Development of Cross-Language Information Retrieval for Resource-Scarce African Languages

Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

by

Kula Kekeba Tune
200299007
kulakk@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed University)
Hyderabad – 500 032, India

July 2015

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "***Development of Cross-Language Information Retrieval for Resource-Scarce African Languages***" by **Kula Kekeba Tune**, has been carried out under my supervision and is not submitted elsewhere for a degree.

July 29, 2015

Date

Adviser: Prof. Vasudeva Varma

This dissertation is dedicated to my parents Tirunesh Yadaye and Kekeba Tune.

Acknowledgements

First and foremost, I would like to thank the Almighty God for enabling me to undertake and complete this research work. I would like to express my profound and sincere gratitude to my research advisor, Professor Vasudeva Varma, Dean (R & D), IIIT-Hyderabad, for his invaluable guidance, continuous support and utmost patience throughout the course of this study. Without his professional guidance, constant support and encouragement, I would not have been able to finish my dissertation. I am deeply grateful to my family members and friends whose love and support has always remained unwavering regardless of my limitations. This dissertation would not have been possible without the consistent help and patience of my brothers and sisters. A special thanks to my friends and colleagues Lachisaa Tolasaa, Fufa Dawo and Mysoul for their endless love, constant support and understanding. I would also like to thank the Ministry of Education, (MoE, Ethiopia), for providing me with the opportunity and financial support to pursue my post-graduate study. Finally, I would like to express my special appreciation and thanks to IIIT-Hyderabad for providing me with the academic and research supports necessary to complete my post-graduate study. I am very grateful to the academic and administrative staff at IIIT-Hyderabad, who have patiently supported and encouraged me throughout my Ph.D. research project.

Abstract

As we move towards an increasingly globalized and knowledge-based economy, the ability to discover and share information across language and cultural boundaries has become more and more crucial. With the advent and rapid development of the Internet, the amount of information generated in different languages and disseminated via the World Wide Web (WWW) and social media platforms is growing exponentially. As the Internet becomes more ubiquitous and pervasive, its users have become linguistically more diverse and culturally more heterogeneous. It is thus of utmost importance to ensure that online information resources and services are efficiently and equitably accessible to all users, regardless of their linguistic and cultural backgrounds. Unfortunately, owing to language barriers and linguistic digital divide, the majority of the world's populations, including native speakers of resource-scarce African languages, have long been denied the opportunity to access and benefit from online resources. Since most of the current major search engines and commercial Information Retrieval (IR) systems have primarily focused on well-resourced European and Asian languages, they have paid little attention to the development of Cross-Language Information Retrieval (CLIR) for resource-scarce African languages. The need for exploring and building more specialized information systems that enable speakers of African languages to discover valuable information beyond linguistic and cultural barriers has, therefore, become more urgent today than ever before. Taking these facts into consideration, this study is aimed at exploring and building an experimental CLIR between one of the severely under-resourced African languages, (i.e. Afaan Oromo) and one of the most commonly used online languages, (i.e. English). We have focused on developing and evaluating the first Oromo-English CLIR (hereafter also referred to as OMEN-CLIR), which is designed to make effective use of limited linguistic resources to search and retrieve relevant information across language and cultural boundaries.

Although Afaan Oromo is one of the major indigenous African languages that is widely spoken by more than 35 million people in the Horn of Africa, it is still considered as one of the most under-resourced languages, especially where being resourced is measured by the extent to which a given language is supported by computational linguistic resources and information access technologies. Afaan Oromo poses a huge challenge to the development of natural language processing and information access technologies, not only because it is one of the most resource-scarce languages, but also because it has very rich and complex morphological processes. Unlike English, Afaan Oromo is a highly synthetic language with a very productive inflectional and derivational morphology. In this study, we have focused on exploring and building basic linguistic resources and IR tools required for designing and developing OMEN-CLIR. Some of the major linguistic resources and translation tools that we have designed and developed during the course of this study include a generic computational model of Afaan Oromo morphology, a machine-readable bilingual dictionary, a rule-based Afaan Oromo stemmer, lists of Afaan Oromo suffixes and stopwords. While our machine-readable Oromo-English dictionary has been adopted and used as a main source of knowledge for query translation, our Afaan Oromo stemmer has played a key role in identifying and normalizing word form variations.

Apart from designing and building the components of OMEN-CLIR, for which the necessary linguistic resources and translation tools have been crafted from scratch, another major contribution of this study is to assess and evaluate the performance of the proposed retrieval system. In order to assess the performances of OMEN-CLIR, we had participated in one of the well-recognized international Cross-Language Evaluation Forums (i.e., CLEF campaign) over the past couple of years. The main focus of our evaluation in two different CLEF annual campaigns was to assess how well OMEN-CLIR performs in international and standard evaluation competitions like the ad-hoc track of the CLEF campaign. Besides a number of official runs that we had submitted to the ad-hoc tracks of the CLEF-2006 and CLEF-2007 campaigns, we have conducted various additional retrieval experiments to improve the performances of OMEN-CLIR. The evaluation results have been found to be very promising and encouraging, given the disparity of the languages involved and the limited amount of linguistic resources used for developing OMEN-CLIR. In one of our official retrieval experiments in which we have applied our Afaan Oromo stemmer, our CLIR system has achieved an average mean precision (AMP) of 29.90%, which is about 67.95% of a monolingual baseline. In general, the evaluation results show that it is viable to design and develop a CLIR system for resource-scarce African languages without relying on very rich linguistic resources that are not yet available for severely under-resourced languages like Afaan Oromo.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

In the digital world we live in today, the ability to instantly access and share relevant information, regardless of the language in which it is recorded or represented in, has become very crucial for sustainable development of a society [1, 2]. With the advent and widespread use of the Internet, the volume of information generated in different languages and made available via the World Wide Web (WWW) and social media outlets is growing exponentially. The WWW has constantly evolved and become one of the largest massive repositories of multilingual and multimedia information resources. More importantly, the WWW has become one of the most popular and powerful communication platforms, not just for exploring and discovering valuable information, but also for sharing and learning new ideas and concepts. With the rapidly increasing penetration of the Internet and mobile devices across the globe, the digital revolution has presented an ideal opportunity for achieving the longstanding goal of universal and equitable access to information resources.

However, the digital revolution has not only offered unprecedented opportunities and possibilities, but has given rise to several serious challenges that urgently need to be addressed by IR researchers and information professionals. As with many previous technological innovations, the digital revolution has not reached and benefited all nations in a uniform and equitable manner. Even though the digital revolution has dramatically narrowed the gap between the “information-rich” nations of the North and the “information-poor” nations of the South, the problem of information poverty is still prevalent in many developing countries [3, 4]. The vast wealth of information available on the Internet, which is instantly accessible to the majority of the population in the Global North, has been extended only to very restricted elites and educated individuals within the Global South. In other words, access to information and knowledge resources is still far from being universal and equitable, especially in multilingual and multicultural developing countries like Ethiopia and India. Since the majority of the world’s populations, including speakers of resource-scarce African languages, have been excluded from the emerging knowledge-based society, [3] argue that the digital revolution might have exacerbated the existing social and economic inequalities. In multilingual developing regions like Africa and Asia, access to online resources and services has been severely constrained by formidable obstacles such as language barriers, linguistic digital divide, and lack of efficient

Cross-Language Information Access **(CLIA)** systems [3, 5]. As noted by [6, 7, 8], language barriers and linguistic digital divide have continued to undermine the potential of the WWW to deliver an efficient and equitable access to online information resources and services.

As the Internet becomes more ubiquitous and pervasive, its global audiences have become culturally and linguistically more heterogeneous. It is, thus, vitally important to ensure that online information resources are efficiently and equitably accessible to all users, regardless of their linguistic and cultural backgrounds. Unfortunately, with the rapidly increasing penetration of the Internet into many African and Asian countries, the limitation of classical search engines and monolingual IR systems, which were primarily designed for a handful of resource-rich European and Asian languages, has become more and more apparent. As noted by [9, 10, 11], the larger the number of languages used on the Internet, the greater the volume of multilingual content on the Web, the more difficult to identify relevant information. Owing to lack of robust *Cross-Language Information Retrieval* **(CLIR)** systems, enormous amount of educational resources published in English have remained inaccessible to the vast majority of African population.

Addressing the challenges posed by language barriers and linguistic digital divide is of special concern to African nations because of their rich linguistic and cultural diversity. Although physical and geographical barriers were drastically reduced or virtually eliminated by the advent and rapid development of the WWW, language barriers and linguistic digital divide have emerged as the major obstacles to information access [6, 12, 9, 13]. Indeed [6, 11], much work remains to be done before language barriers can be surmounted as effectively as geographic ones. The need for exploring and developing more advanced and specialised information systems that allow speakers of resource-scarce African languages to search and retrieve relevant information beyond language and cultural boundaries has, therefore, become more critical than ever before.

Broadly speaking, *language barrier* can be defined as cultural and linguistic problem that impede the free flow of information and ideas across language boundaries. More specifically, it refers to linguistic impediments and obstacles that discourage or prevent users from seeking, discovering and communicating information across language and cultural borders. Although the term *linguistic digital divide* is closely related to language barriers, it is specifically used to describe the disparity in the development and application of computational resources across different languages and linguistic communities [3, 14, 15]. While the term *digital divide* is commonly used to refer to the disparity in accessing and using computing devices and online resources among various communities and social groups, the term *linguistic digital divide* is primarily used to describe the relative advantages of certain languages (or language communities) over others with respect to online content and computational resources such as lexical databases, parallel text corpora, Information Retrieval (IR) tools and Machine Translation (MT) systems.

Due to the massively increasing volume of multilingual content, providing information access is no longer about indexing and retrieving monolingual documents. Normally, providing access to information presupposes the availability of information and IR tools in a form that is comprehensible to the user, that is, in a language that the user can understand [16]. Sadly [8], the quantity and quality of existing linguistic resources and information access technologies vary considerably from one language to another. In particular, there is a substantial gap between a handful of resource-rich European languages and most resource-scarce African languages. Although English has been considered as the lingua-franca of the Web, the volume of online content produced in other major languages of the world, including German, Chinese, Hindi and Arabic, has been growing at a torrid pace in recent years. At the same time, the number of non-English speaking Internet users has been growing at a faster rate than English speaking Internet users (see section 1.4 for more details). On the other hand, since most of the current major search engines and commercial IR systems have primarily focused on better-resourced European and Asian languages, they have not paid much attention to the need for developing CLIA systems for under-resourced African languages. Bridging language barriers and linguistic digital divide is essential to the achievement of universal and equitable access to information resources. To this end, much attention has been recently given to the development and application of CLIR, which is primarily concerned with searching and discovering relevant information beyond language and cultural barriers [2, 17]**.**

As noted by [3, 4], in multilingual and multicultural societies, searching for information is not limited to the native language of the users. There are many situations where users may want to search and retrieve information regardless of the language in which it is written or recorded in. A few examples of such situations where the application of CLIR is critical may include [2]:

- **The desired document or information is not available in the user's native language**: As mentioned earlier, the quantity and quality of existing online resources vary considerably from language to language. In particular, there is a huge gap in the volume of online content produced in better-resourced European languages like English and in severely under-resourced African languages like Afaan Oromo. Most speakers of African languages are facing serious challenges to find reliable information owing to language barriers and linguistic digital divide. The development of CLIR for resource-scarce African languages is critical to enable indigenous African communities to access and benefit from the vast amount of multilingual content available on the Web.

- **The desired document is available in the form of image or multimedia data that can be easily understood by the users**: This is the case of image and multimedia information retrieval, in which the target document collection is catalogued or annotated in unfamiliar or foreign languages. Since most of the multimedia document retrieval methods are largely based on the captions or textual descriptions, there is a growing interest in developing and employing CLIR to search and identify important images and photos.

- **The user may intend to find all the relevant information available, whatever the language is used**: This is a case of recall-oriented retrieval. A typical case is patent retrieval, where an entrepreneur or a patent professional tries to identify if there is an existing patent for a technology or invention across different languages. In such situation, since it is difficult limit the search to only one language, there is a critical need to extend it to many foreign languages. Another typical situation is when a company tries to identify if there is an international competitor or collaborator in the same business sector, where the search should not be limited within the same country and the same language.

In CLIR, because the user request is expressed in a language different from the language of the target collection, it is difficult to match the search query against the index of the documents directly [2, 7]. One of the major tasks of CLIR researcher is, therefore, to bridge the language gap between the query and the documents. But, accomplishing this feat is a non-trivial task. The linguistic and cultural differences between the languages considered should be thoroughly investigated and bridged by designing and building appropriate computational resources and translation tools.

Although there are a wide range of options and methods to implement CLIR, two major approaches are commonly used to overcome language barriers, namely: *translating search requests to the language of documents or translating documents to the language of the queries*. To be more specific, either the queries should be translated into the language of target documents, or the documents should be translated into the language of the queries. The former approach seems more feasible and easier than the latter because query translation is much more efficient and easier than translating a large collection into the language of the user’s query. As indicated by [6], query translation is generally preferred because the translation of the documents, which would require prior knowledge of the languages in which the users are likely to express their requests, is not feasible for very large document collections. Since the translation of extremely large and dynamic online content is very expensive and time-consuming, most CLIR researchers have focused on exploring and developing different query translation techniques and approaches. Search queries are easier to translate, because they are typically short and can be translated as “bag-of-words”, whereas document translations have to obey more complex rules and procedures of natural language processing [10]. Another major advantage of query translation over document translation is that a query translation module can be easily adapted and incorporated into an existing monolingual IR system.

There are a range translation resources and techniques that could be employed for translating a source language query into a target language query. As noted by [10, 18, 19], the three major translation resources that are commonly used for query translation are:

- Multilingual dictionaries;
- Parallel corpora, and
- Machine translation systems.

Although existing machine-translation system or annotated parallel corpora can be easily adopted for query translation, the availability and quality of such advanced linguistic resources varies greatly from language to language. As indicated by [11], while well-established MT systems and annotated parallel corpora are readily available for resource-rich European and Asian languages, such advanced linguistic resources are either very scarce or not available at all for under-resourced African languages. As noted by [20], it may take a considerable amount time before sufficient computational linguistic resources and machine translation systems are readily for indigenous African languages. In other words, core linguistic resources and translation tools necessary for building CLIR must be designed and developed from scratch. One of the major goals of this study is, therefore, to identify and construct basic linguistic resources and translation tools necessary for developing **Oromo-English CLIR (**hereafter also referred to as **OMEN-CLIR)**. As noted in [19, 17], because it is a much cheaper and less resource intensive than the other two options described above, a dictionary-based query translation is often considered as a more feasible and realistic approach in developing CLIR for under-resourced languages like Afaan Oromo.

Over the last two decades, CLIR has significantly evolved and emerged as one of the most challenging and demanding areas of IR research [2, 7]. A number of important CLIR studies have been presented and discussed at well recognized international conferences and workshops such as Text REtrieval Conference (TREC)[1] and Cross-Language Evaluation Forum (CLEF)[2]. Unfortunately, since most of the earlier studies were highly concentrated on a handful of well-resourced European and Asian languages, the need for exploring and developing CLIR for under-resourced African languages has been left unaddressed for quite a long time. While the development of CLIR related to less-resourced Asian languages has received very little attention, the need for investigating and building CLIR for resource-scarce African languages has been left unaddressed [3, 14, 20, 11].

Evidently, the challenges posed by language barriers and linguistic digital divide have continued to threaten and undermine the promise of the digital revolution. While online information resources are abundantly available and accessible to speakers of well-resourced European and Asian languages [11], these online resources are not accessible to the majority of speakers of African languages. The challenges posed by language barriers and linguistic digital divide may persist or even worsen unless concerted research efforts are immediately directed towards the development of computational linguistic resources and information access technologies related to indigenous African languages. To this end, this study is aimed at exploring and developing an experimental CLIR between one of the most resource-scarce African languages (i.e. Afaan Oromo) and one of the most widely used online languages (i.e. English). We have focused on exploring, building and evaluating the first Oromo-English CLIR (OMEN-CLIR), with a view to enhancing access to information and knowledge resources across language boundaries.

Afaan Oromo, which is the native tongue of more than 35 million people, is one of the major indigenous African languages that is widely spoken in the Horn of Africa in general and in most parts of Ethiopia in particular. Afaan Oromo is also spoken in the adjoining parts of neighbouring countries such as Kenya and Somalia. Currently, Afaan Oromo is an official language of Oromia

[1] http://trec.nist.gov/
[2] http://clef.isti.cnr.it/

State, which is the largest and most populous Region in Ethiopia. However, despite being one of the major indigenous African languages, Afaan Oromo is still considered as one of the most severely under-resourced languages, especially where being resourced is measured by the extent to which a given language is supported by computational linguistic resources and information access technologies. Like many other resource-scarce African languages, it is characterized by the lack of computational linguistic resources such as digital data, lexical databases, text processing facilities and IR tools.

On the other hand, Afaan Oromo has a very rich and complex morphology. Unlike English, which is often considered as a moderately analytic language, Afaan Oromo is a highly synthetic and agglutinative language with a relatively flexible word order. While most grammatical categories in English are conveyed through prepositions, auxiliaries and word order, syntactic functions and grammatical relations in Afaan Oromo are indicated by inflectional affixes and postpositions. As noted by [21, 22], most grammatical categories in Afaan Oromo, including number, gender, case and tense are marked by inflectional affixes (or bound morphemes). As described in more detail in section 3, while most Oromo nouns decline for number, gender, definiteness and case, Oromo verbs conjugate for person, gender, number, tense and aspect. Suffixes are predominantly used for both inflection and derivation, and they can be strung together one after another. For example, the Oromo noun "*adurree*" (cat) can take different inflectional suffixes such as "*adurree* + *icha* + *tti*" → "*adurrichatti*" ("cat", definiteness, dative); and "*adurree* + *ota* + *in*" → "*adurrootiin*" ("cat", plural, nominative). The morphological variations and complexities in Afaan Oromo are further exacerbated with very rich and productive derivational processes such as nominalization, verbalization and compounding. For example, a few derivatives of the Oromo verb "*beekuu*" ("to know" or "aware") include: "*beekaa*" (skilled or expert), "*beekumsa*" (knowledge), "*beekamuu*" (to be known), "*beekamaa*" or "*beekamtuu*" (famous or popular).

Such complex morphological processes implies that Afaan Oromo has a large number of distinct word forms, which has a major impact in designing and developing NLP applications such as IR, MT and CLIR. In other words, Afaan Oromo poses a huge challenge to the development of CLIA systems, not only because it is one of the most resource-scarce African languages, but also because it has very complex and productive morphological processes. In a dictionary-based CLIR, where a bilingual lexicon is used as a main source of knowledge for query translation, a search query expressed in a source language must match with the citation form of the lexicon.

However, due to word form variations, many query terms may fail to match with a semantically equivalent lexical entry. In morphologically very rich and complex languages like Afaan Oromo, the entire process of query translation may fail due to the problem of word form variations. Hence, the task of identifying, removing and normalizing morphological variants becomes very critical. As indicated by [23, 24], for highly synthetic and agglutinative languages like Afaan Oromo, removing frequent inflectional and derivational affixes is crucial for improving the performance of CLIR. To this end, we have designed and developed a rule-based stemmer that efficiently identifies and removes very common inflectional and derivational affixes. More detailed description of our stemming algorithm is presented in section 5.4.

As indicated earlier, the performance of a CLIR system largely depends on the quantity and quality of linguistic resources and language processing tools available for its implementation. During the course of this study, we have devoted lots of efforts in identifying and building basic linguistic resources and translation tools necessary for the development and evaluation of OMEN-CLIR. Some of the important linguistic resources and translation tools that we have designed and developed during the course of this study include a machine-readable bilingual dictionary, Afaan Oromo stemmer, list of morphological affixes and stop-words. While our machine-readable bilingual dictionary has been used as a main source of knowledge for query translation, our Afaan Oromo stemmer has played a key role in identifying, removing and normalizing word form variations. We have also adopted a phonetic based transliteration method to handle some of the technical terms and proper names that are not found in the bilingual dictionary.

Obviously, evaluation is essential for the development and improvement of CLIR. Apart from designing and building the basic components of our OMEN-CLIR, for which most of the required linguistic resources and translation tools have been constructed and developed from scratch, another major contribution of this study is to assess the performance of the proposed retrieval system at one of the well-recognized international evaluation forums like the CLEF campaign. As indicated earlier, various international competition and evaluation forums such as TREC and CLEF have been established to support IR researchers in testing and assessing the effectiveness of their retrieval systems. Over the past couple of years, we had participated in the ad hoc track of the CLEF campaign [13, 25, 26, 27]. Over the last one decade and a half, CLEF has been supporting and promoting the comparative evaluations of cross-language information access systems [7, 28]. The participation in such international competition plays a key role in helping the researchers to test and measure the performance of their CLIR system against well-established test-collections and standard evaluation techniques. It also enables the participants to compare the performance of their retrieval system with the performances of other CLIR systems that have been tested on the same datasets.

The main focus of our evaluations in two different CLEF campaigns was to determine how well OMEN-CLIR performs in standard and international competitions like the ad hoc track of the CLEF campaign. We have been also interested in investigating the impacts of the major components OMEN-CLIR on the effectiveness of the retrieval system. In this light, we have conducted a series of retrieval experiments based on the general guidelines and procedures recommended for the CLEF campaign. We have used two different sets of test collections distributed by the CLEF campaign organizers. Besides the official runs that had been submitted to the bilingual task of the ad-hoc track, we have conducted various additional retrieval experiments in order improve the performances of OMEN-CLIR. Most of our evaluation results were found to be very promising and encouraging, given the limited amount of linguistic resources and IR tools that we have employed for developing OMEN-CLIR. We will describe and discuss some of the major resources and approaches that we have adopted for the development and evaluation of OMEN-CLIR, along with the major findings of our study in the subsequent chapters of this thesis.

1.2 Basic Definition of CLIR and IR

IR is a multidisciplinary field of study that deals with the development and application of information systems that are designed to assist users to identify relevant items from a very large collection of unstructured (or quasi-structured) documents [1, 29]. It is mainly concerned with designing, implementing, evaluating and managing information systems that allow users to efficiently discover information from extremely large databases and online repositories. Some of the major activities that are involved in a typical IR system include: classifying, organizing, indexing and managing a large collection of documents with a view to better enable the user to retrieve items pertinent to his/her search request. A search request or a query, which is often expressed in a natural language, is considered as a representation of user's information needs. In a classical IR, unlike CLIR, since search requests and target documents are represented in the same language, they can be processed and matched against one another in order to find relevant items. However, nowadays IR has grown beyond its primary goals of indexing and organizing monolingual documents. With increasing demand for more advanced and specialized information access and retrieval services, many new areas and subfields of IR have recently emerged, including Question Answering, Multilingual Information Retrieval (MLIR) and CLIR.

As indicated by [2, 7], CLIR is a subfield of IR that deals with searching and discovering information across language and cultural boundaries. Some of the major activities involved in CLIR include query translation, document indexing, searching and retrieving relevant items across different languages. As indicated by [6, 17], the main focus of CLIR is to search and retrieve documents written in one or more language(s) in response to a query expressed in another language. Suppose a search query is given in one source language, (e.g. Afaan Oromo), a CLIR system is expected to identify and retrieve documents written in another target language, (e.g. English). The process is called bilingual [18], when dealing with a language pair, i.e. one source language and one target language. In the case of Multilingual Information Retrieval (MLIR), users can express their search requests in one or more source languages in order to retrieve relevant items across multiple languages. While the main focus of CLIR is to process and translate a search request expressed in a source language into a target language query, an MLIR system can accept search requests in two or more languages to retrieve relevant information across multiple languages. The main objective of both CLIR and MLIR is to develop automated information access tools and techniques to discover relevant documents beyond language and cultural boundaries. Since an MLIR is a broader term, it can embrace the major research issues that are involved a CLIR.

Due to its rapidly growing recognition, CLIR has been extensively studied over the recent two decades [7]. However, CLIR has been sometimes incorrectly referred to as an information system that is specifically developed to assist users to retrieve documents published in unfamiliar language(s) [24, 30]. This notion is not strictly true. Even if the user can understand many languages, it is still a burden for him/her to formulate different search queries to find relevant items across multiple target languages. The number of languages and the amount of multilingual content available on the Web far exceeds the ability of any polyglot user to search and retrieve relevant items in real time. By automatically accepting and translating a query expressed in one source language into multiple target languages, a CLIR system can help a polyglot user to discover relevant documents across multiple languages efficiently. For instance, a polyglot researcher who is planning to conduct a formal study on a given topic may use a CLIR system to find out whether the topic of his/her research interest has been already studied elsewhere in other major languages of the world.

