

# **Evaluation of Oromo-English Cross-Language Information Retrieval**

by

Kula Kekeba Tune, Vasudeva Varma, Prasad Pingali

in

*International Joint Conference on Artificial Intelligence (IJCAI)-2007, January 12, 2007, Hyderabad, India*

Report No: IIIT/TR/2008/80



Centre for Search and Information Extraction Lab  
International Institute of Information Technology  
Hyderabad - 500 032, INDIA  
June 2008

# Evaluation of Oromo-English Cross-Language Information Retrieval

**Kula Kekeba Tune, Vasudeva Varma and Prasad Pingali**

Language Technologies Research Centre

IIIT, Hyderabad

India.

kuulaa@gmail.com, vv@iiit.ac.in, pvvpr@iiit.ac.in

## Abstract

This paper reports on the first Oromo-English CLIR system that is based on dictionary-based query translation techniques. The basic objective of the study is to design and develop an Oromo-English CLIR system with a view to enable Afaan Oromo speakers to access and retrieve the vast on-line information sources that are available in English by using their own (native) language queries. We describe the major approaches and procedures that have been used in designing and developing Oromo-English CLIR system together with the information retrieval evaluation experiments recently conducted at CLEF 2006 ad hoc track. The purpose of the current initial evaluation experiments was to assess the over all performance of the Oromo-English CLIR system by using different fields of Afaan Oromo topics. Thus we submitted three official runs (experiments) that differed in terms of utilized fields in the topic set, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN) to CLEF 2006. Since appropriate online language processing resources and information retrieval tools are not available for Afaan Oromo, only limited linguistic resources such as Oromo-English dictionary and Afaan Oromo stemmer that had been designed and developed at our research center were used in conducting the experiments. Yet we found the performances of our evaluation experiments at CLEF 2006 very encouraging which is a good news for development and application of CLIR systems in other indigenous African languages.

## 1 Introduction

Cross-language information retrieval (CLIR) can briefly be defined as a subfield of information retrieval system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query. Facilitating the process of finding relevant documents written in one natural language with automated systems that can accept queries expressed in other language(s) is thus

the major purpose of CLIR system. The process is bilingual when dealing with a language pair, i.e., one source language (e.g., Afaan Oromo) and one target or document language (e.g., English). In multilingual information retrieval the target collection is multilingual, and topics are expressed in one language [Hedlund *et al.*, 2004]. In any of such cases CLIR is expected to support queries in one language with a collection in another language(s) [Allan *et al.*, 2003; Waterhouse, 2003].

According to [Peters and Sheridan, 2001] CLIR is a complex multidisciplinary research area in which methodologies and tools developed in the field of Information Retrieval (IR) and natural language processing converge. Information retrieval is traditionally based on matching the words of a query with the words of document collections. Because the query and the document collection are in different languages, this kind of direct matching is impossible in CLIR. Translation is needed: either the query has to be translated into the language of the documents or the documents have to be translated into the language of the query. Obviously, translating the whole document collection is more demanding, as it requires more scarce resources like full-fledged MT system which is not available for a number of languages in developing countries. Hence query translation techniques become more feasible and common in development and implementation of CLIR system. The basic methods in query translation are machine translation, corpus-based translation and knowledge-based or dictionary-based translation [Lehtokangas *et al.*, 2004]. Since established MT system and/or parallel corpora are not available for Afaan Oromo, we have designed and developed a CLIR system that is based on machine-readable dictionary.

In this paper we discuss the basic techniques and procedures we have adopted and used for designing and developing Oromo-English CLIR together with its performance assessment experimental results at Cross Language Evaluation Forum (CLEF 2006). We had made our debut participation in the ad hoc track of CLEF from Language Technologies Research Center (LTRC) of IIIT-Hyderabad, India. The major purpose of our participation in the CLEF campaign was to obtain hands-on experience in standard evaluations of cross-language information retrieval forum by conducting experiments in three language pairs, i.e. Oromo-English, Hindi-English and Telugu-English. A sets of 50 English topics that are supposed to be representatives of the information needs

of users had been prepared by CLEF and used in each of the three languages in order to search a large size (more than 169,000 items) databases of English documents. This paper focuses on description of OromoEnglish CLIR system and its evaluation experiments at CLEF 2006.

