1.65%
	Charactergram	0.5635	0.5451	0.5766	0.6132	**0.7902**	0.8651†	△ 2.18%
	Wordgram	0.5541	0.5423	0.5679	**0.6206**	0.779	0.8461	▽ 0.05%
	Wordtrigram	0.5442	0.5726	0.5728	0.6046	0.7796	0.8512	△ 0.55%
All	Final	0.5721	0.5571	0.5685	0.6116	0.7819	0.8676†	△ 2.49%

Table 4: Average nDCG scores at different ranks for models using five fold cross validation, and model comparison of performance of NDCG@10. Scores with † are significant for paired t-test at p < 0.1

[26, 34]. Our results on feature selection suggest that cognacy may be a significant predictor of intelligibility among closely related languages. Our intelligibility classification results are promising, and we have shown that it is possible to predict intelligibility automatically from linguistic features. Using Random Forest classifiers provided good prediction accuracy. However, the imbalances in terms of intelligibility classes in the dataset affected prediction performance at class level. Our analysis of the effect of dataset size on performance suggests that, with more data, it might be possible in the future to obtain better improved results.

We have proposed a way to improve the quality of search results for resource-constrained languages by re-ranking results using topical relevance and intelligibility criteria. We extracted features from documents and queries to estimate the similarity relationship between the document language and that of the query and to estimate topical similarity between the query and the document. We trained and tested LTR models using these features. We also used normalised BM25 scores and weighted cosine similarity and normalised BM25 scores. Our evaluation of the models shows slight performance improvements in terms of nDCG. Models using topical relevance features and our proposed metrics, Jensen-Shannon Divergence and Kullback-Leibler, are among the top performing ranking models at all retrieval cut-points considered. The small improvements seen so far provide some evidence that integrating intelligibility in re-ranking of search results written in related languages can improve retrieval effectiveness. Although our results are promising, using document utility judgments as proxies for ranking preference may have affected the results. While using this approach has been effective in other studies [45], document assessment using preference judgments could be more successful [11].

7 CONCLUSION AND FUTURE WORK

We have shown how integrating intelligibility features with topical relevance features can provide a signal for document utility for queries with no or limited relevant documents. Our results on improving retrieval quality through re-ranking are promising. Intelligibility is a difficult attribute to model effectively – both linguistic (e.g., vocabulary, morphology, phonology and phonetics, and syntax) and extra-linguistic (e.g., a speaker's prior language knowledge and experience, and perceptions) factors affect intelligibility. We have shown that it is possible to automatically classify intelligibility with a good accuracy. As a first study to explore re-ranking using intelligibility, there are several opportunities to be explored further. Firstly, our approach assumed a uniform intelligibility score across monolingual individuals speaking the same language as L_1 – we used language features to estimate how intelligible two languages would be without speakers having any prior knowledge of the other language. Future work will add personalization to adapt results to user language knowledge and preference. Secondly, we relied on a predefined relationship between languages: we extracted features of the involved languages to estimate how intelligible the languages were. A more dynamic approach would be to retrieve documents based on intelligibility calculated dynamically without prior knowledge of the languages involved. Studies on retrieval by lexical similarity using deep learning methods on well-resourced languages are needed to provide techniques for such dynamism, and to understand the interplay between topical relevance and intelligibility, which can provide insights on retrieval for low resourced languages.

REFERENCES

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002.

[2] F.Hutton Barron and Bruce E. Barrett. The efficacy of smarter — simple multi-attribute rating technique extended to ranking. *Acta Psychologica*, 93(1):23 – 36, 1996. Contributions to Decision Making II.

[3] Hutton Barron and Bruce Barrett. Decision quality using ranked attribute weights. *Management Science*, 42(11):1515–1523, 1996.

[4] Robert Bayley, Richard Cameron, Ceil Lucas, and Charlotte Gooskens. Experimental methods for measuring intelligibility of closely related language varieties, 01 2013.

[5] Leonid Boytsov, Anna Belova, and Peter Westfall. Deciding on an adjustment for multiplicity in ir experiments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 403–412, New York, NY, USA, 2013. Association for Computing Machinery.

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA, 1984.

[7] 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.

[8] Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research, 2010.

[9] Lyle Campbell. *Borrowing.* Edinburgh University Press, ned - new edition, 3 edition, 2013.

[10] 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.

[11] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval*, pages 16–27, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[12] Benjamin A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. Inf. Syst.*, 30(1), March 2012.

[13] 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.