On the other hand, in addition to IR and MLIR, there is another important concept that is closely related to CLIR, which is referred to as Cross-Language Information Access (or CLIA). In most research studies, **CLIA** is often used to describe a combination of CLIR and MLIR, because it can embrace both concepts. However, CLIA deals with much more general issues that may include not only the academic domain of information access and retrieval, but also many aspects of language processing, text analysis and understanding. In other words, CLIA combines different strategies and technologies used in classical Information Access (IA) and IR systems with methodologies and resources in computational linguistics and NLP [7]. In this study, we use the term CLIA in its narrower sense to describe the activities of querying, accessing and retrieving information across different languages.

1.3 Why CLIA for Resource-Scarce African Languages?

1.3.1 Resource-Scarce Languages: Definitions and Characteristics

Language is fundamental to most aspects of human life, including thinking, learning and communication. As a primary means of communication and social interaction, language is an integral part of human culture and civilization. It is a principal medium through which cultural and social values have been preserved and transmitted from generation to generation. Language can be briefly defined as a medium of communication through which our thoughts, ideas or feelings are expressed and exchanged by using a set of arbitrary symbols such as written texts, vocal sounds, auditory or visual gestures in conventional and understandable ways. On the other hand, culture is a broader concept, language being one of its components. Culture is a cumulative experience, which may include ideas, beliefs, morals, arts, traditions, and any habits acquired by a certain community or a group of people in a society. It represents the total system of habits and behaviour of certain social groups or communities in which language is an essential element. For centuries, language has been used as a primary medium of instruction for education and life-long learning. According to [31, 32], any language is capable of being a vehicle for thoughts and social interaction. Hence, many researchers argue that there is no human trait more pervasive than language. Most of our personal perceptions and reflections as well as our social values are embodied in and expressed through language. Every language carries the collective perspectives and ideas of hundreds of generations who have shaped it. In particular, indigenous languages are strongly associated with the cultural heritages and social values of their native speakers. However, as different linguistic communities come in contact with one another and began to interact with each other, a number of national and regional languages have been adopted and widely used to foster mutual understanding and cooperation among people from different cultural backgrounds. In general, most of the major achievements in human history including philosophy, art, science and technology have been encoded, preserved, shared and communicated through languages [33].

The information revolution that began in the last half of the twentieth century and continues to unfold today has profound implications for science, culture, education, political and economic life. Language plays a fundamental role in this revolution since information is primarily created and accessed in linguistic form. Because information is largely conveyed through spoken and/or written texts, most users are searching and accessing online content by using language. Accordingly, language serves not only as a principal means of communication and bearer of cultural heritages, but also as a vital instrument of seeking and acquiring knowledge. In an increasingly knowledge-based economy [4], language plays a key role similar to that of money in industrial society. While money has played a critical role in acquiring material resources and tangible goods, language plays a key role in acquiring knowledge and learning new ideas. In a highly networked and knowledge-based society, language is the primary medium that is primarily used for sharing and disseminating information.

On the other hand [34], language has many structural levels and layers. From the perspectives of computational linguistics, understanding and analysing a language involve identifying and computing the linguistic properties and structures of language at different levels including phonology, morphology, syntax and semantics. Unfortunately [8], the quantity and quality of existing linguistic studies and computational resources vary considerably from one language to another. In particular, there is a huge gap between better-resourced European languages and severely under-resourced African languages. According to [14], very few of the world’s 7000 languages are currently enjoying computational linguistic resources and language processing tools such as part-of-speech tagger, morphological analyser and machine translation (MT) and IR systems. More specifically [35], only about 10% of the existing human languages have managed to assemble and organize computational linguistic resources such as parallel text corpora, machine-readable dictionary, multilingual thesaurus, part-of-speech tagger, morphological analyser and the like (which are often referred to as Basic Language Resource Kit or BLARK). The remaining, (more than 90% of the world's living languages), lack core linguistic resources necessary for the development of natural language processing and cross-language information access systems. Therefore, they are often referred to as *under-resourced* or *resource-scarce languages*.

The term *under-resourced* or *resource-scarce languages* is used to refer to a group of languages that have very little or no computational linguistic resources required for building natural language processing applications and information access technologies like MT, IR and CLIR systems. Although the term resource-scarce languages is mainly used to denote a group of languages for which insufficient amount digital data are available, sometimes it is also used to refer to a group of languages with very limited commercial and economic influence, especially languages that lack the clout of commercial interest [36]. Since major software companies and online service providers have little commercial interest in building and providing information access technologies related to resource-scarce languages, most African languages have remained underrepresented and underserved on the WWW over the last two decades.

It is important to note that language developers and IR researchers have used different terms to refer to resource-scarce languages. For instance, the term "*minority languages*" has been widely used to refer to resource-scarce languages in regions where there is a dominant language [37]. But, a language that is considered as a minority in one region may not be a minority in another region or country. This is the case for Afaan Oromo, a minority language in Kenya but a majority language in Ethiopia. Another term that is commonly used to refer to under-resourced languages is "*less documented languages*." In this case, the focus is on the volume of written documents or literatures available in a given language. For the purpose of building CLIR, the fact that a given language is widely used or well documented in written and spoken form is relevant but not necessarily essential. The availability of reliable computerized linguistic resources and translation tools are more crucial to the development of CLIR and MT systems than the availability of literature or documents. Most under-resourced languages are characterized by lack of important computational linguistic resources and translation tools.

In summary, some of the major characteristics of resource-scarce languages include [3, 14, 38, 16] lack of:

- lexical resources such as monolingual and bilingual machine-readable dictionaries,
- sufficient online content and linguistic data such as annotated comparable corpora and parallel corpora;
- language processing tools such as tokenizer, stemmer, part-of-speech tagger and morphological analyser;
- text analysis and proofing tools such as spelling and grammar checkers;
- semantic of semantic analysis tools such as Word Sense Disambiguation (WSD) and semantic roles;
- significant online presence and commercial influence;
- translation memories and MT systems;
- cross-lingual and multilingual information access technologies such as CLIR and MLIR;
- detailed linguistic study and computational models due to shortage of language experts and computational linguists.

1.3.2 Rationale for Developing CLIA for Resource-Scarce African Languages

Over last few decades, English has been viewed as the lingua franca of the Web. Like many other parts of the world, most digital devices and communication technologies have been introduced to Africa in major European languages like English and French. However, the vast majority of population in the continent either does not speak English or cannot understand it very well [4]. When the WWW initially began in the mid-1990s, the dominance of English on the Internet was not an apparent problem. As described by [7], the first websites and webpages were almost entirely dedicated to provision of information in English and the first search and retrieval services in the mid-1990s (e.g., Lycos, AltaVista, Yahoo) were exclusively implemented to meet the needs of well-educated users and native English speakers. Those earlier users of online resources had not only good academic backgrounds, but also sufficient English language skills to express their information needs and formulate meaningful search queries in English. Most of these earlier web users can also read and understand the documents retrieved by the monolingual IR system.

However, the rapid development of the Internet over the last two decades has led to an exponential growth in the volume of digital content published in different major languages of the world. As pointed out by [7], information sources published in major online languages such as

English and French are potentially valuable and relevant for many speakers other languages. The major problem, of course, is that the majority of the world population does not have efficient CLIA systems to search and identify relevant information across language and cultural boundaries [7, 39]. Owing to linguistic digital divide and language barriers, the overwhelming majority of African population has remained excluded from the digital world. Lack of efficient access to online resources and services makes it virtually impossible for many developing nations to participate in the knowledge-based economies of the 21st century. Although a number of less-resourced Asian languages have recently made significant progress in building basic computational linguistic resources and IR tools necessary for searching and retrieving information online, the need for building CLIA for resource-scarce African languages has been left unaddressed for quite a long time.

As the Internet expands and reaches multilingual developing nations like Ethiopia and India, the number of non-English speaking users becomes more and more prevalent on the Web. As indicated by [8], more than two thirds of global Internet users are now estimated to be non-English speakers and their number has continued to grow steadily. Most of these users are more comfortable to express their information needs in their vernacular languages [20, 40]. For instance, most speakers of African languages often prefer to use their native language to find valuable information online. This should come as no surprise, as education and communication are generally easier in the first language than in languages that people acquire later [20]. Because indigenous languages are deeply ingrained in the daily lives of African communities, they are considered as more reliable tools for querying and finding relevant information.

Unfortunately, since most of the major search engines and commercial IR systems have been primarily designed for well-resourced European and Asian languages, the need to develop CLIA for resource-scarce African languages has received very little attention. As a result, speakers of resource-scarce African languages are facing daunting challenges in searching and discovering information online. As noted by [8], without the support of multilingual information access technologies such as CLIR and MT systems, language barriers and linguistic digital divide are insurmountable obstacles to the accessibility and usability of online resources. To be more specific, without the development and deployment of efficient CLIR systems, finding relevant information across linguistic and cultural borders is either impossible or must be performed in non-native and unfamiliar languages that may not be understood by the users.

In spite of the massive amounts of electronic documents and digital data available on the Web, the overwhelming majority of native speakers of African languages are still suffering from *information poverty*. According to [41, 42], *information poverty* is defined as the condition of life where a significant number of people lack information that is necessary to meet their social and economic needs for survival. The abundance and richness of digital content produced in English or other well-resourced languages is meaningless to the majority of African communities as long as they cannot search and retrieve relevant items. According to [43], providing the majority population in developing countries with the opportunity to access online resources is essential to alleviate information poverty. Solving the economic and social problems in developing countries is not about choosing between internet access and basic necessities; because both need to complement each other to enable developing nations to participate in the emerging knowledge-based societies. Providing all users with universal and equitable access to information resources may remain an unachievable goal as long as the major search engines and commercial IR systems are exclusively designed for speakers of well-resourced languages. To achieve the desired goal, research efforts should be geared towards the development of CLIA for resource-scarce African and Asian languages.

In summary, some of the major reasons for the development of CLIA for under-resourced African languages include:

- **To bridge linguistic digital divide:** While computerized language processing tools and IR systems are readily available for resource-rich European languages like English and French, most resource-scarce African languages lack core linguistic resources and IR tools necessary for searching and discovering information online. Today, the problem of linguistic digital divide is nowhere more is nowhere more prevalent than in sub-Saharan African countries like Ethiopia. Due to the lack of computational linguistic resources and language processing tools required for the application of information access technologies, the majority of resource-scarce African languages have received little attention from IR researchers. Owing to language barriers and linguistic digital divide, most native speakers of African languages are not yet in a position to access and benefit from online resources. Clearly, language barriers and linguistic digital divide have continued to perpetuate the gap between the “*information-rich*" nations of the North and the "*information-poor*" nations of the South, putting several million indigenous communities in developing countries at very serious economic disadvantages [41]. Addressing the challenges posed by linguistic digital divide is of immediate concern to Africa. Resource-scarce African languages that fail to cope up with the constantly evolving and rapidly changing communication technologies are facing the danger of digital language extinction [8, 3], failing to serve their native speaker effectively in the digital age. As noted by [4], African languages must strive to cope up with the advancements of modern language processing and information access technologies if they are to prosper in the 21st century. Building multilingual information access technologies focused on indigenous African languages is crucial for bridging language barriers and linguistic digital divide.

- **To reduce the risk of cultural breakdown:** According to a number recent linguistic studies, of the estimated 6,900 languages that exist around the world today, more than 80% will have disappeared or will be endangered by the digital language extinction by 2100. The extinction of thousands of languages and associated cultural heritages and values presents a permanent and, if not countered, irrevocable cultural and scientific loss to humanity [4]. African languages, like all other languages of the world, carry with them unique indigenous knowledge and cultural heritages. They possess extremely diverse and rich cultural heritages that have been passed down for many generations amongst the indigenous African communities. Besides unique and rich oral traditions, indigenous African languages are also bearers of social norms and philosophical thoughts that have been preserved and transmitted from generation to generation for many centuries. Since African languages have deep cultural ties with the majority of indigenous African communities, there is a growing concern over their marginalization and underrepresentation in cyberspace [20, 31, 44]. Absence of adequate research on indigenous African languages is, indeed, unfortunate since the investigation of these less-documented languages can lead to many interesting linguistic phenomena, which have never been studied and discovered over the past several centuries. Failure to revitalize and promote severely under-resourced African languages will not only result in the loss of those languages, but will lead to the loss of unique cultural heritages. The development of linguistic resources and information access technologies for resource-scarce African languages is critical to prevent the cultural breakdown of indigenous African communities.

- **To promote linguistic and cultural diversity on the Internet:** The need for the development of cross-language information access systems for resource-scarce languages stems from the fact that linguistic and cultural diversity is crucial for building and maintaining an inclusive and multicultural knowledge-based societies. As noted by [14, 8], any language that fails to cope up with the advancement of modern communication technologies is in danger of becoming irrelevant and extinct in the digital age. As indicated by [37], due to lack appropriate computational linguistic resources and translation tools, the task of building multilingual information access systems related to resource-scarce African languages has been left unaddressed for many years. As noted by [8], equipping resource-scarce languages with sufficient linguistic and efficient CLIA systems is vital to promote linguistic and cultural diversity in cyberspace. Moreover, the development of indigenous African languages is very important to reduce the negative impacts of globalization on the cultural values and social norms of African communities [31]. The Internet will become a truly multilingual and multicultural global medium when severely under-resourced African languages are technologically developed and widely used on the WWW.

- **To accelerate the accessibility of online resources and services:** The Internet has presented unparalleled opportunities for users to share and exchange information across the globe. In principle, the WWW is open and accessible to everyone, provided basic network connectivity requirements are met and in place. Realistically, however, there are various barriers and obstacles. Most resource-scarce African languages are either under-represented or completely missing on the Web. Due to language barriers and lack efficient CLIA technologies, finding up-to-date and relevant information is still a major challenge for the majority population in developing countries. The provision of universal and equitable access to online resources will remain an unattainable in Africa as long as major search engines and IR systems are exclusively designed and optimized for a hand full of better-resourced European and Asian languages. To achieve this desired goal, African researchers must strive towards building CLIA systems for their resource-scarce languages. The development of CLIR that involves severely under-resourced languages like Afaan Oromo can play a pivotal role in enhancing and accelerating the accessibility of information resources across linguistic and cultural borders.

1.4 Motivation

As we move towards an increasingly knowledge-based and innovation-driven society, information has become one of the most valuable resources for sustainable development. According to [45, 46, 47, 48], a knowledge-based society should be an inclusive society where all citizens, regardless of their linguistic and cultural backgrounds, have the right to access and use information resources. Unfortunately, as indicated in Table 1.2, while about a quarter of the world's population enjoys the opportunity to access and benefit from the vast wealth of knowledge available on the Web, the majority of the world's citizens, including native speakers of resource-scarce African languages [4], do not have reliable access to online resources. Due to the ever-increasing amount of multilingual content, the problem of language barriers and linguistic divide has become a severe bottleneck to the accessibility of online resources. Concerning this, [49] has noted that despite the rapid penetration of the Internet into many developing countries, language barriers and linguistic digital divide have continued to impede the accessibility and usability of information resources and services.

Being a home to about $\frac{1}{3}$ of the world's languages, Africa is the second most linguistically diverse continents on the planet. Linguistic and cultural diversity is the norm rather than the exception in Africa. As noted by [8], although linguistic diversity is a source of social prestige and economic benefits, without the availability of robust CLIA systems, it presents a huge challenge in accessing and retrieving relevant information. As shown in Table 1.1, there is a wide gap between the number of languages that are predominantly used on the Internet and thousands of languages spoken around the world. In fact, the number of languages widely used on the Internet is a small fraction of the total number of languages in the world. For instance, out of the estimated 6,900 living languages in the world today, more than 30% of them are found in Africa [50], which is the second largest and most-populous continent after Asia. A large number of indigenous African languages including Afaan Oromo, Amharic, Hausa, Swahili and Somali are widely spoken by several million people [51]. However, due to lack of computational linguistic resources and information access technologies, most of these indigenous African languages are not yet widely used on the Web.

As shown in Table 1.1, the top 10 European and Asian languages that are widely used on the Internet make up around 80% of the total languages that are currently used online. The rest of the world languages (combined) make up only the remaining 20%. As indicated in Table 1.2, while

Europe with about 12.0% of the world's population accounts for more than 27% of the Internet usage, Africa with more than 14.3% of the world's population accounts for only about 3.6% of the Internet usage. As noted by [49, 52], English is still the most commonly used language on the Web. Evidently, the challenges posed by language barriers and linguistic digital divide have continued to threaten the promises of the Internet revolution. While online resources are constantly produced and abundantly available in a handful of well-resourced Western languages [3, 11], thousands of resource-scarce African and Asian languages have remained either excluded from or underrepresented on the WWW. In other words, most speakers of indigenous African languages are denied the opportunity to access and use online information resources that they desperately need. Taking these facts into consideration, the central research question that this study seeks to address is: *How African nations can mitigate and overcome the challenges posed by language barriers and linguistic digital divide?*

Top Ten Lang. on the Internet	**% of all Internet Users**	**Number of Internet Users by Language**	**World Popn. for the Language**
English	30.4%	427,436,880	2,039,114,892
Chinese	16.6%	233,216,713	1,365,053,177
Spanish	8.7.0%	122,349,144	451,910,690
Japanese	6.7%	94,000,000	127,288,419
French	4.8%	67,315,894	410,498,144
German	4.5%	63,611,789	96,402,649
Arabic	4.2%	59,810,400	357,271,398
Portuguese	4.1%	51,180,960	239,649,701
Korean	2.5%	34,820,000	72,711,933
Italian	2.4%	33,712,383	58,175,843
Top 10 Total	84.8%	1,194,454,163	5,218,073,846
Rest of the Languages	15.2%	213,270,757	1,458,046,442
World Total	100.0 %	1,262,032,697	6,676,120,288

Table 1.1 Top Ten languages used in the Web

World Regions	Population	Popn. (% of World)	Internet Users	Penetratn. (% Popn.)	Usage % of World	Usage Growth
Africa	955,206,348	14.3%	51,022,400	5.3%	3.6%	1030.2%
Asia	3,776,181,949	56.6%	529,701,704	14.0%	37.6%	363.6%
Europe	800,401,065	12.0%	382,005,271	47.7%	27.1%	263.5%
Middle East	197,090,443	3.0%	41,939,200	21.3%	3.0%	1176.8%
North America	337,167,248	5.1%	246,402,574	73.1%	17.5%	127.9%
Latin America	576,091,673	8.6%	137,300,309	23.8%	9.8%	659.9%
Oceania	33,981,562	0.5%	19,353,462	57.0%	1.4%	154.0%
WORLD TOTAL	6,676,120,288	100.0%	1,407,724,920	21.1%	100.0%	290.0%

Table 1.2 World Internet usage and population statistics

Some of the major motivating factors for proposing and conducting this study are summarized as follows.

1. **To help speakers of African languages to overcome language barriers:** In principle, all languages, including resource-scared African languages, can be represented and effectively used on the Internet, regardless of their national or international status. However, there are many economic and cultural factors that hinder the application of resource-scarce languages on the WWW. As indicated Table 1.2, in spite of the recent rise in the availability of online access, the penetration of the Internet in most African countries is very low (less than 5%), especially when it is compared with the penetration of the Internet in North America (73.4%) and Western Europe (47.7%). As noted by [49], language barriers, linguistic digital divide and lack efficient cross-language information access technologies are some of the major obstacles that hinder the accessibility and usability of online information resources. Without the support of cross-language information access technologies such as CLIR and MT systems, Africa's rich linguistic diversity is an insurmountable obstacle to information access and knowledge sharing. Providing efficient access to information resources may remain an unattainable goal in Africa as long as the major search engines and commercial IR systems are

exclusively designed for a handful of resource-rich European and Asian languages. In other words, the development of CLIR that allows speakers of resource-scarce languages to search and retrieve information beyond language and cultural barriers is vital to achieve a universal and equitable access to online resources. Unless concerted research efforts are immediately directed towards the development of computational linguistic resources and information access technologies related to under-resourced African languages, several million native speakers of these languages will remain isolated and excluded from the rapidly growing knowledge-based societies of the 21st century.

2. **Rapid growth of non-English speaking users:** Over the last few decades, since English has been considered as the lingua franca of the Web, the challenges posed by language barriers and linguistic digital divide have not received adequate attention from information professionals and IR researchers. Most commercial search engines and IR systems have focused on developing and providing monolingual information retrieval services related to English and other better-resourced European languages. Nowadays, the number of non-English speaking web users is growing at a faster rate than the number of English-speaking web users. But, due to language barriers and lack of reliable CLIR systems, vast amount valuable information produced in English is not yet accessible to the majority of populations in developing countries. This is especially true in multilingual and multicultural nations like Ethiopia. It is, thus, of utmost importance to ensure that the vast wealth of information available on the Web is efficiently and equitably accessible to all users, regardless of their linguistic and cultural backgrounds. The development of a CLIR system for indigenous African languages is crucial to facilitating the accessibility and usability of online resources across linguistic and cultural boundaries.

3. **Increasing penetration of smartphones and mobile devices in Africa:** Over the recent few years, the penetration rate of mobile phones and portable devices has been growing at very fast pace across Africa. As noted by [4, 53], affordable smartphones and mobile devices have recently emerged as an important means of online communication in many Sub-Saharan African countries including Ethiopia. For instance, since 2000, the number of mobile connections in Sub-Saharan African (SSA) has grown by 44%, compared to an average of 34% for developing regions and 10% for developed regions as a whole [53]. Mobile telephones and portable digital devices have the potential to deliver the benefits of the Internet across poorly networked and less connected rural areas in developing

countries. The lack of affordability, limited coverage and unreliability of wired networks across the SSA means that mobile broadband is a good alternative way for consumers to access and share information online. The rapid expansion and growth of mobile broadband across the SSA is expected to continue in the years to come, reaching small towns and less populous rural areas of Africa. Overall, while mobile Internet traffic is expected to grow by 25-fold over the next four years in SSA, 3G and 4G penetration levels are forecasted to grow by 46% through 2016 [43, 53]. Along with the evolution of the mobile technologies to handle multimedia and text messaging, there is increasing interest in building and providing localized digital content in indigenous African languages. The development and availability of reliable CLIR systems for indigenous African languages will enable the local communities to search and retrieve valuable information across language and cultural boundaries. The rapid penetration smartphones and widespread use mobile technologies in African countries have, thus, presented unprecedented opportunities for building and providing multilingual information access systems and services.

4. **Establishment of community-based information centres:** A large number of Community-Based Information Centres have been recently established in many African countries to facilitate the accessibility and availability of online resources and services. Some of these modern community-based information centres include: Multipurpose Community Telecentre (MCT), Community Multimedia Centre (CMC) and E-Health Care Information System. For example [39], the New Partnership for African Development (NEPAD) has launched a very exciting “e-schools” initiative designed to connect 600,000 schools across the continent through the application of web-based technologies. NEPAD has supplied many schools with computers, satellite hook-ups, television monitors, and video conferencing equipment. However, the problem of language barriers has become a major obstacle in searching and retrieving relevant documents. Due to the scarcity of localized digital content and lack of suitable CLIR, most users are not able to access and benefit from the electronic educational resources. The development of a CLIR system that involves major indigenous African languages like Afaan Oromo and the most commonly used online languages like English can play a key role in enabling the users to overcome language barriers and linguistic digital divide.

5. **Encouraging studies on the development of CLIR for resource-scarce languages:** During the last decade, some important studies have been reported on the development and application of CLIR for resource-scarce African and Asian languages [33, 34, 35]. These earlier studies have provided us with the opportunity to review and examine their experimental results. Their contributions were enlightening, giving insight on the need for exploring and developing CLIR for other major indigenous African languages like Afaan Oromo. Inspired by the earlier research works, we are interested in exploring and developing an experimental CLIR system that involves one of the major indigenous African languages (i.e., Afaan Oromo) and one of the most commonly used online languages (i.e., English).

1.5 Statement of the Problem

Nowadays, the volume of digital data generated in different languages and made available via the WWW and social media platforms is growing exponentially. However, the abundance of online content does not guarantee the accessibility and discoverability of relevant information. On the contrary, the vastness and diversity of online content often make information retrieval much more difficult and time-consuming. As pointed out by [11, 49], the greater the number of languages used on the Internet, the larger the volume of multilingual content on the Web, the more difficult it becomes to search and retrieve relevant information. Although the Internet has a suite of networking standards and protocols that have been widely used for data transmission and communication across national and regional borders, it does not have an efficient mechanism for searching and discovering relevant information across different languages and cultures. The task of making online resources efficiently accessible to culturally diverse and underserved populations is one of the major challenges that face IR researchers today.