The rest of this paper is organized as follows. In the next sections we present the major motivations of the study. This is followed by an overview of Afaan Oromo in section 3. Section 4 reviews few of the available related works from the perspectives of African indigenous languages including Ethiopian languages. Section 5 describes the experimental setup of our CLIR system including the basic procedures and approaches that have been adopted for official runs of CLEF 2006. Section 6 presents summary of the results while section 7 provides our general concluding remarks.

## 2 Motivation and Contribution of Study

The amount of accessible electronic information has exploded in recent years thanks to the Internet and other related distributed international networks. Due to the rapidly expanding use of the Internet for communication and dissemination of information through out the world, electronic information sources are now available in an ever-increasing number of languages. Users of such globally distributed networks (including digital libraries and World Wide Web) need to be able to access and retrieve any relevant information in whatever language and form it may have been recorded and stored [Peters and Sheridan, 2001]. Considering the limitations of the existing monolingual IR systems and the need for more research efforts for development of CLIR [Oard, 1997] has noted that we are rapidly constructing an extensive network infrastructures for moving information across national boundaries, but much remains to be done before linguistic barriers can be surmounted as effectively as geographic ones.

Indeed, [Lehtokangas *et al.*, 2004] the more languages there are on the Internet, the more there are language barriers to be crossed. Thus it is understandable that CLIR has become an important area in both research and practice. Consequently, [Peters and Sheridan, 2001] much attention has been given over the past few years to the study and development of tools and technologies for CLIR and Multi-Lingual Information Access (MLIA). To this end, various national and international CLIR evaluation campaigns and research forums have been initiated and conducted in different parts of the world including USA, Europe and Asia. Chief among such organizations and institutions that are strongly supporting CLIR research projects are TREC (in USA), CLEF (in Europe) and NTCIR (in Asia).

In spite of such important recent progresses and developments of CLIR studies, the task of investigating and developing CLIR system for indigenous African languages including Afaan Oromo has been left unaddressed over the past few years. This is due to the fact that most of the CLIR researchers are highly concerned with designing and evaluation of their own CLIR systems in western languages. Considering the need for initiating and undertaking additional CLIR researches in different languages including minority languages like Afaan Oromo [Gey *et al.*, 2005] said we find

that insufficient attention has been given to the Web as a resource for multilingual research, and to languages which are spoken by hundreds of millions of people in the world but have been mainly neglected by the CLIR research community. As indicated above, CLIR system can play a vital role in addressing and resolving the problem of language barrier in accessing online multilingual information sources. Thus, this study is mainly motivated by and aimed at enabling online information sources accessible to more than 25 million Afaan Oromo speakers who are otherwise unable to do so due to lack of English understanding.

Few of the major contribution of this study include:

- Designing and implementation of dictionary-based CLIR system for indigenous language like Afaan Oromo
- Construction and adaptation of basic Afaan Oromo IR tools such as bilingual dictionary, stemmer and stop-word list a and their applications for Oromo-English CLIR
- Analysis and review of CLIR research works related indigenous African languages and sharing of their experiences
- Testing the performance of Oromo-English CLIR system at standard and internationally recognized evaluation forum like CLEF
- Demonstrating the feasibility CLIR application for non-western and resource scarce language like Afaan Oromo

## 3 A Brief Overview of Afaan Oromo

Oromo (also known as Afaan Oromo) is one of the major Languages that are widely spoken and used in Ethiopia [Nefa, 1988]. Currently it is an official language of Oromia state (which is the largest region in Ethiopia). Unlike Amharic, (an official language of Ethiopia) which belongs to Semitic family languages, Afaan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum [Yimam, 1986; Nefa, 1988]. In this Cushitic branch of the Afro-asiatic language family Afaan Oromo is considered as one of the most extensive languages among the forty or so Cushitic languages [Stroemer, 1987]. It is a common mother tongue for Oromo people, who are the largest ethnic group in Ethiopia, at 32.1% of the population according to the 1994 census. Although it is difficult to identify the actual number of Afaan Oromo speakers (as a mother tongue) due to lack appropriate and current information sources, according to some earlier general information sources it is estimated that Afaan Oromo is being spoken by more than 25 million Oromos within Ethiopia.