[14] Catherine Chavula and Hussein Suleman. Intercomprehension in retrieval: User perspectives on six related scarce resource languages. In Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska, editors, *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*, pages 263–272. ACM, 2020.

[15] Jean Josephine Chavula. *Verbal Derivation and Valency in Citumbuka.* PhD thesis, Centre for Linguistics, Leiden University, 2016.

[16] 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.

[17] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 403–412, New York, NY, USA, 2011. Association for Computing Machinery.

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

[19] 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.

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

[21] 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.

[22] Michael Golebiewski and danah boyd. Data voids: Where missing data can easily be exploited. Technical report, Data & Society, May 2018.

[23] 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.

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

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

[26] Charlotte Gooskens and Femke Swarte. Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics*, 40(2):123–147, 2017.

[27] 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.

[28] 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.

[29] Anni Järvelin and fi Sanna Kumpulainen. Dictionary-independent translation in clir between closely related languages. 2006.

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

[31] 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.

[32] Solomon Kullback. *Information Theory and Statistics.* John Wiley and Sons, New York, 1959.

[33] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[34] Sebastian Kürschner, Charlotte Gooskens, and Renée van Bezooijen. Linguistic determinants of the intelligibility of swedish words among danes. *International Journal of Humanities and Arts Computing*, 2(1–2):83–100, Sept 2009.

[35] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, September 2006.

[36] Jouni Maho. *A Classification of the Bantu Languages: An Update of Guthrie's Referential System.* Routledge Language Family Series. Taylor & Francis, 2006.

[37] Fernando Martínez-Santiago, L. Alfonso Ureña-López, and Maite Martín-Valdivia. A merging strategy proposal: The 2-step retrieval status value method. *Information Retrieval*, 9(1):71–93, Jan 2006.

[38] Sam Mchombo. *The Syntax of Chichewa.* Cambridge Syntax Guides. Cambridge University Press, 2004.

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

[40] Winfred Mkochi. The morphosyntactic status of zamu- and ku- in malawian tonga. *South African Journal of African Languages*, 38(3):337–342, 2018.

[41] Marius Mosbach, Irina Stenger, Tania Avgustinova, and Dietrich Klakow. incom.py - a toolbox for calculating linguistic distances and asymmetries between related languages. In *Proceedings of Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, 2-4 September 2019*, pages 811–819, 2019.

[42] Mohammed Mustafa and Hussein Suleman. Multilingual Querying. In *Proceedings of the Arabic Language Technology International Conference (ALTIC), Alexandria, Egypt*, 2011.

[43] Daniel Nettle. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, 17(4):354–374, 1998.

[44] Derek Nurse and Gerard Philippson. *The Bantu Languages.* Routledge Language Family Series. Taylor & Francis, 2006.

[45] Joao 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 '16, pages 965–968, New York, NY, USA, 2016. ACM.

[46] Tiziano Papini and Michelangelo Diligenti. Learning-to-rank with prior knowledge as global constraints. In *Workshop on Combining Constraint solving with Mining and Learning (CoCoMiLe)*, 2012.

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

[48] Stephen E. Robertson and Stephen Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[49] Tetsuya Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, June 2014.

[50] Tetsuya Sakai. Statistical significance, power, and sample sizes: A systematic review of sigir and tois, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 5–14, New York, NY, USA, 2016. Association for Computing Machinery.

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

[52] Ben Steichen, Carla Castillo, and Kevin Scroggins. Personalized multilingual search - predicting search result list language preferences. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 343–347, New York, NY, USA, 2020. Association for Computing Machinery.

[53] 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):175–199, 2017.

[54] Stephanie Strassel and Jennifer Tracey. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[55] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307), 2008.

[56] Morris Swadesh. Lexico-statistic dating of prehistoric ethnic contacts. with special reference to north american indians and eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463, 1952.

[57] Francisca Hendrika Euphemia Swarte. *Predicting the mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors*. PhD thesis, University of Groningen, 2016.

[58] Niek Tax, Sander Bockting, and Djoerd Hiemstra. A cross-benchmark comparison of 87 learning to rank methods. *Inf. Process. Manage.*, 51(6):757–772, November 2015.

[59] John R Taylor and Anthony P. Grant. Lexical borrowing, 03 2014.

[60] Ellen M Voorhees and Donna K Harman. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press, 2005. Cambridge MA.

[61] 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.

[62] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 91–100, New York, NY, USA, 2014. Association for Computing Machinery.

[63] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.

[64] 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.

[65] 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.