Effective management of the ever-growing amount of multilingual data requires the development of more specialized IR systems that can efficiently search and retrieve valuable information, regardless of the language(s) in which it is written or represented in. Unfortunately, since most of the current major search engines and commercial IR systems are focused on a handful of better-resourced European and Asian languages, the need for developing CLIA systems for resource-scarce African languages has been left unaddressed for quite a long time. As described by [8, 7], without the support of robust CLIA systems, language barriers and cultural differences are insurmountable obstacles to the accessibility and usability information resources. In other words, without the development and application of robust CLIR systems, the task of searching and discovering relevant information across different languages is either impossible or must be performed in a foreign language that may not be understood by the majority of the users. As indicated by [20, 4], acute shortage of digital content available in resource-scarce African languages is one of the major factors that drives indigenous African communities to seek for more specialized IR system that can help them to find valuable information beyond language and cultural boundaries.

Due to lack of computational linguistic resources and translation tools, most under-resourced languages have not received very adequate attention from IR researchers over the last few decades. In spite of some linguistic studies that have recently been reported on indigenous African languages [31, 54, 20], the technological gap between well-resourced Western languages and poorly-resourced African languages still keeps widening. Owing to language barriers and linguistic digital divide [20], most speakers of resource-scarce African languages are facing serious challenges to find up-to-date information about critical matters such as education and healthcare services. Clearly, the challenges posed by language barriers and linguistic digital divide have continued to threaten and undermine the promise of the digital revolution. While online information resources are abundantly available and easily accessible to speakers of well-resourced European and Asian languages [11], the need to provide native speakers of African languages with localized digital content and information retrieval tools has been left unaddressed for many years. The problem of information poverty, which most African communities are facing in their daily lives, may persist or even worsen unless appropriate linguistic resources and information access technologies are immediately developed for African languages.

As noted by [3, 11, 16], building CLIA for resource-scarce language is an extremely difficult task for several reasons including:

- Lack of adequate documentation and linguistic description,
- Absence of language resources and annotated linguistic data,
- Lack of reliable multilingual lexicons and translation tools,
- Lack of morphological and semantic analysers.

Currently, most indigenous African languages, including major Cushitic languages like Afaan Oromo and Sidama, do not have basic computational resources and IR tools required for building CLIA [55]. Core linguistic resources and language processing tools such as stemmer, lemmatizer, part-of-speech tagger and morphological analyser are not readily available for resource-scarce African languages. Hence, the task of building CLIR for indigenous African languages is extremely difficult, prohibitively expensive and time-consuming. This task gets even more complex and challenging when the language pairs do not belong to the same language group and do not share similar morphosyntactic features, which is the case for English and Afaan Oromo. As described in more detail section 2.1 and section 2.2, English and Afaan Oromo are such pair of languages that happen to belong to different language groups and apparently have quite different morphological processes and syntactic structures.

In CLIR, unlike a monolingual IR, since the documents are represented in a language different from the language of the user's query, it is impossible to match a search query with documents directly. In other words, the problem of language barriers must be addressed and overcome by developing and deploying appropriate translation tools and techniques. The availability of language processing tools and translation resources such as bilingual dictionaries, part-of-speech tagger, morphological analyser and stemmer is, therefore, crucial for successful development of CLIR. Unfortunately, these essential linguistic resources are not available for severely under-resourced African languages like Afaan Oromo.

As described in more detail in section 3.2, Afaan Oromo has very rich and complex morphological structures. Unlike English, which is often considered as a moderately analytic language, Afaan Oromo is a highly inflective and agglutinative language with a relatively more flexible word order. The morphological complexity of Afaan Oromo, combined with the scarcity of computational resources, creates a serious challenge to the development of CLIR. As noted by [2], one of the major problems in developing CLIR for morphologically rich languages like Afaan Oromo is word form variations. A given stem (or lexeme) may occur in various inflected or derived forms. In a dictionary-based CLIR, the entire process of query translation may fail because of inflectional and derivational affixes. In order for the query translation to succeed, word form variants must be identified, stemmed and normalized. The need for normalizing word form variations is very critical for morphologically very rich and complex languages like Afaan Oromo [44, 56].

In summary, this study seeks to explore and investigate the following basic research questions:

- What are the major constraints in designing and developing CLIR for resource-scarce African languages?
- Which CLIR approaches and techniques are feasible for severely under-resourced African languages like Afaan Oromo?
- Which linguistic resources and tools are crucial for the development of OMEN-CLIR? How they could be easily identified, constructed and implemented?
- What are the major morphological processes in Afaan Oromo? How do they affect the development and effectiveness of OMEN-CLIR?
- Is a rule-based stemmer a realistic approach to deal with the inflectional morphology of Afaan Oromo? What is the impact of such a stemmer on the performance of OMEN-CLIR?
- How well does OMEN-CLIR perform at well-recognized international evaluation forums like CLEF?
- What are the major challenges encountered in developing OMEN-CLIR and how they could be dealt with?

Answers to these basic research questions have been explored, examined and discussed in this study.

1.6 Goal and Objectives of the Study

The goal of this study is to develop and evaluate an experimental Oromo-English CLIR that is designed to make effective use of limited linguistic resources. In addressing this goal, this study seeks to:

- Explore the viability of building a CLIR for severely under-resourced African languages like Afaan Oromo;
- Identify core linguistic resources and translation tools necessary in developing CLIR for resource-scarce languages;
- Understand and model the major morphological processes in Afaan Oromo;
- Review literature pertaining to the development of CLIR for resource-scarce languages;

- Build a machine-readable bilingual dictionary that can be adopted and used as a main source of knowledge for query translation;
- Design and construct a rule-based Afaan Oromo stemmer that can efficiently identify and normalize word form variations;
- Construct a list of Afaan Oromo stopwords and suffixes;
- Assess the impacts of Afaan Oromo stemmer on the performance of OMEN-CLIR;
- Evaluate the performance of OMEN-CLIR at well-recognized international evaluation forums like CLEF campaign;
- Identify the major challenges encountered in developing CLIR for resource-scarce African languages and suggest potential remedies that could be explored further.

1.7 Contributions of the Study

Given the ever-growing volume of multilingual content on the Web, modern IR systems are expected to provide the users with an efficient cross-language search services. Unfortunately, most of the current major search engines and commercial IR systems are primarily focused on a handful of well-resourced European languages like English and French [11]. As pointed out by [57], although major search engines and commercial IR systems can use the top 10 or 15 well-resourced languages to reach about 90% of their online customers, thousands of languages are needed to reach the vast majority of the population in developing countries, a demographic that has not only been marginalized and underserved in cyberspace, but also in urgent need of online information resources and services.

Nowadays [58, 59], the ability to access and share information beyond language and cultural boundaries has become more and more critical for effective communication and global collaboration. In this regard, CLIR have proved to be a crucial tool in helping and enabling users to search and discover valuable information beyond language and cultural boundaries [2, 7]. CLIR allows users to express their information needs in their native language and thereby takes care of identifying and retrieving relevant documents across different languages [60]. In addition to Web-based cross-language search and retrieval, some of the major areas where the application of CLIR has become increasingly more important include: e-governance, e-learning, publishing industries and digital libraries, travel and tourism services, global trading and marketing businesses, international banking and investment, healthcare and medical services [7, 59]. In this study, we have developed and evaluated an experimental Oromo-English CLIR that is designed to make effective use of limited linguistic resources, with a view to enabling Afaan Oromo speakers to search and retrieve valuable information beyond language and cultural boundaries.

As indicated in section 1.3, the performance of CLIR is largely depends upon the quantity and the quality of linguistic resources and translation tools employed for its development. One of the major problems encountered in building a CLIR for under-resourced languages is unavailability of reliable language processing tools and translation resources. Core linguistic resources such as parallel text corpora, bilingual dictionaries, stemmer and lemmatizer are not readily available for the majority of resource-scarce African languages [61]. In fact, it may take a considerable amount of time before sufficient amounts of linguistic resources and translation tools will be made readily available for indigenous African languages. This is especially true in the case of severely under-resourced languages like Afaan Oromo. In this study, we have focused on exploring and building basic linguistic resources and IR tools necessary for developing OMEN-CLIR. Some of the important linguistic resources and translation tools that we have designed and developed during the course of this study include a machine-readable bilingual dictionary, Afaan Oromo stemmer, list of morphological affixes and stop-words. While our machine-readable Oromo-English dictionary has been adapted and used as a main source of knowledge for query translation, our stemmer has played a key role in identifying and removing inflectional affixes of Afaan Oromo. In summary, some of the major contributions of this study include:

1. **Pioneers the development of Oromo-English CLIR**: Although several research reports are available on development and application of CLIR for major European and Asian languages, no comprehensive and formal study has so far been reported on the development and application of CLIR for severely under-resourced African languages like Afaan Oromo. For instance, there is no a CLIR system that can support major Cushitic family languages like Afaan Oromo and Sidama. To the best of our knowledge, no detailed study has so far been reported on the development and evaluation of a dictionary-based Oromo-English CLIR. In this sense, this study pioneers the development of Oromo-English CLIR. We believe that this study will serve as a foundation for designing and developing CLIR for other resource-scarce African languages. We also hope that the evaluation results of our study will motivate and encourage other researchers to conduct similar studies for other major indigenous African languages.

2. **Provides a detailed review of literature on CLIR**: During the course of this study, literatures on the development and evaluation of CLIR has been thoroughly surveyed and extensively reviewed. Core linguistic resources and translation tools necessary for the development of CLIR were carefully examined and analysed. Reference materials on morphological characteristics of Afaan Oromo were also thoroughly explored and reviewed with a view to understand the major morphosyntactic structures of the language [21, 22, 62, 63]. Research works reported on the development and application of CLIR for under-resourced African languages were also thoroughly reviewed and examined to learn from their earlier experiments and findings [15, 14, 64, 65, 66, 67].

3. **Reduces the knowledge gap in developing CLIR for African languages**: Over the last two decades, we have witnessed a significant progress in the development and evaluation of cross-language information access systems for well-resourced Asian and European languages. On the contrary, the need for designing and developing CLIR for resource-scarce African languages has been left unaddressed for quite a long time. Presently, little is known about the challenges and benefits of extending the application of CLIA systems to severely under-resourced African languages like Afaan Oromo. Like many other resource-scarce languages, Afaan Oromo is one of the least researched and poorly understood African languages, particularly from the perspective of computational linguistics and information access technologies. This study seeks to reduce this knowledge gap by exploring and developing an experimental Oromo-English CLIR. We hope that this study expands the empirical base of knowledge on the development and application of CLIR for resource-scarce African languages.

4. **Designing and building a general architecture of OMEN-CLIR**: As noted by [11], most CLIR approaches and techniques have been designed with the assumption that core linguistic resources and translation tools such as parallel text corpora, bilingual dictionary, MT system, part-of-speech-tagger, stemmer and morphological analyser are readily available and accessible to the researchers. Unfortunately, in the case of severely under-resourced African languages like Afaan Oromo, basic computational linguistic resources and automated translation tools are either not available at all or are only available in insufficient quantity and very poor quality. In this study, as discussed in section 5.1 in more detail, we have designed and built a general CLIR architecture that

can work with limited linguistic resources and translation tools such as bilingual dictionary, stemmer, stop-words and query processing modules. This general architecture can be extended to support other closely related indigenous African languages such as Somali and Sidama with minor modification and incorporation of language specific resources.

5. **Exploration and understanding of the morphological processes of Afaan Oromo:** Detailed understanding of morphological processes and inflectional paradigms is crucial in designing appropriate computational models and tools such as stemmer and lemmatizer. After detailed review of various reference materials related to Afaan Oromo, we have identified the inflectional and derivational affixes that are commonly used in the language. In consultation with Afaan Oromo linguists and language experts, we have also tried to build a general computational model of Afaan Oromo inflectional morphology.

6. **Construction of a bilingual machine-readable dictionary:** A translation resource is one of the essential prerequisites and components of building multilingual information access systems like CLIR. Unlike monolingual IR, in CLIR, it is necessary to ensure that the users can access and retrieve relevant information regardless of the language(s) used to express the search request. In other words, since the target documents and the user's request do not share the same language, bridging the language gap between the query and the documents is one of the major challenges that must be addressed by CLIR researchers. As described in more detail in chapter 4, there are different translation resources and approaches that can be adopted in implementing CLIR including MT system, annotated parallel corpora and multilingual or bilingual dictionaries. While such advanced linguistic resources and translation tools are readily available for well-resourced European and Asian languages, they are rarely available for the majority of indigenous African languages. The task of designing and building basic translation resources such as a machine-readable bilingual dictionary has, therefore, become very critical for the development of OMEN-CLIR. During the course of this study, besides building and adopting a machine-readable bilingual dictionary from a printed copy of a human-readable dictionary, we have tried to enhance its coverage by incorporating additional vocabularies and translations from various lexical resources and glossaries. Accordingly, we have been able to build and contribute a medium-size Oromo-English dictionary, which has been used as a main source of knowledge for query translation.

7. **Designing and application of Afaan Oromo stemmer:** As noted by [61, 2], in morphologically rich and complex languages like Afaan Oromo, word form variations pose a serious challenge for the development of CLIR. In the context of a dictionary-based CLIR, a search term in a given query may fail to match with a semantically equivalent dictionary entry because of its inflectional and derivational affixes. In other words, the entire process of query translation may fail due to inflectional and derivational affixes associated with the search terms. The task of identifying and normalizing word form variants has, therefore, become very essential. To this end, different stemming algorithms have been developed and widely used for many European and Asian languages [56, 68]. Although a number of previous studies have confirmed that the application stemming algorithm is very useful for morphologically very rich and inflectional languages like Afaan Oromo, little is known about the role and impact of such stemmer on the performance of CLIR that involves resource-scarce African languages. Hence, the task of designing and evaluating Afaan Oromo stemmer has become one of the major components of our study. We have constructed and adapted a rule-based stemmer that focuses on identifying and normalizing inflectional affixes. Our evaluation experiment on the impact of the stemmer has shown that the application of the stemmer has improved the performance of OMEN-CLIR substantially (by 30% to 50% average precision).

8. **Evaluation of OMEN-CLIR at a well-recognized international evaluation forum:** Performance evaluation is essential for sustainable development and improvement of CLIR. Apart from designing and building our Oromo-English CLIR, for which most of the basic linguistic resources and translation tools have been designed and developed from scratch, another major contribution of this study is to assess and determine the performance of the proposed retrieval system. With a view to assess and improve the performances of OMEN-CLIR, we had participated in one of the well-recognized international CLIR evaluation forum over the last couple of years**, (**i.e. CLEF-2006 and CLEF-2007 campaigns). The main focus of our evaluation experiments in the two different CLEF annual campaigns was to assess and determine how well OMEN-CLIR can perform at one of the well-recognized international competition forums like the ad-hoc track of the CLEF campaign [39, 40]. We have conducted a series of retrieval experiments based on the general guidelines and procedures designed for evaluating the

performances of a bilingual CLIR. We have used two different datasets distributed by the ad-hoc track of the CLEF campaign. Besides a number of official runs that had we had submitted to the ad-hoc track of the CLEF campaigns, we have conducted various additional retrieval experiments in order to test and improve the performances of OMEN-CLIR. Overall, we found the experimental results very promising and encouraging, given the disparity of the languages involved and the limited amount linguistic resources that have been employed to implement OMEN-CLIR. In one of our official retrieval experiments in which we have used our Afaan Oromo stemmer, our CLIR system has achieved an average mean precision (AMP) of 29.90%, which is about 67.95% of a monolingual baseline.

1.8 Organization of the Thesis

The rest of this thesis is structured and organized as follows. Chapter 2 presents an introduction to the genetic and typological classification of African languages along with some of the major challenges and opportunities in developing CLIA for resource-scarce African languages. A summarized review of related studies is also presented towards the end of this chapter. Chapter 3 introduces and describes the basic structures and characteristics of Afaan Oromo morphology. It describes some of the major linguistic phenomena and peculiarities of the language with special focus on nominal and verbal morphology. Moreover, an FST-based computational model of Afaan Oromo inflectional affixes has been proposed, designed and illustrated with various examples. Some of the basic issues and approaches involved in CLIR, including MT-based, corpus-based and dictionary-based query techniques are reviewed and described in chapter 4. The architecture and major components of OMEN-CLIR are introduced and described in chapter 5. The building blocks of OMEN-CLIR, including the construction of Oromo-English dictionary and Afaan Oromo stemmer are described in detail in this chapter. Chapter 6 focuses on the evaluation of OMEN-CLIR and error analysis. It presents the results of evaluation experiments that we have conducted in order to improve the performance of OMEN-CLIR. Finally, chapter 7 presents summary of key findings and draws conclusions. Important recommendations and future directions of the study are also summarized and suggested towards the end of this chapter.

2 Linguistic Diversity in Africa: Challenges and Opportunities for Developing CLIA

As indicated by [5, 69, 70], one of the salient by-products of the digital revolution is the development of HLT, which is concerned with designing and building computer programs that can automatially process, analyse and recognize human languages. As one of the major component of HLT, the main objective of CLIA is to design and develop information systems that can automatically process, index, access and retrieve information across different languages. The development of CLIA systems for resource-scarce African languages is of paramount importance to accelerate the accessibility and usability of online resources and services. More specifically, the development of CLIA for indigenous African language is vital to enable African communities to use their indigenous languages in accessing and retrieving valuable information on the Web.

This chapter introduces genetic and typological diversity of African languages along with some of the major challenges and opportunities for developing CLIA. A review of related studies is also summarized and presented towards the end of this chapter. In this study, we use the term *African language technology* in its narrower sense, as defined by [71]. It is used to refer to the development of natural language processing and information access technologies for sub-Saharan African languages, especially from the four major African language groups: Afroasiatic, Khoesan, Nilosaharan and Niger-Congo languages, (see section 2.1). Consequently, we will not consider some African variants of European languages as well as Arabic since they are already well researched and documented by many linguists and HLT developers.

2.1 Genetic Classification of African languages

According to [72], there are thousands of languages on the planet, all descended from the earlier (or proto) languages that spread and changed and split up into dialects as people moved and spread around the world. Given enough time, the separation of groups and the dialects they speak inevitably leads to the birth of new languages, the way French, Romanian, and Spanish grew out of the Latin spoken by the Romans. In linguistics, the task of classifying languages into different groups or types could be based on different properties and criteria including genetic, historical, typological and areal. Languages are often categorized and compared to one another based on genetic and typological classifications. On one hand, genealogists argue that languages can be categorized into different families or groups in which a community of origin is distinctly traceable. On the other hand, typologists argue that a set of languages that are not (or not necessarily) genetically related to one another may share common linguistic properties such as morphological processes and syntactic structures. In this section and the next section, we will describe some of the major genealogical classification and typological features of indigenous African languages. Although many linguists agree on the classification of languages as well as on its significance in studying human languages, there are some differences as to how language groups should be determined or how language families should be identified and assigned. In this section, we present the major classification of African languages with special focus on the four major groups [54].

As indicated by [73], while linguistic and cultural diversity is a norm, multilingualism is commonplace in Africa. In fact, linguistic and cultural diversity is common in most parts of the world including Asia and Europe. According to [46, 51], today there are over 6,900 living languages in the world, about 60% of which are located in Asia and Africa. While more than 2000 languages are found in Africa, at least 100 of them have over one million native speakers [30]. Being home to about one-third of the world's languages, Africa is the second most culturally and linguistically diverse continent on the planet. From country to country, African nations are characterized by very rich cultural and linguistic diversity. Indeed, Africa is one of the most rich and complex part of the world in terms of the number of languages, the size of the communities speaking them, and the area each language covers. Linguistic diversity and multiculturalism holds true for many African countries including Ethiopia, which is home to several indigenous languages and multicultural communities.

The term *indigenous African language,* which is geographic rather than linguistic classification of languages, is often used to refer to the major groups and families of languages that are native to Africa. Usually, it is used to distinguish Sub-Saharan African languages from non-native and foreign languages that have also been used in the continent. As noted by [4], most of African countries south of the Sahara do not have a single majority language. They are rather characterized by having scores or even hundreds of different languages. Although some of the major European languages introduced during the colonization era are still in use in many African countries, these foreign languages are spoken as second or third language by urbanized and educated segments of the African communities. Since most foreign languages are not spoken by the majority of African population, they do not have strong connection with social values and cultural heritages of the indigenous communities.

According to [74], among about 6900 languages that are spoken throughout the world today:

- more than 2000 languages are found in Africa,
- about 1000 are found in the Americas,
- more than 2250 are found in Asia,
- about 220 are found in Europe, and,
- about 300 are found in Australia and the Pacific.

These languages of the world can be grouped into about 90 major language families [89]. A *language family* is defined as a group of languages with a common origin or a set of languages that can be shown to have common ancestry. Accordingly, Africa is known for being a home to the world's largest language phylum, Niger-Congo. Some of the major language families that have been identified by linguists around the world include [74]:

- Afro-Asiatic (about 353 languages spoken in Africa and Asia),
- Austronesian (about 246 languages spoken in Asia and Oceania),
- Indo-European (about 430 languages spoken in Asia and Europe, and in European settlements in other parts of the world),
- Niger-Congo (about 1495 languages spoken in Africa),
- Sino-Tibetan (about 399 languages spoken in Asia),
- Trans-New Guinea (about 561 languages spoken in New Guinea and adjacent islands).

As indicated in Figure 2.1 and in Table 2.1, the language families of Africa can broadly be divided into four major groups [54]: Afro-Asiatic, Nilo-Saharan, Niger-Congo and Khoisan. For each of these four language phylum, Table 2.1 shows major subdivisions and a rough estimate of the number of languages belonging to the sub-phylum. Note that since Afro-Asiatic languages are not found exclusively in Africa, Figure 2.1 represents only those Afro-Asiatic languages spoken in Africa. Below, we present a brief description of each of these four major groups.

Classification (Phyla)	**Major Subdivision (Sub-Phylum)**	**Number of Languages**
Niger-Congo	Atlantic-Congo	1448
	Mande	73
	Kordofanian	23
Afro-Asiatic	Chadic	194
	Semitic	78
	Cushitic	45
	Berber	26
Nilo-Saharan	Eastern Sudanic	98
	Central Sudanic	65
	Saharan	9
Khoisan	Southern Africa	26
	Hatsa	1
	Sandawe	1

Table 2.1 Major language families of Africa

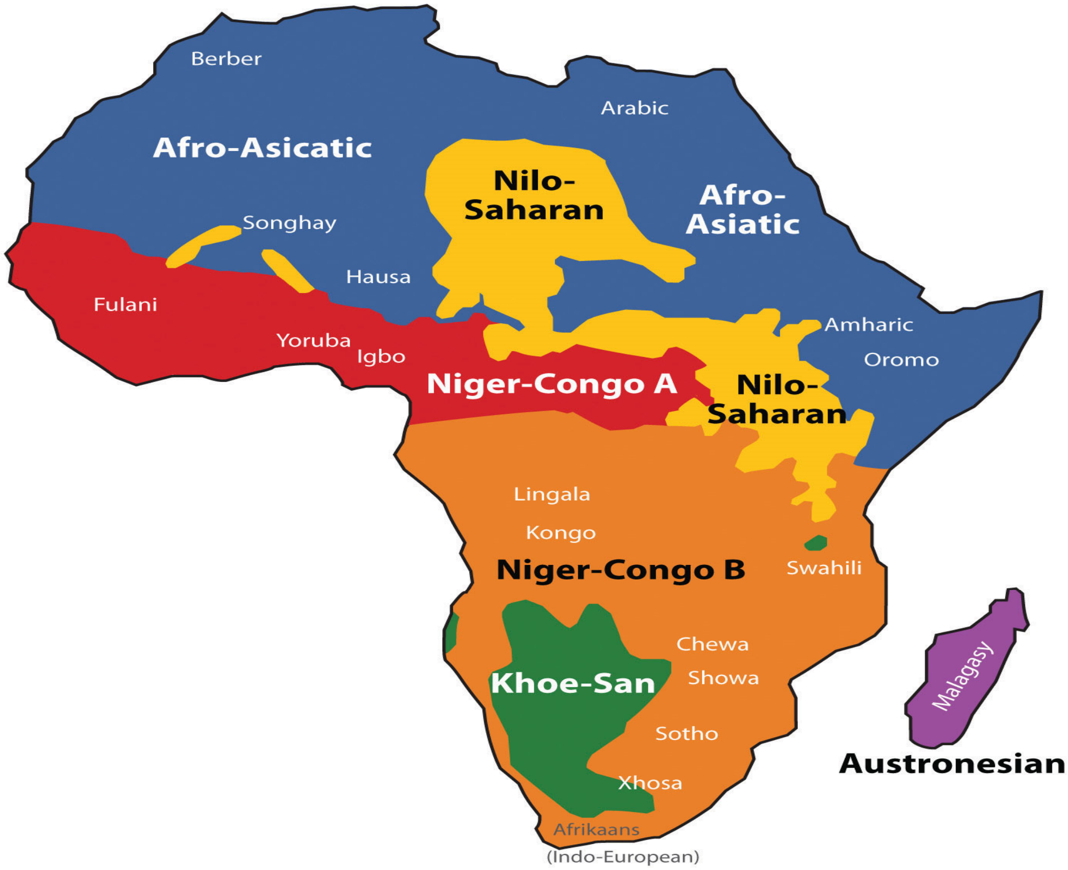

Figure 2.1 Major language families of Africa

1. **Afro-Asiatic:** The Afro-Asiatic, (formerly Hamito-Semitic), group constitutes most of the languages spoken in Northern and Eastern Africa. It is one of the major language families with about 370 living languages and more than 350 million speakers spread throughout North Africa, the Horn of Africa, and Southwest Asia. Some of the Afro-Asiatic languages are also spoken in central and Western Africa. The Afro-Asiatic group can be further divided into the following six major branches: Berber, Chadic, Cushitic, Egyptian, Omotic and Semitic. As indicated in section 3, Afaan Oromo belongs to the Cushitic branch of the Afro-Asiatic languages together with Afar, Somali and Sidama.