With regard to the writing system, Qubee (Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991. Currently, Afaan Oromo is widely used as both written and spoken language in Ethiopia and some neighboring countries, including Kenya and Somalia. Besides being an official language of Oromia State, Afaan Oromo is the instructional medium for primary and junior

secondary schools throughout the region and its administrative zones. Moreover, a number of literature works, newspapers, magazines, education resources, official documents and religious writings are written and published in Afaan Oromo.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very rich morphology. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo, Amharic and Zulu most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afaan Oromo nouns and adjectives are highly inflected for number and gender. For instance, [Oromoo, 1995] in comparison to the English plural marker *s* (-es), there are more than 12 major and very common plural markers in Afaan Oromo nouns (e.g. oota, ooli, -wwan, -lee, -an, een, -oo, etc.). Afaan Oromo verbs are also highly inflected for gender, person, number and tenses. Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo.

Since Afaan Oromo is morphologically very productive, derivations and word formations in the language involve a number of different linguistic features including affixation, reduplication and compounding [Oromoo, 1995]. Obviously, these high inflectional forms and extensive derivational features of the language are presenting various challenges for text processing and information retrieval experiments in Afaan Oromo.

## 4 Related Work

Very limited works have been done in the past in the areas of IR and CLIR in relation to African indigenous languages including major Ethiopian languages. A case study for Zulu (one of the major languages in South Africa) was reported by [Cosijn *et al.*, 2002] in relation to application for cross-lingual information access to indigenous knowledge databases. A dictionary-based Zulu-English CLIR which was supplemented by approximate string matching technique was attempted to be implemented for this purpose. The researchers had indicated that the disparate vocabularies of Zulu and English were one of the main causes for the poor results that they had achieved in their experiments. Another similar study was undertaken by [Cosijn *et al.*, 2004] on Afrikaans-English cross-language information retrieval. The main components of this CLIR were source and target language normalizers, a translation dictionary, source and target language stopword lists and an approximate string matching. A dictionary-based query translation technique was used to translate Afrikaans queries into English queries in similar manner to our approach. By using a combination of bilingual dictionary, Afrikaans morphological analyzer and stopword lists, the source language topics were translated into target language queries and matched against the English corpus database. The performance of our CLIR system was evaluated by using 35 topics from the CLEF 2001 English test collection (employing title and descriptions fields of the topic sets) and achieved an average precision of 19.4%.

More recently, different dictionary-based Amharic-English and Amharic-French CLIR experiments were con-

ducted at a series of CLEF ad hoc tracks [Alemu *et al.*, 2004; 2005; 2006]. The first Amharic-English information retrieval was conducted at CLEF 2004 with emphasis on analyzing the impact of the stopword lists of the source and target languages. While a bilingual dictionary was used to translate Amharic queries to English bags-of-words unmatched (untranslatable) source language terms were translated and added into the dictionary manually. The result of the evaluation showed that the experiment that had employed English stopword list had performed better compared to the other its counter part experiments. Another similar dictionary-based Amharic-English CLIR experiment was conducted at CLEF 2006 by employing Amharic morphological analyzer and part of speech tagging to facilitate more accurate query translation. Out of dictionary terms were handled by using fuzzy matching techniques and Lucene search engine was used for indexing and retrieval of English documents. The mean average precision that was obtained in this evaluation (i.e. 22.78%) has been reported as a better retrieval performance for Amharic CLIR compared to the experiments that had been undertaken in the previous two years at CLEF [Alemu *et al.*, 2006]. On the other hand, to the best of the current researchers knowledge no formal study has been undertaken and found in relation to Afaan Oromo CLIR or a number of other Ethiopian Languages except Amharic. Thus, this is an initial research work that is intended to investigate and develop an Oromo-English CLIR system.

## 5 Dictionary based CLIR

As mentioned earlier, we have used a dictionary-based CLIR method that is similar to some of the experiments that have been carried out at the previous CLEF campaigns [Alemu *et al.*, 2004; Marin and Pascale, 2001]. Oromo-English dictionary [Tilahun, 1989] which was adopted and developed from hard copies of human readable bilingual dictionaries by using OCR technology is used to translate Afaan Oromo topics into bags-of-words English queries. Though this is one of the most popular and widely used Oromo-English dictionaries we had made some effort to enhance it by incorporating additional entries from other bilingual dictionaries and related reference sources. To make the dictionary error free and machine readable, editing and formatting of the dictionary were carried out manually. Currently this bilingual dictionary consists of about 12,800 Oromo entries including base words and their derivational variants.

### 5.1 Stopword Lists and Stemming

In order to define Oromo stopwords, we first generated and created a list of the top 350 most frequent words found in 1.2 million words of Afaan Oromo text corpus by using IDF measures. Then we incorporated additional pronouns, conjunctions, prepositions and other similar functional words in Afaan Oromo. Accordingly, we have obtained and used about 580 stopwords of Afaan Oromo in conducting our experiments. First, the number of words in Afaan Oromo topics were reduced by removing the stopwords. Once the stopwords were removed from the topics, we applied a light stemming algorithm in order to conflate word variants into the

same stem or root. Similarly, after query topics are translated into English the translated queries were also filtered and stopped through the English stopwords analyzer that was implemented in the Lucene search engine. Lucene, which is an open source text search engine that is mainly based on vector space model was adopted and used for indexing and retrieval of the English documents.

A number of previous works on CLIR have indicated the fact that languages that are morphologically rich can benefit from morphological analysis such as stemming and lemmatization [Marin and Pascale, 2001]. Since Afaan Oromo is one of such morphologically rich languages and stemming is often language dependent, we have developed a rule based suffix-stripping algorithms focusing on very common inflectional suffixes of Oromo language. This light stemmer is designed to automatically remove frequent inflectional suffixes attached to headwords (base-word forms) of Afaan Oromo. Some of the common suffixes that have been considered in our current light stemmer include gender (masculine, feminine), number (singular or plural), case (nominative, dative), possession morphemes and other related morphological features in Afaan Oromo.

Broadly it is possible to categorize suffixes in Afaan Oromo into three basic groups: derivational, inflectional, and attached suffixes. Attached suffixes are particles or postpositions like *arra*, *-bira*, *-irra*, *-itti*, *-dha*, *-f*, etc. that are attached to stem/root words. For instance the word “*adunyaarratti*” (in the world) is formed from a stem, i.e. *adunyaa* and two attachment suffixes, i.e. *-irra* + *-itti*.

Inflectional suffixes are a combination of word stem with grammatical/syntactic morphemes, usually resulting in a word of the same class as the original stem. These suffixes include plural noun markers such as *oota*, (e.g. *nama* + *-oota* = *namoota* i.e. person + *-s* = persons); *-lee* (e.g. *jabbi* + *-lee* = *jabbiilee*, i.e. calf + *-es* = calves) and *-wwan* (e.g. *indaaqqo* + *-wwan* = *indaaqqowwan*, i.e. chicken + *-s* = chickens). Inflectional suffixes for other forms of nouns as well as adjectives and verbs are affixed to the stem words in similar manners though they are sometimes more complicated by requiring certain modifications in the stem.

Derivational suffixes enable a new word, often with a different grammatical category, to be built from stem/root other words. For example, the stem verb *qabuu* + *-eenyaa* becomes *qabeenyaa* which is noun while the adjective *gowwa* + *-ummaa* becomes *gowwummaa*, which is also another Oromo noun.

Based on our current observations the most common order/sequence of Afaan Oromo suffixes (right to left) is derivational, inflectional and attached suffixes. Thus, the stemmer is expected to remove from the right end first all the possible attached suffixes, then inflectional suffixes and finally derivational suffixes. To facilitate this task, we have identified and built three different suffixes clusters/lists with respect to the above three major types of suffixes in Afaan Oromo. The identification and classification of these various Afaan Oromo suffixes were based the linguistic studies and grammatical reference sources that we have consulted and investigated earlier as part of our CLIR project.