2. **Nilo-Saharan:** The Nilo-Saharan (including the East African Masai and Luo) is another major family of African languages. Some of the major languages in this group include Luo, Kalenjin and Kanuri. They are mainly spoken in the upper parts of the Chari and Nile rivers, including Nubia.

3. **Niger-Congo:** As indicated earlier, Africa is home to the world’s largest language phylum, Niger-Congo. It is considered as the largest language family of Africa as well as the world in terms of its geographical area, number of speakers, and the number of its distinct languages. It is enormous language branch whose subgroups is found throughout Southern and central Africa as well as in most Western parts of the continent below the Sahara (see Figure 2.1). A number of widely spoken languages of Sub-Saharan Africa belong to this group. While Swahili is considered as the most widely spoken language by total number of speakers, Yoruba, Igbo, Fula and Shona are considered as the most widely spoken languages in terms of the number of native speakers.

4. **Khoisan:** The Khoisan language family (also known as the Khoesan) constitutes many of the languages in the South-eastern Africa. Although they are currently restricted to the regions around the Kalahari Desert, (primarily in Namibia and Botswana), the Khoisan languages might have spread throughout Southern and Eastern Africa prior to the Bantu expansion. Like many other indigenous African languages, most of the languages in this family are considered as either endangered, since they do not have reliable written records. Some of the major Khoisan languages include Nama of Namibia, Sandawe in Tanzania and the Juu language cluster of the northern Kalahari.

Besides the above four major language families, there are some other languages like Malagasy and Afrikaans, which are widely spoken in southern parts of the continent. Malagasy which belongs to Malayo-Polynesian or Austronesian languages family is widely spoken by several people in Madagascar while Afrikaans is used in South Africa. It is important to note that since the linguistic history of Africa is very complex and has not yet been well understood, the above four broad classifications could not be considered as a comprehensive and exhaustive genealogical classification of African languages. In general, linguistic and cultural diversity is very common and the norm in many parts of Africa including the Horn of Africa. For example, it is not unlikely that a typical citizen of Ethiopia will speak at least one or two indigenous languages in addition to the official language (i.e., Amharic). This implies that many African nations and communities need a reliable CLIR system that allows them to access and identify valuable information across multiple languages.

2.2 Typology of languages in Africa

Apart from the genetic classification described in the foregoing section, languages can be categorized into different groups based on their internal structures and linguistic features. A number of linguists and language developers have long been interested in investigating and classifying the structural properties of the languages around the world. Broadly speaking, *linguistic typology* can be understood as a comparative study of the grammatical structures of human languages. It is a subfield of linguistics that focuses on classifying and describing human languages based on their structural and functional features [74, 75, 76]. Typology may involve any structural aspect of a language including phonology, morphology and syntax. The main purpose of typology is not only to identify common linguistic properties, but also to describe the structural variation of languages across the world. In contrast to the study of *linguistic universals*, which is concerned with what human languages have in common, typology tries to investigate and describe the ways in which languages differ from each other [74]. Expressed differently, while the study of typology aims to classify and describe languages according to their linguistic properties and grammatical categories, linguistic universals focuses on identifying basic and fundamental features that are common to all human languages.

As indicated above, typologically, languages can be classified into different groups with respect to their different linguistic properties such as morphological processes, syntactic structures and grammatical categories. In this section, two major typologically significant features of African languages, (i.e., morphological structures and basic word order), are briefly reviewed and presented.

2.2.1 Morphological typology

There are a number of morphological processes that a language might employ in order to build or form words, such as affixation, compounding, reduplication, alternation, and suppletion. While many languages make use of some or most of these morphological processes, others make use a just few or none of these processes. In some languages, most words are composed of a single morpheme, where in others words consist of two or more morphemes. *Morphological typology* is a subfield of linguistic typology that classifies and describes languages on the basis of how they build or form words. It is a method of classifying and describing languages according to their morphological structures or processes. Most linguists often distinguish between two main types of morphological structures: analytic languages from synthetic languages, though the latter has further subdivisions [74, 77], (see Figure 2.2 and Figure 2.3). The morpheme-per-word ratio defines where the language lies in the isolating-synthetic scale, a ratio of one meaning purely isolating language. In this section, we will describe the morphological typology of African languages with their implications for developing and implementing CLIA.

- **Analytic languages**: As noted by [74], languages in which a word tends to consist of only one morpheme (or a very few combined morphemes) are called analytic or isolating. The number and use of bound morphemes in analytic languages is minimal since they do not use affixes to compose words or to express grammatical categories and relations. Syntactic information and grammatical relations in analytic languages are expressed by using adpositions or separate auxiliary words. Since sentences are often made up of free morphemes (or uninflected words), documents written in analytic languages have a very low ratio of morphemes to words. In isolating languages like Afrikaans and Yoruba, the ratio of morphemes to words is nearly one-to-one. As noted by [77, 78], grammatical relations in analytic languages are often conveyed by using word order and function words such as prepositions and auxiliaries. For instance [74, 79], in analytic languages like Afrikaans and English, syntactic information and grammatical categories are indicated by word order (like inversion of verb and subject for interrogative sentences) or by using separate function words instead of morphological affixes that are commonly used in synthetic languages like Afaan Oromo. In analytic languages, context and syntax play more important role than morphological processes [78].

  Due to the abundance of function words and unbound grammatical elements, the task of identifying and removing stop-words is important in developing IR for isolating languages than morphological processing tools like lemmatizer and stemmer. Texts or queries expressed in analytic languages may not require detailed morphological analysis and stemming procedures. In multilingual information retrieval environments like CLIR, it is easy to translated queries expressed in an analytic language into another target language by employing lexical resources such as bilingual dictionaries. Moreover, simple morphological structures of isolated languages can help to counter data sparseness, which poses a serious problem in statistical natural language processing for synthetic languages. Analytic languages are primarily found in East and Southeast Asia (e.g. Chinese, Vietnamese), as well as West Africa (e.g. Yoruba) and South Africa (e.g. Kung, which is also known as Kung-ekoka or Xu) [54, 73, 74].

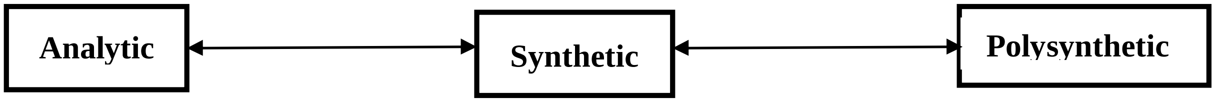

Figure 2.2 Analytic vs. Synthetic

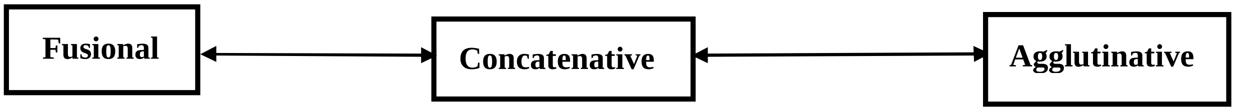

Figure 2.3 Fusional vs. Agglutinative

- **Synthetic languages**: Languages in which a word tends to consist of more than two morphemes are called synthetic [74]. Synthetic languages build words by adding or affixing a series of bound morphemes to a given root word or stem. In synthetic languages, since morphological affixes are frequently used to express syntactic information and grammatical relations, most words tends to have multiple grammatical affixes. On the other hand, word order and syntactic structures are more flexible and less important in synthetic languages like Afaan Oromo than they are in analytic languages like Afrikaans. Languages in which words are composed of many morphemes and complex morphological structures are sometimes called polysynthetic languages. In a polysynthetic language a word may contain multiple morphemes equivalent to both verb and noun in English, which means that an English sentence can be expressed by a single word. For example an inflected Afaan Oromo word: "*deemaniru*" stands for the English sentence "They have gone." While most European languages including English are considered as moderately synthetic or fusional, most African languages including Afaan Oromo are considered as highly synthetic.

As noted by [74], synthetic languages can be further divided into agglutinative and fusional languages based on whether their morphemes are clearly differentiable or not (see Figure 2.3). A synthetic language is purely agglutinative, if the boundaries between

> morphs are clear, and it is fusional, if the morphs are overlaid in a way so that they are difficult to segment. The degree of fusion determines where the language falls in the agglutinative-fusional scale. Though agglutinative languages often build words by a series of affixes (by stacking them one after the other), there is a correspondence between the grammatical meaning and bound morphemes. Broadly speaking, words in agglutinative languages may have several bound morphemes that are easily distinguishable and segmentable. In highly agglutinative languages like Turkish, affixes are not only transparent and distinguishable from the root words, but also among themselves. To be more specific, each morpheme often represents one grammatical meaning and the boundaries between the morphemes are easily demarcated; that is, the bound morphemes are affixes, and they can be individually identified, interpreted or analysed. Unlike fusional languages, agglutinative languages tend to have a number of morphemes per word, and their morphology tends to be regular, with certain notable exceptions. This is not the case in fusional languages, where bound morphemes can be fused with stem or among themselves and multiple pieces of grammatical information may potentially be packed into a single morpheme.

It is important to note that the differences between analytic and synthetic languages as well as agglutinative and fusional languages cannot be clearly defined. In fact, there is a cline or gradient of structural variation. As shown in Figure 2.2, while the ideal types may be placed at either end of the continuum scale, the vast majority human languages are located at different points along the cline. As indicated by [74, 77], the problem of classifying languages according to the familiar morphological typology of isolating, synthetic, agglutinating and inflectional has occupied linguists for many years. As shown in Figure 2.2 and Figure 2.3, this involves determining a language's place along a continuous scale or cline: from analytic to polysynthetic or from agglutinative to fusional, and vice versa. The continuum from isolating to polysynthetic focuses on the number of morphemes per word, (an isolating language having, ideally, one morpheme per word and a polysynthetic language having multiple and complex morphemes per word). As indicated in Figure 2.3, the agglutinative to fusional continuum focuses on the extent to which there are clear boundaries between morphemes within a word. While the boundaries between morphemes are easily demarcated or distinguishable in an agglutinating language, a fusional language lacks clear boundaries between morphemes or affixes. As noted by [77], the two main measures for determining the level of agglutination versus fusion are invariance of the morphemes and the segmentability of the morphemes. The closer a language is to the agglutinating end of the continuum, the more invariant and easily segmentable the morphemes will be. Languages closer to the fusional end have morphemes with more morphophonemic variation and less segmentability.

2.2.2 Basic Word Order

In addition to morphological systems, typologists often divide languages into different categories according to "*basic word order*", which is often based on the order of basic constituents in a declarative sentence. Elaborating this concept further, [78] pointed out that the basic order of constituents in a language is typically defined by the position of subject (S), verb (V) and object (O) in declarative sentence or main clause, which are often abbreviated and labelled as SVO, SOV, VSO, etc. Accordingly, while the vast majority of the world's languages fall into one of three groups: SOV, SVO and VSO, less than 5% of the world's languages belong to one of the three remaining possible types: VOS, OVS and OSV [74].

According to [79, 76, 78], the proportion of languages with SVO constituent order is much higher in Africa than globally; it is the sentential order of approximately 71% of African languages. SVO languages are common in the four major African language families described in section 2.1. Regarding this, [74] stated that a noteworthy feature of Sub-Saharan Africa is the predominance of SVO word order. The majority of the Niger-Congo languages are SVO. In fact, this constituent order is almost without exception in the West Atlantic and Bantu branches. SVO languages in the Nilo-Saharan phylum include some Central Sudanic and Western Nilotic languages. Relatively, SOV constituent order is less common among African languages than worldwide. [78]. There are only a few SOV languages in Niger-Congo, e.g. the Ijoid and Dogon languages, and the Kordofanian language Tegem. Kanuri, Maba, Kunama and the Nubian languages are prominent examples of SOV languages in the Nilo-Saharan phylum. The Ethio-Semitic, Omotic and Cushitic languages of Afro-Asiatic, including Afaan Oromo, are considered as SOV or verb-final languages, a word order that is quite common cross-linguistically, but somewhat less common in Africa. In Modern Standard Arabic, the basic word order is verb-subject-object (VSO), a word order which is considerably less common than SVO or SOV word orders in Africa [74, 78].

Even though the basic sentence structure of Afaan Oromo follows SOV format, it allows greater variations in word order. Since it is a pro-drop language, the subject of a sentence may not overtly expressed in conversations. Moreover, Oromo adjectives usually follow the noun, not precede it as in English. Adjectives are considered as noun phrase modifiers since they agree with the head noun in terms of their number and gender. Adverbs that modify adjectives go before the adjective. Generally, indirect objects follow direct objects in Afaan Oromo. Although it has both prepositions and postpositions, the latter is more common in Afaan Oromo.

2.2.3 Implications of linguistic typology for CLIR

Since most languages cannot be completely fitted into a specific typological category, many linguists, including [54, 74], have pointed out that typological classification is theoretical rather than practical or empirical in nature. In reality, each of the above morphological and syntactic classifications, (see section 2.2.1 and section 2.2.2), are blurred and overlapping, as they do not exist in a pure state. For instance, most indigenous African languages, including Afaan Oromo, may fit into one or another category, since they have mixed types of morphological processes and syntactic structures. In comparison with highly analytic languages like Afrikaans, English may be considered as a less analytic or moderately synthetic language, as it has a very productive derivational morphology. However, English is considered as an analytic language when it is compared with highly synthetic African languages like Afaan Oromo, Amharic and Swahili, which have very rich and complex inflectional and derivational morphology. As a result, it is important to treat the typological classifications discussed in the foregoing two sections as continuous and relative rather than absolute.

On the other hand, understanding the basic typological properties (morphological processes and syntactic structures) of African language is important for the development and application of CLIR. A detailed understanding of the morphological processes and inflectional paradigms in African language is crucial in designing appropriate computational models and IR tools such as stemmer and lemmatizer. For instance, in contrast to analytic languages, queries expressed in synthetic languages would require detailed morphological analysis and normalization procedures. Understanding the morphological structures of synthetic languages like Afaan Oromo is very important to decompose the agglutinative affixes into their constituent bound morphemes.

As noted by [62], in morphologically rich and complex languages like Afaan Oromo, word form variations pose a serious challenge for the development and effectiveness of CLIR. In the context of a dictionary-based CLIR, a search term in a given query may fail to match with a semantically equivalent dictionary entry because of inflectional and derivational affixes associated with it. In other words, the entire process of translating a source language query into a target language query may fail due to inflectional and derivational affixes attached to the search terms. The task of identifying and normalizing word form variants is, therefore, very essential for synthetic and inflectional languages like Afaan Oromo.

As it is described in more detail in the next chapter, the morphological structure of Afaan Oromo presents a huge challenge to the effectiveness of CLIR because most words are composed of a stem and multiple affixes. In multilingual information retrieval environments like CLIR, since queries expressed in a synthetic language may contain many inflected words, it is difficult to identify and determine the stem or base form of a search term without the application of stemmer or lemmatizer. On the other hand, synthetic morphological structures are often considered as the major cause of data sparseness problem, which poses a huge challenge for language modelling in morphologically very rich and complex languages like Afaan Oromo. Most resource-scarce African languages suffer from data sparsity and out-of-vocabulary words problems. The sparseness of data often affects the performance of many data-driven natural language processing applications including CLIR and MT systems. In order to counter the data sparseness problem in synthetic languages, a stemmer or morphological parser can be employed to identify and normalize word form variations.

2.3 Development of CLIA for African Languages: Challenges and Opportunities

As indicated in [20, 4], the digital revolution is a pervasive and unavoidable fact of life for African nations, no less than for other nations, although some of the specific issues and problems faced by different countries may differ in scope or proportion. In the context where basic social and economic needs are not met and digital literacy rate is very low, it may appear to be a luxury to spend time and money on the development of computational linguistic resources and information access technologies for indigenous languages. To consider doing so, however, is not only an expression of hope and optimism, but an affirmation of the value and relevance of Africa's linguistic and cultural heritages [4]. Moreover, it is important to consider the developments of African language technologies as an integral part of the broader economic and social development strategies that are being planned and undertaken in the continent.

For a sustainable development to take root in Africa, the majority of the population must be involved and participated in social and economic development activities. It is important to recognize that the needs of the majority of indigenous African communities would be best served through their indigenous languages. Since indigenous languages are considered as the cornerstones of African culture and history, it is difficult to expect the vast majorities of African communities to learn and use dominant online languages like English while their own indigenous languages are severely underrepresented and neglected in the digital world. In other words, African languages must be investigated, revived and recognized as an indispensable instrument for sustainable development of African nations. The development of indigenous African languages must be recognized as a basis for the future empowerment of African peoples.

The advent of the Internet has presented Africa, a continent struggling with many aspects of its economic development, with an unparalleled opportunity to turn its vast human resources into economic and competitive advantages. The Internet has a huge potential to accelerate and transform the natural resource-based economy of Africa into a knowledge-based economy of the 21st century. In other words, effective use of the Internet and related digital technologies can facilitate the transition of African nations from an economy based on agriculture and natural resources, which are typically very scarce and thus cannot be easily reproduced and shared, to the one based on knowledge, where information can be digitally generated and disseminated to spur the economic growth of knowledge society. An increasing number of national and regional government agencies as well as non-profit and non-governmental organizations have already identified the Internet as an effective medium to accelerate the economic development of African nations.

As indicated by [40], users are more comfortable to express their information needs in their own native languages. For the majorities of indigenous African communities, their language of comfort to search and retrieve information is their indigenous languages. This should come as no surprise, as education and communication are generally easier in the first language than in languages that people acquire later [4, 20]. Emphasizing the benefits of extending the application of natural language processing and cross-language information access technologies to under-resourced languages, [15] pointed out that national and global excellence in the new millennium shall be measured by the extent to which the digital technologies in general, and the Internet in particular, can deliver their full potentials in languages familiar to indigenous communities in developing countries. The development of CLIR that is designed to support indigenous African languages is essential to enable African communities to use their own native language in searching and retrieving valuable information beyond the boundaries of language and cultures.

Although the task of developing and implementing CLIA for indigenous African languages is very difficult, expensive and time consuming, the benefits are enormous. Given the great cultural diversity of African nations and the large number of languages present in the continent, the development of CLIA for resource-scarce languages will help the majority of African communities to gain access to the vast wealth of information produced in better-resourced like English and French. In summary, the major reasons for developing CLIA for resource-scarce African languages may include the following.

- **To alleviate information poverty**: Due to language barriers and lack of suitable information access technologies, the majority of native speakers of indigenous African languages have long been denied the opportunity to share and contribute to online information resources and services. Today, owing to language barriers and linguistic digital divide, the majority of native speakers of African languages, including speakers of Afaan Oromo, are still suffering from information poverty, which is characterized by lack of reliable information sources and services. In many African countries, including Ethiopia, the challenges posed by language and cultural barriers are further aggravated by the acute shortage of digital content produced in indigenous African languages. As indicated by [4], the majority of native speakers of indigenous African languages are facing serious challenges to find reliable information about critical matters such as education, employment, healthcare, financial assistance and market prices. The abundance of information produced in English is not significant for the majority of African communities as long as they cannot search and retrieve relevant items. Delivering online information resources to the disadvantaged population of Africa requires the development of appropriate computational linguistic resources and information access technologies tailored to the specific needs of indigenous communities. African researchers must use the opportunities presented by the digital technologies to alleviate information poverty by tackling the problems of language barriers and linguistic digital divide.

- **To promote linguistic and cultural diversity**: Although linguistic and cultural diversity poses a huge challenge to information access and exchange, it should not be considered as a threat or bane that should be avoided in the digital age. Rather, it should be welcomed, embraced and promoted. In other words, linguistic diversity is not a threat per se, but, rather a challenge that can be managed to the benefit of the increasingly multilingual and multicultural global society. As noted by [4], Africa's linguistic diversity is not only a challenge, but also an asset that can be exploited for the development of language industry and online services such as social media networks, e-commerce, online recommendation systems and translation services. The linguistic diversity of Africa must be preserved and used as an instrument not only to demonstrate the cultural heritages and social values Africa, but also to accelerate the development and prosperity of African nations. In multicultural continent like Africa, multilingual communication is an integral part of social and economic activities. Hence, CLIA could be considered as a strategic tool for enhancing and speeding-up the economic growth of African nations. In this regard, the main focus of CLIA should be to build and provide efficient multilingual information access services to ensure the participation and competitiveness of African citizens in the digital age.

- **To revitalize indigenous and endangered African languages**: No matter how severely under-resourced and endangered they are, indigenous languages of Africa have been and will be one of its distinct and unique cultural heritages. Hence, there is a widespread concern among linguists and language technology developers that many indigenous

African languages may remain neglected and ignored in the digital world [20]. Indeed, the current state of resource-scarce African languages is matters of utmost concern. It is important to recognize that the survival of indigenous African languages largely depends on their ability to cope with modern communication technologies. Indigenous languages are a precious component of the cultural wealth of African communities and thus, they deserve future-proofing. The revitalization of resource-scarce African languages calls for not only deeper and comprehensive linguistic studies, but also for concerted research efforts to explore and develop core computational linguistic resources and tools. In particular, it calls for the development and application of more specialized information access systems with which native speakers of African languages will be able not only to access and discover information, but also contribute to the digitally accumulated knowledge repositories.

2.4 Related Works

In this section we present a review of related studies with special emphasis on research works on development of CLIR for under-resourced languages. As has been indicated earlier, CLIA is crucial to discover valuable information beyond language and cultural boundaries. During the last two decades, CLIR has been widely studied and developed for widely used European and East Asian languages [10, 19, 80, 17, 81, 82, 83]. According to [82], information access across languages challenges many interdisciplinary researchers and practitioners, including the developers of CLIR and MT systems. After a detailed description about the importance of CLIR in general, and the significance of Google’s cross-language search services in particular, the authors have proposed and suggested various strategies that digital libraries can apply for developing and implementing multilingual information access system. They identified and proposed the strategies based on a case study they had conducted in five different multilingual digital libraries.

On the other hand, very little works have been done in exploring and developing CLIA for indigenous African languages [84]. Due to the lack of adequate literature and comprehensive study on African languages, especially from the perspectives of CLIA, it is difficult to find detailed descriptions about the morphological processes and syntactic structures of resource-scarce African languages. Since most resource-scarce languages, including Afaan Oromo, came into the digital world lately from economically less developed regions, they do not have automated translation systems and information access technologies that can support them online. According to [21], the quality and quantity of existing linguistic studies on major indigenous African languages like Afaan Oromo ranges from fairly rich to extremely poor or nil. With its approximately 2000 different languages, Africa is a complex multicultural continent that presents a huge challenge for the development of CLIA systems. Taking these facts into account, some IR researchers and language technology developers, both from Africa and elsewhere, have come forward to share the common goal of building CLIA for resource-scarce African languages [85, 67]. As indicated by [71], a regional workshop on development and evaluation African Language Technology have been annually organized and conducted as a forum to bring together a wide range of researchers and practitioners working on this area since 2009.