Our Afaan Oromo stemmer first starts with consulting the

attached suffix list and try to remove all possible such suffixes if any. Then it consults the inflectional suffix list and strips any inflectional suffixes. Then the Oromo-English dictionary is consulted and the translations of all matching entries are returned. In case there is no matching entry in the bilingual dictionary, the stemmer would try to identify and remove any of the most common derivational suffixes for final dictionary look up. The following simple stemming example illustrates some of the major steps described above.

Stemming example for “*sootroowwanitti*.” (i. e. by the drugs).

- *Step-1.* The stemmer consults the attached suffix list and tries to find any longest matching suffix entry in the list with the right end substrings of the given word. Accordingly, it first identifies “*-itti*” and strip it from “*sootroowwanitti*” which then becomes “*sootroowwan*.” If there are more one attachment suffixes, they are all removed by repeating the same procedure.
- *Step-2.* Next the stemmer consults the inflectional suffix list and tries to find any longest suffix entry that matches with the right end substrings of the given word, i.e. “*sootroowwan*.” Accordingly, it then identifies “*-wwan*” and strip it from “*sootroowwanitti*” which then becomes “*sootroo*.” If there are more one inflectional suffixes (which is rare unlike attachment suffixes in Afaan Oromo), they are all removed by repeating the same procedure.
- *Step-3.* The dictionary is searched to find remaining stem (base word), i.e. *sootroo* and all possible translation of its matching entries are returned and passed to the search engine.

## 5.2 Query Translation

Topics in the CLEF ad hoc track are structured statements representing information needs; the systems use the topics to derive their queries. Each topic consists of three parts: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. Sets of 50 topics were created for the CLEF 2006 ad hoc mono- and bilingual tasks. Below we give an example of Afaan Oromo topic from CLEF 2006.

```
<top>
<num> C308 </num>
<OM-title> Gaaddiddeeffamuu Aduu </OM-title>
<OM-desc> Dokumantoota guutumaan gaaddiddeeffamuu ykn cinaan gaaddiddeeffamuu aduu gabaasan barbaadi. </OM-desc>
<OM-narr> Dokumantootni gaaddiddeeffamuu aduu irratti odeeffannoo kamiyyuu kennan fudhatama ni qabu. Dokumantootni waayee gaaddiddeeffamuu baatii ykn sosochiiwwan pilaaneetootaa ibsan as keessa hin galan. </OM-narr>
</top>
```

Topics can be converted into queries that an IR system can execute in many different ways. Both manual and automatic query constructions approaches are possible. Automatic query construction is used in our experiments. In

our case, the original CLEF topic set for English were initially manually translated into Oromo topics by a group of translators who are native speakers of Afaan Oromo. We then automatically translated the resulting Oromo topics back into English queries using Oromo-English dictionary that was adopted and developed from human readable printed bilingual dictionaries by using OCR technology [Tilahun, 1989].

After eliminating of the stopwords by consulting the stopword list, the rest content words of Oromo topics were stemmed by using our light stemmer. Then the stemmed keywords words of Afaan Oromo topics were automatically looked up for all possible translations in the bilingual dictionary. Therefore, the resulting English queries were bags-of-words, taking into account all possible translation of keywords found in the bilingual dictionary. One of the major problems in this translation process was related to handling out of dictionary or unmatched words which are about 120. While most of these words were proper names, few of them were foreign and borrowed words. Since the transliteration of some of the unmatched proper names or foreign borrowed words have similar spelling with the corresponding English words (e.g. Iraaq, Kurdi, Buush, filmi) they have been directly copied and added to the English query for approximate string matching by the search engine. The transliteration of the rest more complex proper names (about 68) was manually analyzed and added in a separate dictionary for automatic consultation by the retrieval system.

## 6 Evaluation

### 6.1 Experimental Setup

All of our Oromo-English information retrieval experiments have been carried out by using standard CLIR evaluation resources provided by CLEF for ad hoc track bilingual tasks. The main objective of CLEF is to promote research in the field of multilingual system development [Peters, 2006]. According to [Nunzio *et al.*, 2006] the ad hoc retrieval track is generally considered to be the core track in the CLEF. The purpose of the track is to promote the development of monolingual and cross-language textual document retrieval systems. The ad hoc track adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s. The test collection used consists of a set of topics describing information needs and a collection of documents to be searched to find those documents that satisfy these information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures.

In all cases the basic task in the ad hoc track is to retrieve relevant documents from the chosen target collection and submit the results in a ranked list. For our Oromo-English CLIR we used the English test collection (provided by CLEF) from two newspapers, the Glasgow Herald 95 and the Los Angeles Times 94. Table 1 summarizes the numbers and sizes of the English test collection that we have used for our experiments.

Collection	Number of Docs	Collection Size (MB)
LA Times - 94	113,005	425
GH Herald - 95	56,472	154
Total	169,477	579

Table 1: Summary of English Test Collection

### 6.2 Experimental Results

In this section we describe the results of our experiments. As mentioned earlier, we had conducted and submitted three official runs (experiments) that differed in terms of utilized fields in the topic set. We use a prefix ‘OM’ to mean the query language is ‘Oromo’, and ‘T’ for including title, ‘D’ for including description and ‘N’ for including the narration. Therefore we have named our run-ids as OMT, OMTD and OMTDN, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN) for Oromo-English bilingual task in the ad-hoc track of CLEF 2006. The Mean Average Precision (MAP) scores for our three runs are shown in Table 2. The total number of relevant documents (Relevant-tot.), the retrieved relevant documents (Rel.Ret.), and the non-interpolated average precision (R-Precision) are also summarized and presented in Table 2. Like wise, Table 3 shows summary of Recall-Precision results for the three runs.

Run-label	Relevant-tot.	Rel. Ret.	MAP	R-Prec
OMT	1,258	870	22.00%	24.33%
OMTD	1,258	848	25.04%	26.24%
OMTDN	1,258	892	24.50%	25.72%

Table 2: Summary of average results for the three runs

Recall	OMT	OMTD	OMTDN
0%	48.73%	58.01%	59.50%
10%	39.93%	47.75%	46.45%
20%	34.94%	42.33%	37.77%
30%	30.05%	32.15%	31.17%
40%	26.41%	28.55%	28.27%
50%	22.98%	24.90%	24.72%
60%	18.27%	20.19%	19.40%
70%	15.10%	16.59%	15.61%
80%	11.76%	12.87%	12.70%
90%	8.58%	8.37%	8.56%
100%	6.56%	6.05%	6.58%

Table 3: Recall-Precision scores for the three runs

Table 2 reveals OMTD (title and description) run and OMTDN (title, description and narration) run have achieved almost the same level of performance (with about MAP of 25 %). The title run has slightly lower performance with MAP of 22%. We feel this is due to the fact that most of the title fields in Afaan Oromo topics were very short.

## 7 Concluding Remarks and Future Work

In this paper, the basic approaches and evaluation results of Oromo-English CLIR has briefly described and reported. We

attempted to show how very limited language resources can be used in designing and implementation of a CLIR system for resource scarce languages like Afaan Oromo. We obtained significant average precision results for all of the three official runs, given the limited CLIR resources we have used in our experiments. The performance of the OMTD run was found to be better than the other two runs achieving a mean average precision of 25.04%. The title run has slightly lower performance with mean average precision of 22%. We feel this is due to the fact that most of the title fields in CLEF topics were very short. In general, the evaluation results of our CLIR system are very encouraging for development and application of the CLIR system in a number of other indigenous African languages (including other Ethiopian languages).

Obviously, there is much room for improvement of the performance of our CLIR system. Various important research problems have been emerged in the course of the evaluation of the performances of the experiments, e.g. the need for handling of out of dictionary terms including proper names or borrowed foreign words and phrasal terms of Afaan Oromo. Currently we are working on evaluation of different components of the CLIR system including the impacts of stopword lists and light stemmer of Afaan Oromo. Automatic query expansion by using Pseudo-Relevance Feedback (PRF), query structuring, proper names handling and application of a crude disambiguation methods are some the tasks that we will consider in our future experiments. The task of phrasal terms identification and compound words handling, which is not considered in our current experiments, will be another important research issue in order to improve the performance of the current Oromo-English CLIR system.