One of the initial important case studies on the development of CLIR for indigenous African languages (i.e. English- Zulu CLIR) was reported by Cosijn et.al [66, 67]. It was an experimental study aimed at facilitating the accessibility of indigenous knowledge through the application of CLIR. A dictionary-based Zulu-English CLIR was designed and implemented for this purpose. The researchers have indicated that the disparate linguistic features of Zulu and English were among the major factors that had affected the effectiveness of the retrieval system. Two years later, another similar study was performed on Afrikaans-English CLIR by the same researchers [65]. A dictionary-based query translation technique was adopted to translate Afrikaans queries into English queries. Chief among the linguistic and translation resources that were employed for this CLIR include bilingual dictionary, Afrikaans morphological analyser and stop-word lists. After the source language query terms were translated into English queries through the bilingual dictionary, the translated search queries were matched against the English text corpus for retrieval of relevant documents. The performance levels of Afrikaans-English CLIR system was tested and assessed by using about 35 topics obtained from the ad hoc track of the CLEF-2001 campaign. The researchers had reported an average precision of 19.4% as the optimal performance of Afrikaans-English CLIR.

Over the past few years, various Amharic-English CLIR experiments had been conducted and reported by [86, 87, 88]. The first Amharic-English information retrieval experiment was conducted in 2004 with a focus on analysing and determining the impacts of removing stop-words from the Amharic query topics. Besides a bilingual dictionary that was used to translate Amharic topics to English search queries, a crude stemmer was designed and employed to normalize the word form variants of Amharic query terms. The evaluation results of the experiments showed that the removal of Amharic stop-words had slightly helped to improve the performances of Amharic-English CLIR. Another Amharic-English CLIR experiment was conducted in 2006 by incorporating Amharic morphological analyser and part of speech tagger in order to improve the performances of the cross-lingual retrieval system. Unmatched and out of vocabulary query terms were handled by using fuzzy matching techniques. The researchers had reported a mean average precision of 22.78% as one of the best performance for Amharic-English CLIR during their evaluation experiments. More recently, an experiment on English–Oromo MT was reported by [89], with a focus on translating of English documents into Afaan Oromo using statistical approach. A few of the major objectives of this study were: to explore applicability of Statistical Machine Translation (SMT) systems between Afaan Oromo and English as well as to assess how far the proposed MT system can perform using very limited parallel and monolingual corpora. According to the researchers, a translation accuracy 17.74% (in terms of BLEU Score), was achieved by the experimental English-Oromo MT system.

In summary, due to very limited studies conducted on indigenous African languages, especially from the perspectives of computational linguistics and IR, it is difficult to determine the current development state of CLIA for indigenous African languages. Generally, research on the development of computational resources and CLIA systems for African languages have very limited practical impacts. Even though increasing number of researchers and academic institutions have recognized the benefits of building CLIA systems for indigenous African languages, most of the current research initiatives are highly fragmented and poorly coordinated. As a result, despite some progresses that have been observed in recent few years, research projects on African language technologies have rarely succeeded in moving from a pilot phase into open-source software and commercial applications.

3 Afaan Oromo

Since the task of searching and retrieving information largely depends on natural language, understanding the linguistic properties of African languages is crucial for the development and effectiveness of CLIR. As discussed in section 2.2, the performance and effectiveness of CLIR is not only affected by the quantity and quality of existing language resources, but by the linguistic feature and properties of the languages being considered. In this chapter, the major morphological features and grammatical categories of Afaan Oromo are introduced and described. But, it is important to note that the goal of this chapter is not to cover and present detailed discussion about all linguistic phenomena in Afaan Oromo, which is beyond the scope of this study. Instead, it is intended to give an overview of Afaan Oromo grammar with a special emphasis on morphosyntactic structures that are relevant to the development of OMEN-CLIR.

As indicated in section 1.1, like many other resource-scarce African languages, Afaan Oromo remained poorly documented for quite a long period of time. In spite of being one the most widely spoken indigenous African languages, Afaan Oromo is still a language for which very limited linguistic resources and translation tools have been developed. The descriptions provided in this chapter are based on the reference materials available to the researcher, which are very limited and sometimes inconsistent with each other in their descriptions and discussions of linguistic phenomena in Afaan Oromo. More detailed study and linguistic description of Afaan Oromo are necessary to understand and analyse the morphological processes and grammatical features of the language. We believe that many linguistically interesting features and computationally important findings will emerge from more detailed and comprehensive studies on Afaan Oromo in the near future.

3.1 An Overview of Afaan Oromo

Afaan Oromo, which is sometimes also referred to as Oromo, is one of the major indigenous African languages. It is considered as one of the five largest languages on the African continent. Afaan Oromo is widely spoken in most parts of Ethiopia by more than 34% of the population [90]. Accordingly, while over 100 indigenous languages are spoken in Ethiopia, Afaan Oromo is the most widely spoken language in the country. In addition to the Oromo people, for whom

Afaan Oromo is a mother tongue, many speakers of other Ethiopian languages also speak Afaan Oromo as their second language. Afaan Oromo is also spoken by a large number of minorities in neighbouring countries, including Kenya and Somalia [91, 21, 63]. Moreover, it is spoken by a significant number of Oromo diaspora in different countries around the world, including the USA, Canada and Germany.

As discussed in section 2.1, among the four major groups of African languages, Afaan Oromo belongs to the Cushitic family of languages, which also includes Somali, Afar and Sidama among other languages, under the Afro-Asiatic language phylum. It is the most spoken Cushitic language in the world. Confirming this, [63] stated that Afaan Oromo is the most prominent constituent of the Cushitic languages. Like many other African countries, the cultural and linguistic diversity of Ethiopia is staggering, with its more than 80 different ethnic groups and over 100 indigenous languages. Ethiopia is not only one of the most populous countries in Africa, but also one of the most culturally and linguistically diverse nations in the continent. Currently, Ethiopia is divided into about nine regional states, including Oromia state. Although each region has its own official language, Amharic is the official language of Ethiopia. Amharic is also used for intercultural communication in most parts of the country. Genetically, unlike Afaan Oromo, Amharic belongs to the Semitic family of languages. Amharic is considered as the second most spoken Semitic language in the world (after Arabic) and the second largest language in Ethiopia (after Afaan Oromo). In Ethiopia, English is the most widely spoken foreign language. English is used as a medium of instruction in most secondary schools and higher education institutions in the country.

With regard to orthography, a modified Latin alphabet (known as *Qubee*) was formally adopted and become a standard script of Afaan Oromo in 1991. Hence, Oromo has had a standard writing system of its own for just about two decades now. Although there are several Afaan Oromo documents including books, manuscripts and religious scriptures that had been written in different scripts over the last many centuries, Afaan Oromo had largely been a vernacular for several generations and centuries, with little in the way of literary expression. Most of the documents published in Afaan Oromo during the last two centuries used a mixture of Ge'ez (Ethiopic) script and various transliteration schemes that were crafted by the individual authors. Over the recent couple of decades, *Qubee* has replaced the earlier transliteration schemes (or scripts) and helped to standardize the spelling of Afaan Oromo words. However, there are still some spelling differences, which may partly reflect dialectical differences in pronunciations of certain Afaan Oromo words, but does not affect or change the meaning of the text. Like other

Eastern Cushitic languages, Afaan Oromo has set of five short and five long vowels, indicated in the orthography by doubling vowel letters. The difference in vowel length is contrastive in Oromo, for example, *"rafuu"* ("to sleep") vs. *"raafuu"* ("cabbage"). Gemination is also significant in Oromo. That is, consonant length can distinguish words from one another, for example, *badaa* ("bad") vs. *baddaa* ("highland"). In general, the adoption and use of *Qubee* as a standard writing system have significantly increased the quantity and quality of Afaan Oromo publications in recent years.

In the beginning of the 1990s, Afaan Oromo was formally adopted as an official language of Oromia state and becomes a medium of instruction in primary schools throughout the region. Currently, Afaan Oromo is a medium of primary education and the language of government and public administration in Oromia state, which is the largest Regional State in Ethiopia in terms of both geographic area and population size. Nowadays, Afaan Oromo is the language of court and mass media (including print, broadcast and electronic media) as well as the language of business and general day-to-day interactions in Oromia state. It is also taught as a subject in many high schools and higher education institutions in Oromia. The adoption of Afaan Oromo as an official language of Oromia state and a medium of instruction in primary schools have not only facilitated the study and acquisition of the language, but considerably enhanced its standardization and development.

However, like many other indigenous African languages, Afaan Oromo is not yet in a position to take full advantage of the rapidly advancing language technologies. While computational linguistic resources and IR systems have been extensively investigated and developed for well-resourced European and Asian languages, very little work has been done on developing linguistic resources and CLIA for resource-scarce Cushitic languages like Afaan Oromo and Sidama. From the computational linguistics point of view, the majority of indigenous African languages, including Afaan Oromo, are considered as severely under-resourced languages, with very limited online content and lexical databases available to them, let alone annotated parallel corpora, Treebank, and automated translation systems. As discussed in section 1.3.1, resource-scarce languages are characterized by having a low language density, with very limited linguistic resources such as machine-readable dictionaries, parallel text corpora and language processing tools. According to [21], the quality and quantity of existing studies on Afaan Oromo ranges from fair to extremely poor. In terms of the availability of computational linguistic resources and online translation tools required for the development of CLIA and MT systems, Afaan Oromo is one of the most resource-scarce African languages. Due to the lack of linguistic resources and information access technologies, the majority of African languages, including Afaan Oromo, are either poorly represented or completely missing from the WWW.

Like many other indigenous African languages, Afaan Oromo, which belongs to Afro-Asiatic family of languages, does not share common linguistic roots with English, which belongs to Indo-European family of languages, (see section 2.1 for more detailed descriptions). Oromo and English have quite different morphological and syntactic structures. Although, many technical words, scientific terms and names of foreign origin, including places and people names, are often derived or borrowed from English, Afaan Oromo and English do not share many vocabularies or cognates in the way that, say, English and French do. Unlike French and English, which have many common Latin roots, English and Afaan Oromo have none. Consequently, their morphological and syntactic structures are almost completely different. Generally, there are many morphological and grammatical differences between Afaan Oromo and English, which make the task of building CLIR for this language pair even much more difficult and challenging.

In summary, some of the major morphological properties and syntactic structures that distinguish Afaan Oromo from English include:

- Whereas Afaan Oromo has a flexible or variable word order (i.e., SOV), English has a fixed word order (i.e., SVO);
- While Afaan Oromo is considered as a highly synthetic and agglutinative language, English is a moderately analytic language;
- While Afaan Oromo belongs to the Cushitic branch of the Afro-Asiatic group of languages, English belongs to the West Germanic branch of the Indo-European group of languages
- While prepositions and auxiliaries are widely used in English to express syntactic functions and grammatical relations, most of the syntactic functions and grammatical relations in Afaan Oromo are encoded and conveyed through inflectional affixes;
- While prepositions are very common and widely used in English, postpositions are more common in Afaan Oromo;
- While Afaan Oromo adjective agree with its head noun, this is not the case in English.

3.2 Afaan Oromo Morphology

Understanding a language involves identifying and analysing the linguistic properties of the language at different levels including phonology, morphology, syntax and semantics. It requires a wide range of knowledge including pronunciation and sound system (phonology), word formation and structure (morphology), grammar rules and sentence structure (syntax). The aim of this section is to introduce and describe Afaan Oromo morphology. Morphology pertains to the internal structure of a word, called morpheme, which is considered as the smallest meaningful grammatical unit of language. A morpheme can be **free**, i.e., it can stand alone as a word, without another morpheme like the Afaan Oromo noun "*mana*" (“house"), or adjective "*guddaa*" ("*big*"). A morpheme can also be **bound**, i.e. it can only be used in combination with other independent morphemes. For example, Afaan Oromo plural markers of such as*, “-oota”, “-wwan”*, and “*-oollee”* can only be used in combination with free morphemes like "*mana*", i.e. "*mana + oota*→ *manoota*" (meaning “houses”).

Morphologists often focus on two major aspects of morphological processes: how words are constructed or formed (derivational morphology) and how words interact with syntax (inflectional morphology). While derivational morphology is concerned with the basic principles that govern word formation, inflectional morphology focuses on identifying and analysing how words are modified or inflected in order to express syntactic functions and grammatical categories. In this section we focus on Afaan Oromo inflectional morphology. As indicated in section 3.1, it is difficult to find a systematic and comprehensive study on morphological processes of Afaan Oromo, let alone a recent book on its computational morphology. Most of the morphological descriptions presented in this study are obtained from various published and unpublished reference materials such as [63, 22, 21, 91, 62, 92]. Morphology is chosen as a focus in this study because of the synthetic and agglutinative properties of Afaan Oromo, which pose a huge challenge to the development of CLIR.

Unlike English, which has a relatively simple morphology, Afaan Oromo has very rich and agglutinative morphological structures. Inflected words are formed by affixing various bound morphemes such a case, number, gender and tense markers to the stem or the root word. In agglutinative languages like Afaan Oromo, multiple bound morphemes (affixes) can be attached to a stem like “beads on a string.” As indicated by [91, 62, 21], most of the syntactic functions and grammatical relations in Afaan Oromo are encoded and conveyed through inflectional affixes. Unlike English, which often uses word order and prepositions to express grammatical categories, most of the syntactic functions and grammatical relations in Afaan Oromo are encoded and conveyed through inflectional affixes and attached postpositions. Grammatical categories such as number, definiteness, gender and case are marked and indicated through inflectional suffixes. While Afaan Oromo nouns declines for number, gender, definiteness and case, Afaan Oromo verbs are also conjugated to show grammatical relations such as gender, number, tense, aspect and mood. As a result, the surface form of words may be composed of prefixes, a root/stem, and multiple suffixes. Since the use of prefixes is less common in Afaan, we have focused on identifying, segmenting, removing and normalizing inflectional suffixes that are frequently used in Afaan Oromo.

Although other combinations of Subject-Object-Verb are possible, the predominant word order is SOV in Afaan Oromo. Because Afaan Oromo is an inflectional language, (i.e., the surface form of a noun or pronoun declines or changes depending on their role in the sentence), its word order is quite flexible. But Oromo verbs tend to come after their subjects and objects in most declarative sentences and main clauses. Hence, it is possible to consider Afaan Oromo as a verb-final language. When Afaan Oromo words are formed by adding bound morphemes (or affixes) to the stem or root word, various morphophonemic changes may take place. The order in which affixes are attached to the stem or base form is determined by the morphotactics of the language that is briefly described section 3.3. The morphological properties of Afaan Oromo nouns and verbs are introduced and described in the subsequent subsections.

3.2.1 Oromo Nominal Morphology

Generally, Oromo nouns are words used to name a person, animal, place, thing, or abstract ideas. As noted by [21], Afaan Oromo nouns are marked to indicate various syntactic functions and grammatical categories, including number, gender, definiteness and case. Inflectional suffixes are predominantly used for marking most of these grammatical categories. The majority of Oromo nouns end with a vowel. The accusative form, which is not marked for grammatical case is used as the root-noun or base-noun. For example, the base-form or root-word of the inflected noun "*manoota*" (i.e., houses) *is "mana"*. Inflectional affixes are attached to this base-form to express syntactic functions, (e.g., "*mana* + *icha*" → "*manicha*" (which means "the house").

3.2.1.1 Number marking

Number marking in Afaan Oromo consists of a binary distinction between singular and plural for countable nouns. Like in English, the singular noun is not marked in Afaan Oromo. This unmarked (singular) noun is often used as citation form (or the accusative case). The plural form is indicated by various inflectional affixes. Generally, the plural form is used to specify that the speaker or writer is concerned with more than one objects or entities. Oromo plural nouns are marked with different suffixes. While there are several suffixes that are used to indicate a plural noun in Afaan Oromo, the most commonly used suffixes are "*-(o)ota*" and "*-wwan*". In order to indicate the plural form, the final vowel of noun stem is dropped before the appropriate number marker is attached to the stem. Some of the common Afaan Oromo plural marker include: "*-oota*", "*-ooli*", "*-wwan*", "*-lee*", "*-ooti*", "*-an*", "*-een*" and "*-oo*" and "*-yyii*". It is possible to mark a noun with more than one or alternative plural suffixes in Afaan Oromo. For example, the plural form of noun "*mana*" (i.e., "house") can be indicated with the following different plural suffixes:

- *manoota* → *(mana + oota),*
- *manneen* → *(mana + een),*
- *manawwan* → *(mana + wwan).*

Unfortunately, the correct plural suffix cannot be predicted from the structure or properties of the noun stem. In certain contexts, an alternative suffix or multiple suffixes could be used to mark a plural noun. In other words, a noun could be marked by more than one plural suffixes, concatenating one after the other, as in:

- *manneenota* → *(mana + een + ota)* or,
- *manneenotawwan* → *(mana + een + ota + wwan).*

When a plural noun is modified by an adjective, the adjective is marked by the appropriate plural suffix. The use of plural suffixes tends to be more common in written Afaan Oromo than in spoken conversations. Marking plural noun may not be required in Afaan Oromo if the number of the noun can be distinguished from the context or by other means of grammatical functions.

3.2.1.2 Gender of nouns

Nouns in Afaan Oromo show two dimensions of gender: masculine and feminine. In other words, Afaan Oromo nouns are treated as either male or female. Nouns derived from verbs are specifically marked to show the gender of the noun. In such cases, where the masculine noun is marked by "*-aa*", the feminine noun is marked by "*-tuu*". Table 3.1 shows examples of verbal nouns that are marked for gender. For other nouns, the gender is often expressed through a demonstrative pronoun, a definite article, a gender-specific adjective, or the verb form (if the noun is in nominative case).

Verb (Stem)	**Gender**		**Glosses**
	Masculine	**Feminine**	
beekuu ("to know")	*beek**aa***	*beek**tuu***	intelligent
barreesuu ("to write")	*barrees**aa***	*barrees**tuu***	writer
fiiguu ("to run")	*fiig**aa***	*fiig**duu***	runner
barsiisuu ("to teach"	*barsiis**aa***	*barsiis**tuu***	teacher

Table 3.1 Example of gender marking in Oromo verbal nouns

3.2.1.3 Definiteness

Unlike English, Afaan Oromo does not indicate indefiniteness, (corresponding to English "*a*"). While "*the*" is used to indicate definiteness for both masculine and feminine nouns in English, Afaan Oromo has separate inflectional suffixes to indicate definiteness for each gender. When "*-(t)icha*" used to mark masculine nouns, "*-(t)ittii*" is often used to mark feminine nouns. If the noun stem ends with a vowel, the vowel is dropped before the definiteness suffix is attached to it. For example: "*harree*" (which means "donkey") is inflected or declined as follows to indicate definiteness:

- *"harr**icha**"* → *("harr + icha")* "the donkey", (Masc.),
- *"harr**ittii**"* → *("harr + ittii")* "the donkey", (Fem.).

In comparison with English, marking definiteness appears to be less common in Afaan Oromo. For instance, unlike in English, Oromo noun can be either definite or plural, since it is grammatically wrong to use plural marker along with definiteness marker for the same noun. Moreover, definite nouns are rarely modified by demonstrative pronouns or possessive pronouns. Marking definiteness is not required by the syntax if its role can be indicated or specified by other grammatical functions in Afaan Oromo.

3.2.1.4 Case marking

A grammatical case marker pertains to a set of inflectional forms added to nouns and pronouns to indicate their role (function) in a sentence. The case of a noun or pronoun shows its relationship with the other words in the sentence. A nouns or pronoun can be marked with different case affixes depending on its function in a given sentence, such as subject (nominative), object (accusative) indirect object (dative), etc. In English, word order and prepositions (or prepositional particles) are often used to express the grammatical cases such as such accusative, dative, and ablative. Afaan Oromo nouns and pronouns are inflected to indicate various grammatical cases. In other words, the surface form of a noun or pronoun declines depending on its function (case) in a sentence. A noun may indicate its grammatical case by a specific case suffix or by lengthening a final vowel (if it ends in a short vowel). Commonly used Afaan Oromo cases include: *nominative*, *accusative*, *genitive*, *dative*, *instrumental*, *locative*, and *ablative*. Normally the accusative case, which is not marked, is used as a citation form (or base form). While a few examples of Afaan Oromo noun cases are illustrated in Table 3.2, cases of pronouns in Afaan Oromo are given in Table 3.3. Note that the attached case suffixes are shown in boldface in both tables. Many Oromo nouns ending in a short vowel and indicate different cases by lengthening that final vowel. This common case marking situation is not illustrated in Table 3.2. If a given noun has a plural or definite suffix, the case marker (or suffix) is attached after the plural or definite suffix. As indicated in Table 3.2 and in Table 3.3, for most of the grammatical cases, there are alternative suffixes that can be used in different contexts. Though a noun that ends with short vowel, e.g. "*mana*" (which means "house") and a noun that ends with long vowel, e.g. "*saree*" (which means "dog") are used to illustrate most of the case suffixes in Table 3.2, a few other nouns are also used in the context where these two nouns cannot be used properly.

Case	Case Marker	Conditions or Contexts of Use (Noun's Endings)	Examples	Glosses
Accusative		A base or citation form	*mana, saree*	house, dog
Nominative	*-ni*	Noun ends in a short vowel and has a single penultimate consonant.	*man**ni***	house (Nom.)
	-i	Noun ends in a short vowel and with two penultimate consonants.	*fard**i***	horse (Nom.)
	-n	Noun e ends in a long vowel.	*saree**n***	dog (Nom.)
Dative	*-f*	Noun ends in a long vowel or with a lengthened short vowel.	*saree**f*** (*manaa**f***)	for dog (Dat.) (for house)
	-iif	Noun ends in a consonant	*bishaan**iif***	for water (Dat.)
	-dhaa(f)	Noun ends in a long vowel or in a lengthened short vowel.	*saree**dhaa(f)***	for dog (Dat.)
	-tti	Noun ends in a vowel.	*saree**tti*** (*mana**tti***)	to dog (Dat.) (to house)
Instrumental	*-n*	Noun ends in a long vowel or in a lengthened short vowel.	*saree**n*** (*manaa**n***)	with dog (with house) (Inst.)
	-iin	Noun ends in a consonant	*afaan**iin***	with mouth
	-dhaan	Noun ends in a long vowel or in a lengthened short vowel.	*saree**dhaan***	with dog
Ablative	*-dhaa*	Noun ends in a long vowel or in a lengthened short vowel.	*Finfinnee**dhaa***	From Finfinnee
	-ii	Noun ends in a consonant	*Hyderaabaad**ii***	From Hyderabad
	-rraa	Noun ends in a vowel.	*mana**rraa***	from house

Table 3.2 Example of common case markers in Afaan Oromo

Base Form	Glosses	Nom.	Dative	Instr.	Ablative	Locative
ana	I	*a**ni*** (*an*)	*na**a***, (*naa**f***)	*naa**n***	*na**rraa***	*na**tti***
nu	we	*nu**ti***	*nuu**f***, (*nu**tti***)	*nuu**n***	*nu**rraa***	*nu**tti***
si	you (sg.)	*ati*	*sii*, (*si**tti***)	*sii**n***	*si**rraa***	*si**tti***
isin	you (pl.)	*isin*	*isin**ii***, (*isin**iif***)	*isinii**n***	*isini**rraa***	*isini**tti***
isa	he	*in**ni***	*isa**a***, (*isaa**f***)	*isaan**iin***	*isa**rraa***	*isa**tti***
ishii	she	*isii**n***	*ishii**f***, (*ishii**tti***)	*ishii**n***	*ishii**rraa***	*ishii**tti***
isaan	they	*isaan*	*isaan**ii*** (*isaan**iif***)	*isaan**iin***	*isaan**irraa***	*isaan**itti***

Table 3.3 Afaan Oromo pronouns case

As indicated in Table 3.2 and Table 3.3, common Afaan Oromo case suffixes include “*–n*” and “-*ni*” for nominative, “*-dhaa(f)*” and “*-f*” for dative, “*-dhaan*”, “*-n*” and “*iin*” for instrumental. Declension of nouns in Afaan Oromo is not limited to common nouns. This is because proper nouns such as place names (or geographic names) and personal names are also inflected just like other nouns. For example, a personal name “*Bulchaa*” may take various case markers such as:

- *Bulchaa + n* → “*Bulchaa**n***” (nominative);
- *Bulchaa + f* → “*Bulchaa**f***” (dative or beneficiary, i.e., for *Bulchaa*);
- *Bulchaa + dhaa* → “*Bulchaa**dhaa***” (dative or beneficiary, i.e., for *Bulchaa*);
- *Bulchaa + tti* → “*Bulchaa**tti***” (dative or beneficiary, i.e., to *Bulchaa*);
- *Bulchaa + dhaan* → “*Bulchaa**dhaan***” (instrumental, i.e., by *Bulchaa*).

As it can be seen from the above examples, it is important to remember that most prepositions cannot be literally translated from English into Afaan Oromo, since their functions are expressed through inflectional affixes in Afaan Oromo.

3.2.2 Oromo Verbal Morphology

Like many other Cushitic languages, Oromo makes a two-way distinction in its verb system between the two major aspects: perfect (or past tense) and imperfect (or present tense). A typical Afaan Oromo verb consists of a combination of two or more dependent morphemes. An Oromo verb consist minimally of a root-morpheme (base-verb), representing the lexical meaning of the verb, and one or more inflectional affixes, indicating syntactic functions such as tense, aspect and subject agreement. The simplest verb form (or infinitive) is composed of a mandatory dependent root-morpheme (verb-base) and a long final vowel, i.e. “*-uu*”. Oromo verbs are normally listed in dictionaries in infinitive form, which is not marked for a tense or person. For instance, the verb “*baruu”* (which means “to learn or to know” and the verb “*deemuu”* (which means “to go”) are appropriate verb-lexemes in most Oromo dictionaries. Generally, since all Oromo infinitives end in “*-uu*”, it is easy to identify the verb-base of any verb, which is the infinitive form without the final long vowel “*-uu*”. For example, the root-verb of the Oromo verb “*baruu” is “bar-”,* the root-verb of the Oromo verb “*deemuu” is “deem-”*. But, it is important to note that, without certain appropriate linguistic contexts, some Oromo nouns, adjectives and inflected words that end with “*-uu*” can make the infinitive form very ambiguous.

As indicated in the preceding sections, since Afaan Oromo is a highly synthetic and an agglutinative language, most of words decline according their syntactic functions in a sentence. An Oromo verb is conjugated by adding one or more affixes to a verb-base or stem (if it is already in an inflected or derived form). Most Oromo verbs are treated as “regular”, that is, they attach the regular person- and number-based markers to their verb-root with very minor orthographic or phonological modification of the base form. In particular, verbs whose root does not end with a: “*double consonant*”, “*ch*”, “*vowel*”, “*y*”, or “*w*” are considered as regular verbs in Afaan Oromo. As an example, the present tense and past tense conjugations for a regular verb “*baruu*” (which means “to learn”) are shown in Table 3.4. Note that the suffixes are indicated in boldface.

Pronoun (Nom.)	**Present Tense**		**Past Tense**	
	Oromo	**Gloss**	**Oromo**	**Gloss**
ani	_bar**a**_	(I learn)	_bar**e**_	(I learned)
nuti	_bar**ra**_	(we learn)	_Bar**re**_	(we learned)
ati	_bar**ta**_	(you learn)	_bar**te**_	(you learned)
isin	_bar**tu**_	(you learn)	_bar**tan**_	(you learned)
inni	_bar**a**_	(he learns)	_bar**e**_	(he learned)
isiin	_bar**ti**_	(she learns)	_bar**te**_	(she learned)
isaan	_bar**u**_	(they learn)	_bar**an**_	(they learned)

Table 3.4 Example of Conjugation for a regular Oromo verb

	Examples					
Oromo Sandhi Rules	**Verb Root**		**Suffix**		**Inflected Form**	**Gloss**
b- + -t → bd	qab	+	ta	→	qabda	('you have')
[t- | d- | x-] + -n → nn	bit	+	na	→	binna	('we buy')
r- + -n → rr	jir	+	na	→	jirra	('we are')
l- + -n → ll	ilaal	+	na	→	ilaalla	('we see')
t- + -n → nn	nyaat	+	na	→	nyaanna	('we eat')
x- + -t → xx	fix	+	ti	→	fixxi	('she finishes')
s- + -t → ft	baas	+	tu	→	baaftu	('you remove)

Table 3.5 Example of Afaan Oromo sandhi rules

According to [21, 22], the morphological structure of Afaan Oromo verbs is much more complicated than that of nouns. There are many inflection affixes that are used to mark tense and aspect as well as person, gender and number agreements with the subject. As indicated in Table 3.4, even a simple and regular verb like "*baruu*" is conjugated to show several grammatical properties. As an additional example, a few of the inflected forms for an Afaan Oromo regular verb *“himuu”*, (i.e., to tell) include: *“hima”, “himi”, “himan”, “hime”, “himte”, himne”, “himeera”, “himneera”, “himteetta”, “himteetti”, “himtanittu”,* and *“himaniru”*.

Afaan Oromo verbs often mark gender, number and person agreement, when the subject of the sentence is third person singular, and number and person agreement for the rest. Oromo verbs are also marked for tense and aspect of the action as well as the mood of the speaker. Moreover, Afaan Oromo verbs conjugate for causation and passive voice. Broadly, while a tense can be classified into past tense and present tense; aspect can be classified into perfective, imperfective and progressive in Afaan Oromo. Mood can also be classified into indicative, imperative, interrogative and conditional. As indicated Table 3.5, a set of sandhi (morphophonemic) rules may need to be applied during the process of attaching suffixes to a root-word or other inflected form a word.

In summary, some of the morphological processes and grammatical features of Afaan Oromo verbs may include:

- tense – present and past;
- gender - masculine and feminine;
- number - singular and plural;
- person - first, second and third;
- polarity - positive and negative;
- aspect - perfect, habitual and progressive;
- mood - jussive and imperative.

3.2.3 Other Morphological phenomena in Afaan Oromo

As described in the preceding sections, owing to its very complex and agglutinative inflectional morphology, Afaan Oromo poses a huge challenge to the development of natural language processing applications and information access systems such as IR and CLIR. Unlike in English, in Afaan Oromo, adjectives decline and agree with the noun-heads they modify. For example, the plural form of adjective *“cimaa”* (i.e. “strong”) can formed as: “*cim**oota**”* → *(cimaa + oota)*, or *cim**oo*** → *(cimaa + oo)*. Unlike in English, Oromo adjectives follow the noun they modify. For instance, in the sentence “*Isaan bart**oota** cim**oo** dhaa*”, (which means “They are strong students”), the adjective “*cim**oo***” agrees in number with the plural head-noun “*bar**toota***” (i.e., “students”). As indicated in Table 3.5, a set of sandhi rules or morphophonemic rules, (which may involve deletion and insertion of vowels as well as modifying or germination of consonants) are also often applied during the process of affixation [62, 91].

In addition to extensively productive nominal, verbal, and adjectival inflectional morphology, Oromo has many other morphological processes and properties. Although postpositions, pronouns, conjunctions, auxiliaries and other related particles in Afaan Oromo are grammatically free morphemes and could be used as independent word, usually, they are treated as clitics or enclitics. A clitic is a morpheme that has the syntactic characteristics of a free morpheme but is phonologically bound to another word (or host). They are often attached to the stem of nouns, adjectives and verbs. In Afaan Oromo, different grammatical cases are also marked by postpositions, which may attached to the nouns or the adjective they are associated with. As shown in Table 3.6, a few examples of Afaan Oromo postpositions, pronouns, auxiliaries and particles that may attached (or affixed) to the content words and treated as enclitics include: *irra, irraa, dhaa, fi, jira, jiru, kaa, keessa, itti, moo, woo,* and *yoo*. Possessive pronouns such as *koo, kee, isaa,* and *ishi* are also sometimes attached to a noun they modify [62].

Oromo Postpositions	Gloss	**Oromo Prepositions**	Gloss
irra	on	*gara*	towards
bira	beside, around	*wal*	with
irraa	from	*haga (hanga)*	till, until
itti	to, towards	*hamma*	up to, as much as
koo	my	*akka*	like, as
kee	your	*waayee*	about, in regard to
isaa,	he	*waan*	about

Table 3.6 Example of Afaan Oromo postpositions and prepositions

Besides inflectional morphology discussed in the preceding two sections, Afaan Oromo has also extensive derivational morphology. Afaan Oromo nouns have various derivational suffixes such as *–aa, achuu, -eenyaa, -ee, -ii, -ina, -iisa, -iinsaa, -maata, -noo, -ooma, -tuu, -umsa* and *–ummaa*. Verbal derivations in Afaan Oromo may take suffixes such as *-s, is, -sis, -siis* for causative; *-am* for passive and *-adhf* for reflexive. Other major types derivational suffixes of

Afaan Oromo include: *-achu, -amuu, -isu, -eessu, -uu*. These derivational suffixes are often used for formation of new words or lexemes in the language [21, 91, 62]. Earlier studies on effects of derivational morphology on the performance of IR and CLIR have shown that removing derivational affixes may not significantly improve the effectiveness of retrieval systems [93, 81, 58, 19, 24, 94]. Since addition (or attachment) of derivational affixes of often changes the categories and semantics of the original words, it is not a good idea to remove derivational prefixes and derivational suffixes from words for the purpose of IR. Taking this into account, we have mainly focused on identifying and normalizing inflectional suffixes in this study.

In general, due to its synthetic and agglutinative morphology, Afaan Oromo poses a huge practical challenge for natural language processing applications like IR and CLIR. Particularly, in the context of a dictionary-based CLIR, keywords in a source language query may fail to match with the lexical entries in a bilingual dictionary because of the word form variations. The abundance of inflectional affixes increases the likelihood of mismatches between the source language query terms and the lexical entries provided in a bilingual dictionary. Owing to word form variation, a number of source language search terms may not be found in a bilingual lexicon. This results in a partial or incomplete translation of a search query. The presence of many inflected words in a query can, thus, considerably reduce the performance of a CLIR. Hence, the need for identifying and normalizing word form variations is critical for morphologically very rich and complex languages like Afaan Oromo.

3.3 Towards Finite-State Morphology of Afaan Oromo

3.3.1 Overview of Computational Morphology and FST

An important part of developing natural language processing (NLP) applications such as CLIR and MT systems is the designing of computational models that capture the morphological processes and grammatical properties of the languages being considered. The complexity of such modelling task is in part dependent on the morphological properties of the intended language, that is, how words of the language are inflected and derived. Since languages often differ in what grammatical information they encode in words and how they encode the information, the linguistic phenomena and peculiarities of each language need to be carefully examined and studied in order to come up with an appropriate computational model. In morphologically simple languages such as English and Afrikaans, it is often sufficient to use words (i.e. surface forms) as basic units of text analysis and indexing. However, this is not the case in morphologically very rich languages, which are characterized by high degree of inflection, agglutination and compounding that may produce a very large number of word forms from a single root-word or lexeme. Since Afaan Oromo is a morphologically very rich language, it has a huge number of distinct word forms which makes using words as basic units of indexing and searching very difficult and inefficient. Hence, there is a compelling need for designing a method that can identify and normalize the morphological variants of lexical items. To this end, we propose and present a general computational model for Afaan Oromo inflectional morphology. Since the finite state morphology is one of the most successful methods applied in a wide variety of languages over the last few decades, we have explored and adopted a Finite State Transducer (FST-based) approach to model and describe Afaan Oromo inflectional morphology.

According to [34], two of the major challenges that are frequently encountered in a computational morphology are:

1. The morphemes that make up the surface forms of words do not combine at random; their combinations and orders are often language specific and selective. A morphological processing tool needs to know which sequence of combining morphemes are valid or invalid.
2. Morphemes may be realized in different forms and ways, depending on their phonological and morphological contexts. Morphological processing tools and analysers need to recognize the morphophonological changes between lexical and surface forms (i.e. alternation rules).

More specifically, the two fundamental challenges in the morphological analysis are morphotactics and orthography. While the morphotactics deals with how the morphemes combine together to form a word and express syntactic information, the orthography is concerned with modelling how morphemes modify or change their spellings when they combine together. In order to analyse and process different components of natural language, computational linguistics have developed different formal approaches and models. For instance, a computational models based on a finite state automat (FSA) have been commonly used for phonological and morphological processing. In particular, computational models based on Finite-State Machine (FSM) and its extended version called FST have been widely used to model the relationship between upper language (lexical level) and lower language (surface level). Basically, an FST is a model of linguistic behaviour that consists of [34]:

- a finite number of states;
- transitions from each state to another state and;
- actions to be taken at each transition.

Accordingly, an FST is just another type FSA where arcs are labelled by pairs of symbols, a lexical (underlying) level symbol and a surface level symbol. For example, an arc or a transition arrow labelled '*a:b*' means '*a*' stands for lexical level representation while '*b*' denotes the surface level representation. A transition from current state to another state through an arc "*a*" can be taken if the current symbol on the top tape matches the lexical symbol and the current symbol on the bottom tape matches the surface form representation. One may start with the surface level representation tape containing an inflected word (as it appear in a text), and then tries to identify or determine the lexical forms that are licensed or specified by the FST.

As discussed in the foregoing section, inflected words are formed by affixing or attaching bound morphemes such as case, number, gender, tense and aspect markers to their stem or base form. In morphologically rich and agglutinative languages like Afaan Oromo, inflectional morphology is extensively productive, meaning they can produce extremely large number of *word form variants* from a single word-root or lexical item. Indeed, it is very difficult to list all inflected Oromo words in a general purpose dictionary. Consequently, inflected words are excluded from most of Oromo lexicons, including bilingual dictionaries and thesauri. Apart from the sheer amount of time and resources needed to include all inflected words, incorporating all possible word forms in a general purpose dictionary would make the database highly susceptible to error and extremely difficult to maintain. Hence, it is crucial to devise an efficient mechanism that can identify a lexical-root (base-word) from which word form variants (or surface forms) are derived. In order to obtain the stem, morphological affixes must be identified and analysed by appropriate morphological processing tools such as stemmer, lemmatizer and morphological analyser.

To this end, an FST-based approach provides an efficient method of synthesizing and parsing words. FST has been successfully employed to model both morphological generator and morphological analyser for many languages. In particular, the *two-level morphology* model has been proved successful for modelling and formalising the morphological structures of different languages including English, German, French Arabic and Swahili (a major member of Bantu languages in Africa). A morphological analyser processes a word by segmenting it into its component morphemes and assigning them with appropriate morphological information, including the category of the word class along with grammatical information. A morphological generator is the opposite or reverse of a morphological analyser. It takes a lexical-root or lemma as input and generates the appropriate word form variants. A morphological generator is also responsible for combining or synthesizing different morpheme elements into a word (surface form).

As described in more detail in the next two sections, the task of analysing a word requires a well-established morphotactics, which guides and governs the combination of the stem and inflectional affixes. However, when a particular phone appears in specific positions within a words or when combinations of phones come together, the surface form may be different from what the morphotactics would dictate. As shown in Table 3.6, in addition to morphotactics, the morphophonemic changes (sandhi rules) occurring when a root word or stem concatenates with affixes should be applied. Sandhi rules are phonological and morphological alternations that are triggered at junctures and at junctions of base-words and morphemes. In this study, we have tried to model the morphotactics that govern the order of Afaan Oromo inflectional suffixes. The surface forms and morphological structures of Afaan Oromo nouns have been carefully examined in order to come up with their morphotactics and morphophonemic rules. A shown Table 3.6, in Afaan Oromo, morphophonemic changes are often occurring in the end phonemes of the stem and the initial phoneme of inflectional suffixes.

	Example of Sandhi rules				
Oromo morphophonemics Rules	**Root_word**		**Suffix**		**Inflected Form**
b- + -t → bd	*dhab-*	+	*-ta*	→	*dhabda*
[t- | d- | x-] + -n → nn	*but-*	+	*-na*	→	*bunna*
r- + -n → rr	*bir-*	+	*-na*	→	*jirra*
l- + -n → ll	*bul-*	+	*-na*	→	*bulla*
t- + -n → nn	*galt-*	+	*-na*	→	*galinna*
x- + -t → xx	*fix-*	+	*-ti*	→	*fixxi*
s- + -t → ft	*buus-*	+	*-tu*	→	*buuftu*

Table 3.7 Example of Afaan Oromo morphophonemics rules

3.3.2 An FST Model for Oromo Nominal Inflection

As indicated in section 3.2, Oromo nouns and verbs form the two main lexical categories in the language. Thus, they often take different types of inflectional and derivational affixes to express syntactic functions. Generally, word formation and inflectional processes are not arbitrary since affixes are attached to a root-word or stem in a certain order or sequence. But, the task of generalizing and representing inflectional morphology is very challenging. For instance, formalizing and generalizing a nominal morphology involve the following major steps:

- Noun morphotactics description;
- Morphophonemics rules description;
- FST (or transducer) network formation.

As indicated in the foregoing section, morphotactics is concerned with the principles that govern the combination of stems, affixes and other morphological properties to form a grammatically acceptable word in a given language. The first major step in defining the morphological structure and morphotactics of a language is, therefore, categorization of different morphemes into certain paradigms, groups and sequences, primarily based on the affixes that a particular morpheme can follow and the affixes that can follow it. Therefore, the term *morphotactics* is used here to refer to a set of rules that are concerned with the ordering and realization of inflectional affixes. It governs which (types of) morphemes can or cannot follow other morphemes within the structure a given paradigm. The following four figures (i.e., Figure 3.2, Figure 3.3, Figure 3.4 and Figure 3.5), illustrate the morphotactics of Afaan Oromo regular nouns and verbs. A circle and double square are used to represent final states (i.e., legal word). An arrow indicates transition from one state to the next upon receiving the input symbol indicated near the arrow. A transition can be made if the item (string) currently being parsed satisfies the symbols specified near the arrow (or edge of the arc). As shown in Figure 3.1, Figure 3.2 and Figure 3.3, Afaan Oromo nouns consists of a ***stem*** (*root-noun)* that may be followed by one or more optional suffixes. More specifically, an Oromo *noun-stem* or *root* (labelled as "*N_STEM*" in Figure 3.2) can occur in isolation or may take inflectional suffixes in the following order (if any):

1. A plural or definiteness markers;
2. Grammatical case markers and,
3. Postpositions or enclitics.

The morphotactics of Oromo noun inflectional is summarized and represented in Figure 3.1 as follows:

***N_STEM* + [*plural* | *definiteness*] + [*case*] + [*enclitics*]**

Figure 3.1 Morphotactics of nominal inflection in Afaan Oromo

Generally:

- A plural or a definiteness marker, should be the first suffix to be appended to the noun stem directly (if any), while singular noun is determined by the absence of plural suffix;
- A given Oromo noun can be either plural or definite, but not both simultaneously;
- A case marker is attached or suffixed after a plural or definiteness suffix (if any), otherwise it can be appended to the noun-stem directly;
- One or more postpositions or enclitics can be suffixed after a case marker (if any), otherwise following a plural or definiteness marker (if any).

Figure 3.1 and Figure 3.2 illustrate the ordering of nominal suffixes in Afaan Oromo.

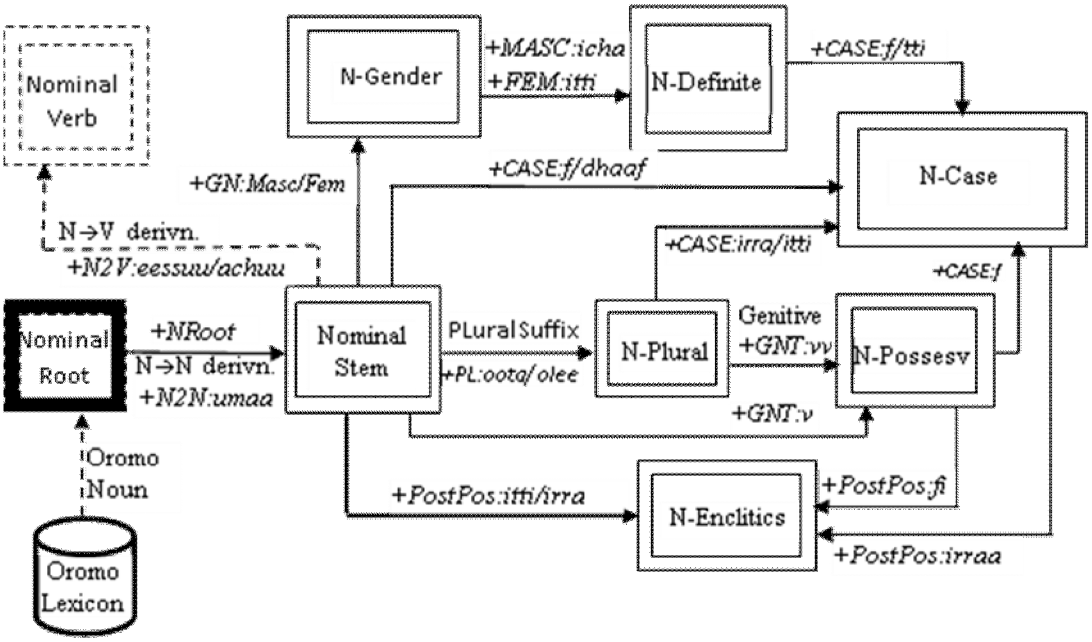

Figure 3.2 An FST representation for Oromo nominal inflection

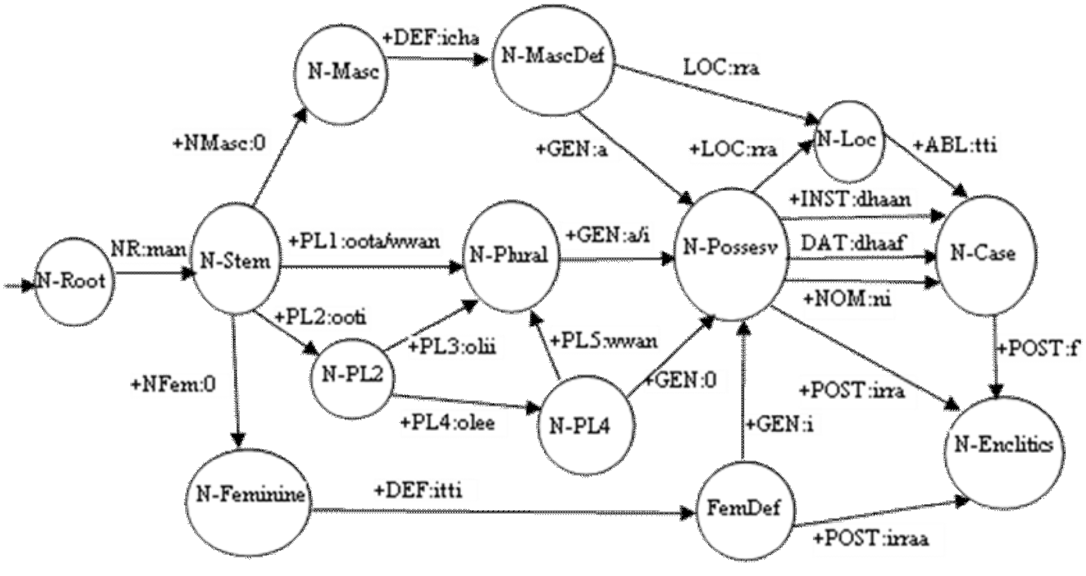

Figure 3.3 Example of FST representation for Oromo nominal inflectional suffixes

As indicated in Figure 3.3, there are multiple plural suffixes in Afaan Oromo, which may be concatenated in certain order to form a plural noun. Although it easy to observe that multiple and alternative plural markers can be used with a single root-noun, there is no obvious way to predict or determine which plural marker(s) could be used (or cannot be used) with a particular noun-root. A more detailed study of morpho-phonological properties of Afaan Oromo words is necessary to come up with certain linguistic constraints. For instance, to indicate the plural form of the Oromo noun "*mana*" (i.e. house), the following number different makers can be used:

- *"–oota"* (as in "*manoota*") or,
- *"–wwan"* (as in "*manawwan*" or,
- *"–oolee"* (as in "*manoolee*") and,
- *"–ooti"* (as in "*manooti*").

More surprisingly, a combination of these plural markers (as in "*manootawwan*") or ("*manootoleewwan*") is also possible, which is not common but certainly acceptable or permissible in Afaan Oromo. As discussed in section 3.2.1, the definite suffixes such as *"-icha"* (masculine) and *"-itti"* (feminine) varies with the gender of the noun they modify. In Figure 3.2 and Figure 3.3, we have treated that the plural and definiteness markers cannot co-occur together because we haven't seen any example of their co-occurrences in Afaan Oromo text and references materials we have consulted.