## References

- [Alemu *et al.*, 2004] Atelach Alemu, Lars Asker, Rickard Coster, and Jussi Karlgén. Dictionary Based Amharic English Information Retrieval. In *CLEF 2004 Bilingual Task*, 2004.
- [Alemu *et al.*, 2005] Atelach Alemu, Lars Asker, Rickard Coster, and Jussi Karlgén. Dictionary Based Amharic French Information Retrieval. In *CLEF 2005 Bilingual Task*, 2005.
- [Alemu *et al.*, 2006] Atelach Alemu, Lars Asker, Rickard Coster, and Jussi Karlgén. Dictionary Based Amharic English Information Retrieval. In *CLEF 2006 Bilingual Task*, 2006.
- [Allan *et al.*, 2003] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, and ChengXiang Zhai. Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, 37(1):31–47, 2003.
- [Cosijn *et al.*, 2002] Erica Cosijn, Ari Pirkola, Theo Bothma, and Kalervo Jrvelin. Information access in indigenous languages: a case study in Zulu. In *Proceedings of the fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, Seattle, USA, 2002.
- [Cosijn *et al.*, 2004] E Cosijn, H Keskustalo, and Ari Pirkola. Afrikaans - English Cross-language Information Retrieval. In *Proceedings of the 3rd biennial DISSAnet Conference*, Pretoria, 2004.
- [Gey *et al.*, 2005] Fredric C. Gey, Noriko Kando, and Carol Peters. Cross-language information retrieval: the way ahead. *Inf. Process. Manage.*, 41(3):415–431, 2005.
- [Hedlund *et al.*, 2004] Turid Hedlund, Eija Airio, Heikki Keskustalo, Raija Lehtokangas, Ari Pirkola, and Kalervo Jrvelin. Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000/2002. In *Information Retrieval*, 2004.
- [Lehtokangas *et al.*, 2004] Raija Lehtokangas, Eija Airio, and Kalervo Jrvelin. Transitive dictionary translation challenges direct dictionary translation in clir. *Inf. Process. Manage.*, 40(6):973–988, 2004.
- [Marin and Pascale, 2001] Carpuat Marin and Fung Pascale. Simple Dictionary-Based Query Translation. In *CLEF 2001 Bilingual Task*, 2001.
- [Nefa, 1988] Abara Nefa. Long Vowels in Afaan Oromo: A Generative Approach. In *M.A. Thesis. School of Graduate Studies, Addis Ababa University*, 1988.
- [Nunzio *et al.*, 2006] Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. CLEF 2006 Ad-hoc Track Overview. In *CLEF 2006 Working Notes*, 2006.
- [Oard, 1997] Douglas W. Oard. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. In *D-Lib Magazine*, 1997.
- [Oromoo, 1995] Gumii Qormaata Afaan Oromoo. Caasluuga Afaan Oromoo, Jildi. In *Komishinii Aadaaf Turizmii Oromiyaa*, Finfinnee, Ethiopia, 1995.
- [Peters and Sheridan, 2001] Carol Peters and Paraic Sheridan. Multilingual information access. In *ESSIR '00: Proceedings of the Third European Summer-School on Lectures on Information Retrieval-Revised Lectures*, pages 51–80, London, UK, 2001. Springer-Verlag.
- [Peters, 2006] Carol Peters. What happened in CLEF 2006: Introduction to the Working Notes. In *CLEF 2006 Working Notes*, 2006.
- [Stroomer, 1987] H.A. Stroomer. Comparative Study of Southern Oromo Dialects in Kenya: Phonology, Morphology and Vocabulary. Burke, Hamburg, 1987.
- [Tilahun, 1989] Gamta Tilahun. *Oromo-English Dictionary*. Addis Ababa University Printing Press, Addis Ababa, 1989.

[Waterhouse, 2003] E Waterhouse. Building Translation Lexicons for Proper Names from the Web. In *Thesis, Department of Computer Science, University of Sheffield*, 2003.

[Yimam, 1986] Baye Yimam. The Phrase Structure of Ethiopian Oromo. In *Ph.D. Thesis. School of Oriental and African Studies, University of London*, 1986.