3.3.3 An FST Model for Oromo Verbal Inflection

Like an Oromo noun, Oromo verb consists of a stem and one or more suffixes. But, the morphological structures and processes in verbs seem much more complex than nouns. As indicated in section 3.2.2 (see Table 3.4) and in Figure 3.4, Figure 3.5 and Figure 3.6, an Oromo verb is conjugated to indicate Person, Number and Gender (i.e., PNG or subject agreement) as well as to mark tense, aspect and modality (TAM). Hence, the verb-stem is appended with different inflectional suffixes, which are necessary to express grammatical functions such as subject-verb agreement, tense, aspect, mood, causation and passive voice. In Figure 3.5 and Figure 3.6), a circle or a double square are used to represent final states (i.e., legal word). An arrow indicates a transition from one state to the next state upon receiving the input symbol indicated on or near the edge of the arrow. A transition can be made if the item (string) currently being processed matches the symbols specified by using the syntax of the two-level morphology. As indicated in Figure 3.5 and Figure 3.6, the basic structure of Oromo verbs consists of:

1. A ***verb-stem*** (or *root-verb)* that must be followed by,
2. ***Subject agreement*** (a fused person, gender and number or PNG marker), which in turn must be followed by,
3. A "***tense***" marker (or TAM suffix), and which can optionally followed by,
4. A ***case marker***, as well as
5. ***Postpositions*** (or verbal enclitics).

While marking a verb with the last two suffixes is optional, marking a verb for subject agreement and TAM is mandatory in Afaan Oromo. To illustrate this, a general morphotactics of verbal Oromo inflectional morphology is summarized and represented in Figure 3.4 as follows:

V_STEM + SubjAgr (PNG) + TAM + [***CASE***] + [***PostPos***]

Figure 3.4 Morphotactics of verbal inflection in Afaan Oromo

As shown in Figure 3.5, verbs like "*kuutu*", (which means to "cut"), must be marked to indicate subject agreement (i.e., a verb must agree with the gender, number and person) when the subject of a sentence is third person singular. A verb must be marked to show number and person agreement when the subject of a sentence is not third person singular. Moreover, a verb must be marked to express tense and aspect of the action as well as the mood of the speaker. Broadly, while a tense can be classified into past tense and present tense, aspect can be classified into perfective, imperfective and progressive in Afaan Oromo. Mood can also be classified into indicative, imperative, interrogative and conditional.

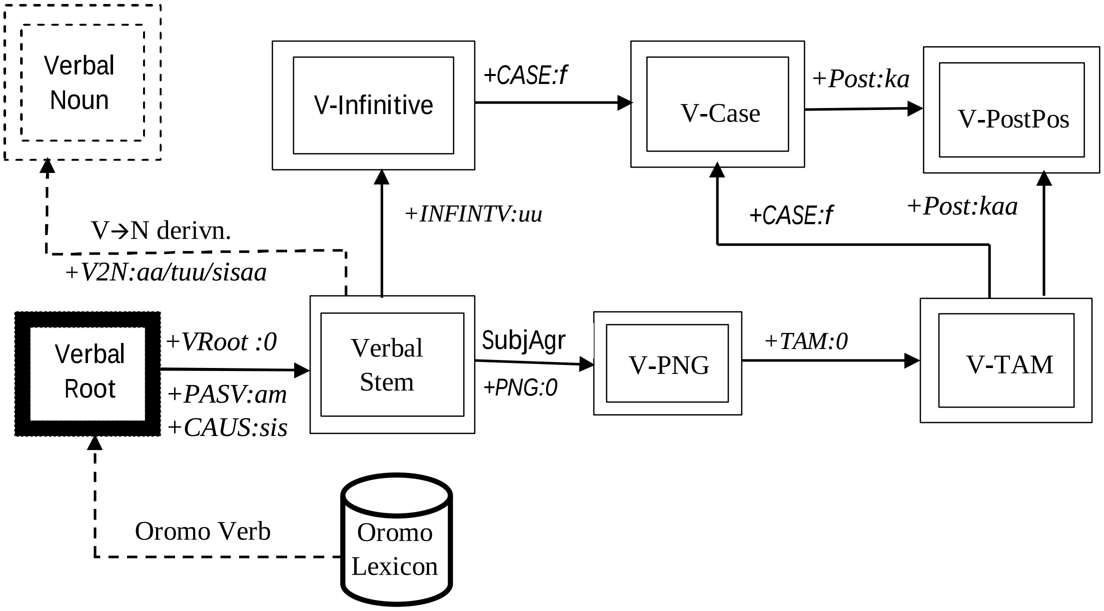

Figure 3.5 FST representation for Oromo verbal inflection

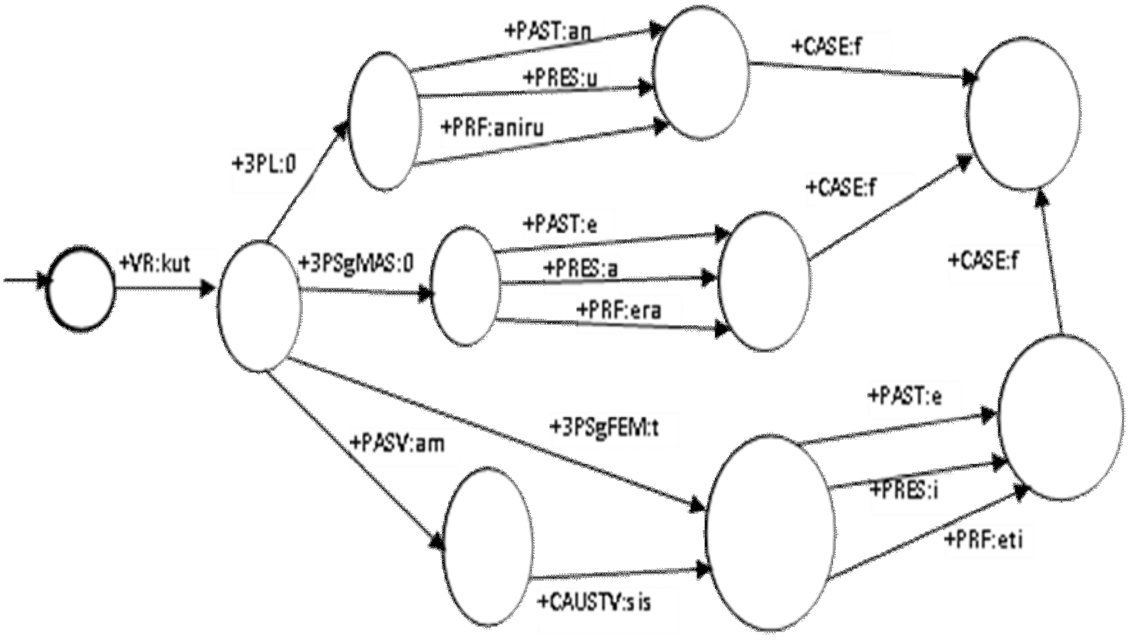

Figure 3.6 Example of FST representation for Oromo verbal inflectional suffixes

In summary, we have proposed and presented a general FST-based computational model for Afaan Oromo inflectional morphology, which is an essential prerequisite for developing Afaan Oromo morphological analyser and morphological generator. It can also play a pivotal role in designing and constructing IR tools such as stemmer and lemmatizer. As described in section 5.4, we have adopted and used this generalized computational model not only to design and construct our Afaan Oromo stemmer, but also to refine and improve its performance.

4 CLIR Issues and Approaches

4.1 CLIR vs. IR: A Historical and Conceptual Background

As described in section 1.2, IR is an interdisciplinary field of study that deals with systematic organization and management of very large unstructured documents. Its goal is to organise and retrieve documents relevant to the users' request. Traditionally, IR systems used to organize and compare bibliographic data (metadata) of documents against the users search requests to locate and identify relevant items in huge collections of documents. Systematic organization and management of bibliographic records (or surrogates) were considered as crucial to better enable the users to retrieve relevant items. Consequently, most of the initial IR methodologies were focused on classification and organization of manually indexed documents and bibliographic records such as library catalogues and bibliographic references. These IR systems were primarily used by librarians, information professionals and search experts. Unfortunately, the task of searching and finding relevant items was severely hampered by the need to use highly complicated Boolean queries. As automatic indexing and natural language queries gained popularity in the 1970's, IR systems became increasingly more accessible to non-experts and end users. Gradually, IR has evolved and become an information system that deals with indexing and discovering information from a very large collection of unstructured (or quasi-structured) documents [1, 29]. Electronics documents are often indexed by considering all terms used in them as independent keywords, in what is known as the Bag-of-Words (BOW) representation, a very simple but powerful model that is at the heart of most modern text search engines. As a result, the searching strategy has considerably simplified. Users can express their information needs in very short natural language queries.

According to [2, 7], IR is an academic discipline that researches models and methods to access and organize large amounts of unstructured and structured information. Its applications ranging from Web-based search to enterprise search, digital libraries and personalized retrieval systems. IR systems and search engines are increasingly being seen as critical technologies to find valuable information sources and services. Today, in the context where the problem of "*information overload*" is a day-to-day experience of most users, the role a robust IR system that instantly identifies and retrieves quality documents, cannot be overstated.

To satisfy the rapidly increasing and constantly changing needs of their users, IR systems must be able to manipulate and organize massive amounts information efficiently. Broadly, an IR system is composed of three major components [29]:

- I**ndexing** is concerned with the representation and organization of the information sources, to allow a swift and an instant access to relevant objects or items.
- **Searching** deals with identifying and retrieving items that can satisfy the information needs of the users.
- **Ranking** has been considered as an optional task, but it is very important for the retrieval task. It is in charge of sorting and ranking the search results based on certain heuristics that try to determine the items that better satisfy the information needs of users.

In an increasingly globalized and knowledge-based economy, the ability to seek and discover relevant information across different languages has become more important than ever before. Access to relevant information, regardless of the language in which it is represented, is crucial to understand and cope up with the constantly evolving and rapidly changing multicultural world. Unfortunately, classical IR systems and search engines are primarily designed to provide monolingual retrieval services in well-resourced languages. In other words, in order for the monolingual retrieval system to succeed, the search requests and the documents must be expressed or represented in the same language. As result, a vast amount of valuable information produced in different languages that are not supported by monolingual search engines is treated as irrelevant or unwanted content. Usually, information sources published in foreign (unsupported) languages are either ignored or excluded from the search results.

Nowadays, the influx of multilingual content on the Web comes from an unprecedented myriad variety of languages. Massive amounts of multilingual and multimedia content is increasingly created and disseminated via the WWW and social media networks. Correspondingly, the information need of users is no longer limited to their geographic and cultural boundaries. Rather, it transcends and extends beyond the linguistic and national borders. Hence, the task of searching information across different languages is not just become more important, but increasingly complex and challenging. In order to address this challenge, CLIR systems offer technically feasible and computationally efficient mechanism through which relevant information resources and services can be discovered and delivered to the rapidly growing number of web users beyond cultural and language boundaries [2, 95].

CLIR is a branch of IR that has gained considerable attention among IR researcher and language technology developers since the dawn of the WWW in the mid-1990s. As indicated earlier, while IR researchers have been focused on addressing the problem of organising and retrieving monolingual information resources and services during the last five decades, there is a rapidly growing interest in developing and providing cross-language search services in recent years. As noted by [9, 38], like IR, CLIR is an interdisciplinary field of study. It is an interdisciplinary research area in which methodologies and tools developed for information retrieval and natural language processing converge. In other words, CLIR is built upon more than five decades of research and development in IR and MT systems. Broadly, the purpose of CLIR is to enhance information access across language and cultural boundaries. It is based on the fundamental principle that online information resources and services should be efficiently and equitable accessible to all users, regardless of their linguistic and cultural backgrounds.

While there are numerous search engines that are currently in existence, a few of them have recently started to provide cross-language search services in well-resourced European and Asian languages. Although most search engines are still monolingual, some of them have already recognised the significance CLIR and have added the functionality to carry out cross-language search services in resource-rich European and Asian language pairs. For instance, Google's cross-language search can be considered as a typical model for integrates MT and CLIR technologies to help users find information on the Web that is not written in their familiar languages. According to [82], the launch of the cross-language search by Google was considered as a breakthrough event in the field of CLIA since it has signified the transition from CLIR research to its actual application on the Internet. It was the first time that CLIR and MT were functionally integrated to provide a real time retrieval and translation services online. In recent years, a number of more specialized areas of CLIR have been emerged to address the problem of language barriers in accessing multimedia content or rich media data such as images, videos and music. In particular, there is an increasing interest in cross-language image and video retrieval. Confirming this fact, [2] pointed out that one of the rapidly growing application area of CLIR is the retrieval of images or rich media data that are provided with brief textual descriptions in any language.

Some of the major milestones and turning points in the growth of the IR/CLIR research over the last five decades is briefly summarized and presented as follows [7]:

- **1970/73**: Gerard Salton’s initial CLIR experiments by using hand coded bilingual term lists or dictionary, an experiment on English-German CLIR.
- **1978**: Publication of ISO Standard 5964 Multilingual Thesauri.
- **1996**: SIGIR Cross-Lingual IR workshop.
- **1997**: TREC6-CLIR track had started cross-language retrieval evaluation experiments in English and other major European Languages. In early 2000s, TREC-CLIR track had expanded to include Arabic and other major East Asian languages.
- **1998**: NTCIR (National Institute for Informatics Test Collections for IR) CLIR tracks for East-Asian (CJK) languages was launched in Japan.
- **2000**: Cross-Language Evaluation Forum (CLEF) was launched in Europe, with special focus on major European Languages.
- **2008**: Forum for Information Retrieval Evaluation (FIRE) for Indian Languages.
- Information access across languages on the WWW:
    - ➢ **2006**: Yahoo offered cross-language retrieval services in some of the major European languages such as French, German and English.
    - ➢ **2007**: Google started query translation services as well as translation of retrieved webpages.
    - ➢ **2008**: MS Bing offered query and document translation services.

In summary, although CLIR shares a number of IR models and techniques for organizing and retrieving information, it differs from IR in many significant ways. While managing the massively growing volumes of monolingual collection is very difficult, this task becomes even more difficult and complicated in CLIR owing to the problem of language barriers and linguistic digital divide. Over the last two decades, CLIR researchers have addressed many important topics related to document and query translations as well as issues related multilingual document retrieval. Unfortunately, since the main focus of most the earlier studies have been on better-resourced European and Asian languages like English, Arabic, Chinese and French, there has been very little work done on the development and application of CLIR for resource-scarce African and Asian languages. There is a compelling need for the development and application of CLIR for resource-scarce African and Asian languages to enable their native speakers to access information beyond language barriers. Otherwise, the vast wealth information available on the Web remains inaccessible to the vast majority of the world citizens.

4.2 Basic Issues and Problems in CLIR

The main objective of cross-language search is to discover relevant information beyond the boundaries of languages and cultures. Yet, there many issues that have been hindering efficient and equitable information access across different languages and culture. This section describes and outlines some of the fundamental research issues in IR in general and in CLIR in particular.

4.2.1 Word Form Variations

One of the major problems in IR in general, and a dictionary-based CLIR in particular, is related to word form variations, which is caused by inflectional and derivational morphology. In analytic languages like English, the occurrence of word-form variants is very limited. On the other hand, in highly synthetic and agglutinative languages like Afaan Oromo, most words in written or spoken texts are occurring in their inflected forms. As noted by [95], most IR and CLIR systems suffer from the problem of word form variations since it results in the failure of direct query-term-to-text-word matching or query-term-to-lexical-item matching. In order mitigate this problem, CLIR systems need to be able to distinguish different variants of the same lexical item.

As indicated in section 3.2, most inflectional morphemes in Afaan Oromo consist of suffixes. But, there are some prefixes that are used to indicate various grammatical categories. Unlike in English, since inflectional morphology is extensively productive in Afaan Oromo, lexical resources do not cover word form variants as part of their vocabulary entries. As a result, most search terms or key words expressed in Afaan Oromo may not be found in general lexical resources such as a bilingual dictionary. Detecting and normalizing word form variations in a search query is, thus, considered as an essential prerequisite for achieving a good CLIR performance.

As discussed in section 3.2, inflected words are formed by affixing or attaching bound morphemes such as case, number, gender, tense and aspect markers to the stem or root-word. In morphologically very rich and agglutinative languages like Afaan Oromo, inflectional morphology is extensively productive, meaning they can produce extremely large number of word form variants from a single word-root or lexical entry. Hence, it is very difficult to list all inflected Oromo words in a general dictionary. In other words, inflected words are excluded from most Oromo lexicons, such as bilingual dictionaries and multilingual thesauri. Apart from the sheer amount of time and resources needed to include all inflected words, incorporating all possible word forms in a general dictionary would make the database highly susceptible to error and extremely difficult to maintain. Hence, it is crucial to devise an efficient mechanism that can identify a lexical-root (base-word) from which word form variants (or surface forms) are derived. In order to obtain the stem of inflected words, morphological affixes must be identified and analysed by appropriate morphological processing tools such as stemmer, lemmatizer and morphological analyser. As noted by [81], the challenges posed by inflectional morphology can be reduced or tackled by stemming and lemmatization, where every word is reduced to its uninflected stem or lemma. Stemming is a procedure where different grammatical forms of a word or lexical item are reduced to a common base-form (but not necessarily the lemma) called a stem, through successive removal of agglutinative suffixes. In this study, we have designed and developed a rule-based stemmer in order to deal with the problem of inflectional morphology in Afaan Oromo.

4.2.2 Phrasal Terms and Compound Words

For the effectiveness of CLIR, translation of phrasal terms or multi-word expressions (MWEs) in their entirety, rather than individual word-for-word translation, is very important. Phrasal terms matched against a manually built multi-word (phrase) dictionary showed higher precision than those translated by single word-based dictionaries [81, 19]. While dictionary-based CLIR method provides a simple and efficient approach for query translation it suffers from limited coverage of phrasal terms and compound words. A compound word is a word formed from two or more words. Although compound words are not widely used in English and Afaan Oromo, the use of compound words in Afaan Oromo texts has been significantly increased in recent years. Compound words and phrasal terms are often used to capture the meaning of new technical terms and scientific concepts in Afaan Oromo. Since some of these compound words are separated by hyphen or space, they can be easily decomposed to two or more words, where each word may have related or unrelated meaning (compositional compound vs. non-compositional compound words). For example, an Oromo phrasal term or compound word "*galmee jechoota*", (which means a "*repository of words*", when each of the words are independently or literally translated, stands for "dictionary" when the phrasal term is translated as a unit). Hence, such compound words cannot be accurately translated when each of the word is treated as an independent query term.

In a dictionary-based CLIR, a word-level translation, or word-for-word substitution is commonly employed to convert the source language query terms into the target language search terms. In other words, the CLIR engine attempts to find a translation for each query word in the source language. Other levels of translations that can be employed in MT-based or corpus-based approach include phrase level and sentence level translations, where both syntax-based and semantic-based analysis techniques can be employed to achieve more accurate query translation. A knowledge-based translation, which also considers the structure and semantics of the search query, could be employed to translate phrasal nouns and compound words more accurately. The desired translation is the one that captures and expresses the exact meaning of a compound word in the source text with correct syntax and semantics [81]. Due to the lack appropriate linguistic resources, we were not able to address the problem of phrasal terms and compound words in our current study.

4.2.3 Named Entities and Acronyms

The translation of proper names is quite different from the translation of other common words and lexical items that can be found in bilingual dictionaries. It is also difficult to generalize such issues across different languages and to find a general solution. In other words, cultural and language-specific knowledge is needed to recognize and translate named entities properly [81]. In English, it is relatively easier to recognize proper nouns and abbreviations since they are not often inflected for grammatical functions. Like in English, Oromo proper nouns start with a capital letter whereas acronyms or abbreviations are formed with capital letters which are often separated with period symbol. However, proper nouns in Afaan Oromo are marked with different inflectional affixes, just like any other Oromo nouns (see section 3.2.1). For example, an Oromo personal name "*Bulchaa*" may take various case markers such as:

- *Bulchaa + n* → "*Bulchaa**n***" (nominative);
- *Bulchaa + f* → "*Bulchaa**f***" (dative or beneficiary, i.e., for Bulchaa);
- *Bulchaa + dhaa* → "*Bulchaa**dhaa***" (dative or beneficiary, i.e., for Bulchaa);
- *Bulchaa + itti* → "*Bulchaa**tti***" (dative or beneficiary, i.e., to Bulchaa);
- *Bulchaa + dhaan* → "*Bulchaa**dhaan***" (instrumental, i.e., by Bulchaa).

The task of identifying, normalizing and translating proper nouns and acronyms are, therefore more difficult in Afaan Oromo than in English. CLIR lexicons and general bilingual dictionaries lack a good coverage of proper names and acronyms [19, 81]. A common method used to handle untranslatable named entities is to include them untranslated in the target language query. If this word does not exist in the target language, the query will be less likely to retrieve irrelevant documents [81]. Alternative methods exist to deal with this problem for languages of the same writing system like English and Afaan Oromo, which may include transliteration and transformation rule based translation. In transliteration techniques, a word in one language is often matched to a word in other language based on regular correspondences between the characters of the two languages. In other words, the source language query terms and the target language query terms are regarded as spelling variants of each other. For example, the word "*komputara*" in Afaan Oromo can be matched with "computer" in English by replacing "*k*" with "*c*" and "*ara*" with "*er*". Similarly "*filmii*" can be matched with "film" by deleting the final long vowel of the first word. Usually transliteration is used in conjunction with certain phonetic rules or fuzzy matching such as the n-gram matching techniques. In the n-gram method, search keywords are decomposed into n-grams (sub-strings of length n), then the degree of similarity is computed by comparing their n-gram sets [81]. As described in section 5.5, in this study, we have employed a transliteration technique in order to convert untranslated Afaan Oromo query terms into English search terms.

4.2.4 Technical Terms and Concepts

In most professional journals and research papers, technical terms and scientific words are considered as very important elements of a document. Hence, dictionary coverage of special terms and latest technological concepts is very important for the effectiveness of CLIR [19, 24, 80]. Usually, special terms are not widely covered in general dictionaries. But, special terms can be matched against a special dictionary, (in case such a highly specialized and domain specific lexicons are available for the language pairs being considered). For instance, query terms related to medicine can be matched against lexical entries in a medical bilingual dictionary. Combining both general and specific domain dictionaries enhances the performance of CLIR [81]. Two techniques are used to combine both dictionaries [93]. Sequential translation translates the query keywords against the specific domain dictionary. If it fails to match, it uses the general dictionary, and a parallel translation that matches query keywords against both general and specific dictionaries. Both these techniques reduce the special terms translation problem but do not solve it altogether. For instance, translating a newspaper article that contains scientific terms, technical terms, political terms etc. needs multiple and multilingual domain specific dictionaries [81]. As indicated in the preceding section, we have employed a transliteration technique to convert untranslated Oromo technical terms into English search terms.

4.2.5 Lexical Ambiguity

Ambiguity is another source of translation errors in CLIR and MLIR. As noted by [81], two of the major causes of lexical ambiguity in natural languages are homonymy and polysemy. Where homonymy refers to a word that has at least two entirely different meanings (such as “bark”, which can mean the skin of a tree or the voice of a dog), polysemy refers to a word which can take on two distinct, but related meanings (such as the “head” of the body, and the “head” of a department). However, the distinction between homonymy and polysemy may not clear cut. Lexical ambiguity covers both homonymy and polysemy. In CLIR, translation ambiguity arises owing to source and target language lexical ambiguity. Search queries are usually too short to provide sufficient contextual information about the user’s interest or needs. This makes the task of disambiguation search terms even harder and complex. According to [81, 58] most queries are unclear and very short because it is not easy for users to explicitly express their search intents. For instance [81], some users do not choose appropriate words for a web search, and others omit specific terms needed to clarify search intents. This gap between the users’ search intents and search queries results in queries that are ambiguous and broad. For ambiguous queries, users may get results quite different from their intentions; for broad queries, results may not be as specific as users expect [81]. Many methods have been developed to decrease the ambiguity in query translation, such as part of speech tagging, corpus based disambiguation methods, query structuring [24, 80], and the most probable translation strategy. Since such core linguistic resources are not readily available for under-resourced African languages, we have not addressed the problem of WSD in the current study.

4.3 Query Translation Approaches and Techniques

As described in sections 1.2 and 4.1, CLIR is concerned with systematic organization and retrieval documents across language boundaries. Unlike in IR, the search queries and the documents do not share the same language in CLIR. Elaborating this fact, [2, 95] stated that there are huge cultural and linguistic differences that prevent users from browsing and retrieving relevant information. Bridging linguistic disparity between the query and the documents is one of the fundamental issues that must be addressed by CLIR researchers. Indeed, one of the major research tasks in CLIR is concerned with devising an efficient mechanism to overcome language barriers. Usually, the translation of either the documents or the search queries is necessary in order to overcome language barriers. A few of the basic questions that must be addressed in CLIR are thus:

- What should be translated (queries or documents)?
- How to translate them? and,
- Which translation resources and techniques are more feasible and cost effective?

As noted by [19, 94], either the search queries or the documents must be translated into a common representation in order to enable users to search and retrieve relevant documents across the language boundaries. Since documents are represented in a language different from that of the query, the basic strategy of monolingual IR, which directly matches documents against queries, cannot succeed in searching and identifying relevant items. In a typical monolingual IR, there is no language barrier that will require the process of translating queries into the language of the target documents or vice versa. A general overview of CLIR approaches is illustrated in figure 4.1.

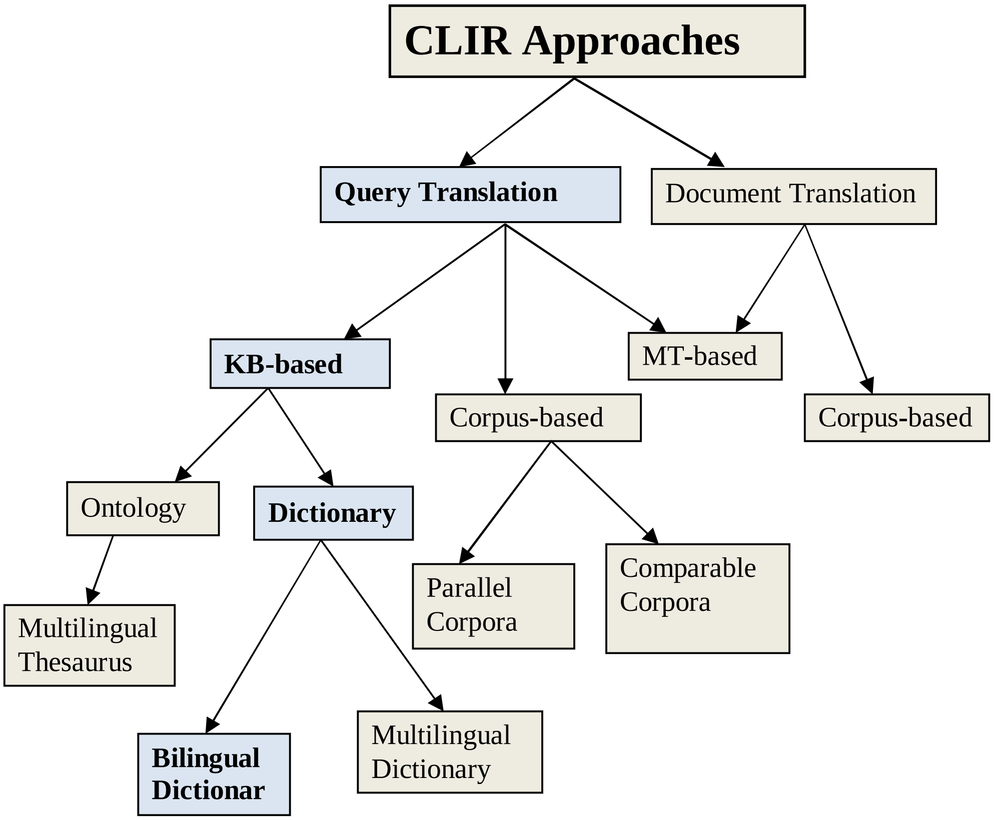

Figure 4.1 Overview of CLIR approaches and techniques

4.3.1 Query Translation vs. Document Translation

One of the main research questions arising in developing CLIR is, how the language gap between the search requests and the documents could be efficiently bridged and crossed. Since translation is crucial for successful development of CLIR, the IR research community has developed a range of translation resources and techniques over the last two decades. In a query translation based approach, the CLIR system translates the user's queries into the language that the documents are written in. This method presupposes that the query can be translated in a reasonably accurate way and that monolingual retrieval systems are available for the languages of the target documents [9, 17]. Document translation is the reverse of query translation where documents are translated into the query language. Among these two approaches, query translation technique has been applied by most CLIR researchers because of its simplicity and effectiveness [52, 58].

As noted by [2], it is generally believed that query translation is the most efficient CLIR approach: given a query, the user is allowed to choose the language of interest, and then the query will be translated into the desired language. However, query translation often suffers from the problem of translation ambiguity, which is amplified due to the limited amount of context in short queries. From this perspective, document translation seems to be more capable of producing more precise translation due to richer contexts. The availability of MT systems also makes the document translation approach possible for well-resourced language pairs. However, it is not obvious that the current MT systems can take full advantage of the rich contexts in a document during the translation process. Generally, translating the whole document collection to the languages of the users' queries is more demanding and very expensive, as it requires very scarce resources like full-fledged real-time MT system, which is not available for most under-resourced African languages.

As noted by [19], query translation is more feasible and efficient than document translation. Queries are easier to translate because they are typically short and can be translated as "bag-of-words", whereas document translations have to obey more complex rules and procedures of natural language processing. In addition, query translation may offer the flexibility of adding cross-lingual capability to an existing monolingual IR engine by incorporating appropriate translation resources and modules. A general overview of CLIR approaches is presented in figure 4.1.

According to [2, 7], while the major disadvantage of query translation is lack of sufficient linguistic contexts for word sense disambiguation, the major disadvantage of document translation (where context is not a problem) is the overwhelming cost in terms of computing resources needed to translate extremely large number of documents. In fact, translating a very large document collection may become very difficult or even impractical in the context of open domain CLIR. Taking these facts into account, we have adopted a dictionary-based query translation approach in developing our first Oromo-English CLIR. As indicated in figure 4.1, although various language resources and translation techniques can be employed for the purpose of query translation, three major approaches are more dominant in CLIR [80], namely:

- dictionary-based approach: translation of query topics using machine readable bilingual dictionaries;

- MT-based approach: translation of query topics by using existing Machine Translation system, and;
- Corpus-based approach: translation of query topics by using parallel or comparable corpora.

Each of these methods of accomplishing query translation has a different set of strengths and weaknesses and each requires different resources to build (see Table 4.1). These three major query translation techniques are described in more detail with special focus on their advantages and disadvantages in the subsequent sections.

4.3.2 MT-Based Query Translation Approach

As indicated by [95, 2], a query translation based CLIR system may use various translation resources, such as bilingual dictionaries, MT system, parallel texts, or a combination of these linguistic resources to translate queries from a source language into another target language. Machine translation (MT) system is one of the major natural language processing applications that can be easily adopted for the purpose of query translation. MT automates the process of language translation, which normally includes analysing and understanding information in one language and expressing it in another language. During the last half century, MT systems have slowly evolved and transformed into commercial applications and translation services. Current commercially available machine translation services, although still not good enough to replace human translations, are able to provide useful and reliable support in many multilingual natural language processing applications including CLIR and cross-language web browsing [2, 7].

However, building a reliable MT is a very difficult and complex task because it involves understanding and interpretation of the connotative meaning in a source (original) language and its expression in a target language using correct terminology and syntax. In particular, translating a search query is not identical to the task of full-text translation in MT. The goal of translating a query in CLIR is not to produce a human-readable translation, but a translation suitable for searching and retrieving documents across language boundaries.

As indicated in [81], MT-based query translation approach has been widely used for development of CLIR for many European and Asian languages. Unfortunately, since the scope of most of the existing MT systems is limited to better-resourced European and Asian languages, the majority of resource-scarce languages lack well-establish MT systems. Moreover, because the search requests submitted by the users tend to be very short, a single translation output per query may become less accurate. Queries are often too short to provide sufficient contextual information for detailed linguistic analysis and natural language processing. Because a typical user's query is often represented as a set of 2 or 3 keywords, it is difficult to expect MT systems to work well against such very short and unstructured search phrase. While MT-based CLIR approach may works relatively well for some languages, [2] argues that it is less optimal in most CLIR and MLIR settings.

According to [95, 2], an ambiguity problem exists in the MT components, since the translated query does not necessarily represents the sense of the original query. For instance, translating the English query *big bank* to another language could produce an inappropriate translation since it is not clear whether "*bank*" means the institution or the edge of a river. MT systems normally attempt to determine the correct word sense for translation by using context analysis. However, a typical search engine query lacks context as it consists of a small number of keywords. Short queries are usually insufficient to describe the need of the user in a precise and unambiguous way, and this makes MT-based query translation even harder and sometimes insufficient. Moreover, commercial MT systems tend to deny CLIR researchers the opportunity to modify and improve the query translation accuracy.

While most of the existing MT systems are trying to translate search queries into a well-governed word order and syntactically correct statements, most of the existing IR systems are not sensitive to word order and grammatical structure of the search queries. In fact, most of the existing IR systems are based on bag-of-words models. Unlike MT system, the goal of query translation in CLIR is not to generate syntactically correct representation of a search queries, but to cull through the tremendous number of documents and to select those which are pertinent to the user's requests. Hence, some of the detailed linguistic analysis and natural language processing procedures that are very important in MT system may not be relevant for CLIR. Instead, an approximate translation of important keywords that captures the gist of the original user's query is typically sufficient for CLIR. This implies that less sophisticated translation resources such as bilingual dictionaries and multilingual thesaurus could be sufficient and more efficient for translation of search queries. In situations where there is a large collection of documents or when searching for documents on the web, machine translation might become impractical or infeasible, especially for many resource-poor African languages like Afaan Oromo.

In general, although MT system can be easily adopted and used for query translation, the translation output that is obtained through such advanced natural language processing application may become less accurate or not appropriate for the purpose of CLIR. For instance, MT systems usually try to select only one translation from different alternative translations available for a given search query. By limiting the possible translations of a search query to only one candidate equivalent translation, the MT system might have restricted or prevented the CLIR system from expanding the original search query by incorporating additional synonyms and other related keywords. Thus, MT-based query translation approach may need to be complemented by other linguistic resources and translation techniques in order to produce more appropriate translation of search queries. On the other hand, since building a reliable MT system is very expensive, it is not yet readily available for most resource-scarce African languages like Afaan Oromo. In other words, MT-based query translation approach is not yet a viable technique in developing CLIR for indigenous African languages.

4.3.3 Corpus-Based Query Translation Approach

Broadly, parallel corpora can be defined as textual documents accompanied by their corresponding translations in one or more other languages. It is collection bilingual or multilingual electronic records (texts) that have been translated to each other and placed together for the purpose of natural language analysis and statistical applications [96]. In a rapidly expanding multilingual environment like the WWW, the role of bilingual collections and linguistic data such as parallel and comparable text corpora cannot be overestimated because they makes the effort to develop, train and test multilingual information access systems more efficient and effective. In the context of multilingual natural language processing and CLIA, parallel corpora are the cornerstone to the development of robust language processing tools and applications, such as parts-of-speech taggers and syntactic parsers as well as CLIR and MT systems. The value of very large-scale parallel texts is widely accepted and there have been lots of efforts to make such collections available online, although most of them are focused on well-resourced European and Asian languages [97]. Unfortunately, resource-scarce African languages are not yet in a position to enjoy the luxury of large parallel corpora. For severely under-resourced African languages like Afaan Oromo, lexical translation resources are the only viable option.

One area that crucially depends on parallel data is the task of building language models for Statistical Machine Translation (SMT) [14]. A corpus-based CLIR approach also makes use of the statistical information of term usage in a parallel or comparable corpus to automatically construct bilingual dictionaries and thesauri. Many empirical studies have suggested that the performance of most information systems based on statistical approach is often a function of corpus quantity and quality [58, 96]. This unfortunately also means that most resource-scarce languages are disadvantaged since the full potential of such linguistic resources cannot be released without the development and availability of sufficient text corpora. Most researchers who are working on better-resourced European languages have been able to make use of high-quality corpora consisting of books, well-edited national newspapers, and the like. Sadly, corpus-based methods are not yet realistic approaches for most indigenous African languages. Parallel linguistic data, even though very useful, is expensive to obtain and hence it is not available in sufficient amount for most of resource-scarce African languages, including Afaan Oromo.

Parallel corpora are valuable resources for obtaining query translation knowledge [96]. The growing availability of online multilingual texts has given rise to the application of corpus-based CLIR systems. Parallel and comparable corpora are increasingly become important translation resources for development and application of CLIR. Over the past few years, various aligned parallel and comparable corpora have been established for a number of European and Asian languages. These aligned parallel and comparable corpora have been widely used as a primary source of knowledge for query translation. Nowadays, most of the corpus-based CLIR methods are trying to exploit the availability of very large multilingual texts on the Web. One of the major advantages of employing corpus-based query translation approach is that it can provide multiple equivalent translations of a given search query. As a result, the translated search query may contain exact translations as well as related translations of the original search query. Accordingly, a corpus-based query translation approach can easily provide a good opportunity for relevance feedback and query expansion [96, 24].

Unfortunately, multilingual corpora, including comparable and parallel corpora are seldom available for severely under-resourced language like Afaan Oromo. Hence, one of the major problems of adopting this approach for query translation is the scarcity or unavailability of sufficient parallel corpora for under-resourced languages. Since the task of designing and building reliable parallel corpora is very expensive and time consuming, the employment a corpus-based query translation approach is not feasible for severely under-resourced African languages like Afaan Oromo. Another limitation related to corpus-based query translation is the coverage and quality of linguistic data. Poor quality corpora lead to less accurate translation, which will decrease the performance of a CLIR system. Due to lack of parallel corpora and MT systems for Afaan Oromo and English, we have focused on building and adopting a dictionary-based query translation approach in this study.

4.3.4 Dictionary-Based Query Translation Approach

In dictionary based query translation technique, keywords in a given query are translated into the target language using bilingual Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. A bilingual machine-readable dictionary is one of the most important lexical resources that is widely used by the developers of CLIR to overcome language barriers. Employing a bilingual or multilingual dictionary as source of knowledge for query translation is one of the most commonly used approaches in CLIR [18, 94], especially when reliable MT systems and parallel corpora are not readily available for the language pairs that are being considered. In this study, we have used this approach to design and develop our Oromo-English CLIR. Over the last two decades, a dictionary-based query translation approach has been widely used for development and application of CLIR systems [80, 10, 65, 26]. Since various bilingual and multilingual dictionaries are increasingly made available on the Web, a dictionary-based query translation approach remains very popular in CLIR. Even if a reliable online bilingual dictionary is not available, converting existing human-readable or printed bilingual dictionary into a machine-readable bilingual dictionary is much cheaper than building a new MT system or parallel corpora. Since human-readable bilingual dictionaries are widely available for many less-resourced languages, digitizing these printed dictionaries and adopting them for the development of CLIR is much more feasible than the other two alternative approaches.

Dictionaries are organized according to different principles. For CLIR, dictionaries are usually considered as a word list, together with their translations [2]. A dictionary-based query translation involves various steps to process and transform a search request expressed in a source language into a target language. Initially, the search request expressed in a source language has to go through various pre-processing stages such as segmentation, tokenization and stemming. Then, each of the query term is looked up in the bilingual dictionary and translated to the target language whenever a matching is found in the bilingual dictionary.

According to [75], dictionaries are usually used for a word-by-word translation. Given a source-language word in a query, the first question one should ask is what translation is appropriate and should be chosen. Unfortunately, many available bilingual dictionaries do not contain useful information to help select the appropriate translation words or expressions. In such a situation, two basic approaches have been proposed [2]:

1. Using all the translations for each query word;
2. Using the first translation listed in the dictionary.

The first approach is motivated by the fact that when all the translations are used, one can include all the possible expressions in the target language and obtain a query expansion effect. However, this is done at the cost of introducing incorrect translations due to ambiguities. In fact, many words in a language may have more than one meaning. The fact that incorrect translations may be included in the query translation will lead to retrieving irrelevant documents related to the incorrect meaning of the original word. As a result, the increase in recall is often gained at the cost of decrease in precision [2].

The second strategy, which uses the first translation listed in the bilingual dictionary is motivated by the fact that the first translation is often the most important and thus frequently used one. This is, of course, dependent on the way in which the bilingual dictionary that is being used for query translation is organized. In doing so, one expects to have a higher chance to obtain the appropriate translation. Similarly, when frequency information is available, one can also choose the most frequent translation word. This strategy is similar to the idea of using the default translation in MT when no additional information is available [2]. However, the assumption on the organization of the dictionary is not true in many dictionaries. For the dictionaries that are organized according to the frequency of translation words and phrases, this strategy can help filter out some incorrect and rarely used translations. However, it also prevents one from having multiple translations for the same word. Since our Oromo-English bilingual dictionary is a medium-size dictionary with limited multiple translations for each lexical entry, we have applied the first approach in translating Afaan Oromo queries into English queries.

Even though a dictionary-based query translation approach has fewer disadvantages that are associated with MT-based approach, it has its own drawbacks. For instance, query terms translated by using a bilingual dictionary may become unstructured, incoherent and ambiguous. Yet, the impact of such noisy translation may not be significant in the context of CLIR since most of the existing IR systems can operate on bag-of-words model, (i.e., without considering the grammatical structure and word order of search query.) In fact, most of the existing major retrieval systems are still based on bag-of-words principles, in which both query statements and document texts are decomposed into a set of words (or phrases) through the indexing process. Unlike MT system, what is more important in CLIR is not the syntactic correctness or coherence of the search queries but the coverage and translation of important terms or keywords.

Generally, terms available in a dictionary are always limited because language is constantly evolving, and there are new words being created from time to time. Many technical terms, abbreviation, names of persons, organizations, and events may not be included in the dictionary. As shown in Table 4.1, some of the major limitations associated with a dictionary-based query translation technique include [19, 94, 17]:

- **Limited coverage**: Proper nouns, technical terms or domain specific special terms may not be included in general bilingual dictionaries. Since such important terms cannot be translated by consulting the bilingual dictionary, additional translation mechanisms and strategies may need to be devised to deal with out-of-vocabulary words.
- **Word form variations**: Queries submitted in morphologically complex languages like Afaan Oromo may not match with the vocabulary entries provided in the bilingual dictionary due to word form variations. Word form variations are often resulted from inflectional and derivational affixes. Hence, morphological tools such as stemmer and lemmatizer should be designed and employed for normalization of inflected query terms.
- **Multiword expressions and compound words**: Most of multiword expressions including phrasal terms and compound words are not covered in general bilingual dictionaries. Failure to identify and translate multiword concepts such as phrasal terms and compound words will reduce the effectiveness of a CLIR system.
- **Lexical ambiguity**: Sometimes the translation outputs of a bilingual dictionary might become ambiguous. They might also add extraneous information to the translated search

queries. Thus certain disambiguation mechanism may need to be devised to reduce the problem of lexical ambiguity that rises during the query translation process.

A number of CLIR researchers have investigated and proposed various tools and techniques to tackle some of the above major limitations. While morphological analysers and stemming algorithms are commonly used to handle word form variations such as inflectional and derivational affixes [97, 56], various statistical and transliteration schemes are adopted to deal with out-of-vocabulary (OVC) words. Moreover, relevance feedback and query expansion techniques have been proposed to reduce the impacts of ambiguous and extraneous translation of query terms. Different phrasal term identification and de-compounding techniques have been also developed to facilitate the decomposition and translation of multiword concepts in a search query for many well-resourced European and Asian languages.

In summary, some of the limitations and strengths of the major three query translation approaches in CLIR are briefly summarized and presented in Table 4.1.

CLIR Approaches	Strengths	Limitations
MT System	The task of processing and translating a search query is almost automatic.	Less optimal in CLIR settings where the bag-of-words model is used.
	Can be easily adapted for both document and query translation purposes.	Suffers from the negative aspects of single-selection translation.
	More effective for document and long query translations	Short queries lack sufficient context for MT since they are often composed of 2 or 3 keywords.
	Problems of word form variations and WSD can be handled by the system.	It is very expensive and prohibitive for resource-scarce language pairs. It is not readily available for most resource-scarce African languages.
	Problems of multiword expressions and phrasal translation are less challenging.	Off-the-shelf or commercial MT systems deny CLIR researchers the opportunity to modify the system to improve the query translation accuracy.
Parallel Corpora	Translation cab be done ether at the word level or phrase level.	Computationally more expensive and difficult to implement. Does not available for most resource-scarce African languages.
	Can reflect the dynamic nature of language by incorporating latest terms or concepts.	Extracting translation model (or knowledge) from parallel corpora is very difficult.
	Provides a mechanism to expand the original query by incorporating related keywords from corpora	Require large enough amounts of parallel corpora and annotated linguistic data
	Statistical or linguistic techniques can be employed to derive topic-specific technical terms and named entities from parallel corpora.	A corpora with a very limited quantity and poor quality may results in the failure of the entire query translation.
	Provides extremely rich context information to handle WSD and compound word translation.	Relatively more expensive than bilingual dictionary. High quality parallel corpora are not available for most resource-scarce language pairs.
Bilingual Dictionary	Relatively less expensive to develop and much easier to obtain. Available for many resource-scarce languages.	A limited vocabulary coverage, absence of named entities and technical terms may leads to the failure of query translation process.
	Effective in CLIR settings where the bag-of-words model is used.	The problem of word form variations can considerably reduce the performance of CLIR.
	Can reduce the negative impacts of single-selection translation by incorporating alternative translations.	Translation errors due to ambiguous search terms, i.e., it lacks an efficient tool for WSD.
	Computationally less expensive and easier to implement.	Some multiword expressions and phrasal terms cannot be translated accurately.

Table 4.1 Limitations and strengths of query translation approaches

4.4 CLIR Evaluation Methods

4.4.1 Why and How to Evaluate CLIR?

Performance evaluation is crucial for developing and improving CLIR systems. It allows the researchers to assess and determine the effectiveness of the retrieval system [7, 98]. There is general consensus among researchers on the significance of evaluation for IR systems. Basically, there is a duality between research and evaluation [7, 28]. Good research is validated by evaluation and good evaluation environments stimulate further research. With the rapid development and expansion of the multilingual Web, the development and evaluation CLIR have received a considerable attention from IR researchers in recent years. The problem of how to conduct CLIR evaluation to determine its effectiveness has not only been one of the most active research areas over the recent few decades, but also the subject of much discussions and debates [2, 98]. Different methods have been proposed for assessing the performance of IR and CLIR systems, ranging from more user-oriented approaches to those more focused on evaluating system performance. A few of the fundamental questions that should be considered in CLIR evaluation may include:

- Why conduct an evaluation,
- What should be evaluated,
- How the evaluation should be conducted.

Evaluation of CLIR involves identifying suitable criteria that can be measured in a certain level of quantity and quality. Such evaluation might related to whether a CLIR system retrieves relevant (compared with non-relevant) documents; how quickly results are returned and whether users are satisfied with the results [99, 98]. For judging the relevance of the retrieved items, it is possible to study if the results returned in response to a given query are related to it or not. This can be done by determining, given a query and a set of documents, the ones that are related (i.e., are relevant) and the ones that are not, and then comparing the number of relevant results returned by the retrieval system [2]. In IR and CLIR, there has been a special focus on measuring system effectiveness: the ability of the retrieval system to discriminate between documents that are relevant and not relevant for a given query set. In order to measure the effectiveness of a CLIR in a standard manner, it is often necessary to obtain a test collection consisting of three sets [98, 99]:

- A document collection
- A test suite of information needs, expressible as queries
- A set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair.

As indicated by [98], test collection-based evaluation is highly popular as a method in both IR and CLIR evaluation. Benchmarks can be used by multiple researchers to evaluate in a standardised manner and with the same experimental set up, thereby enabling the comparison of results. On the other hand, although user-oriented evaluation is highly beneficial, it is not only costly and complex but often difficult to replicate. It is the stability and standardization that makes the system-oriented evaluation so attractive.

To this end, the establishment of evaluation campaigns at both international and regional levels (e.g., TREC, CLEF, NTCIR, and FIRE) constitute a research activity that has been widely credited with contributing tremendously to the advancement of the field. Evaluation campaigns enable the reproducible and comparative evaluation of new approaches, algorithms, theories, and models, through the use of standardised resources and common evaluation methodologies within regular and systematic evaluation cycles. Motivated by the need to support users from a global community accessing the ever growing body of multilingual and multimodal information, the Cross-Language Evaluation Forum (CLEF) annual evaluation campaign, launched in 1997 as part of TREC, became an independent event in 2000 with the goal to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information [7]. The goal of CLEF is to provide an evaluation infrastructure and benchmarking facilities for the testing and tuning of monolingual and cross-language information retrieval systems operating on European languages. To this end, it provides an infrastructure for [7]:

- The comparative evaluation of multilingual and multimodal information access systems,
- The creation of reusable resources for such benchmarking purposes,
- The exploration of new evaluation methodologies and innovative ways of using experimental data, and
- The exchange of ideas experiences among IR researchers.

Since 2000, CLEF has played a leading role in stimulating research and innovation in a wide range of key areas in the domain of multimodal and multilingual information access. It has promoted the study and implementation of evaluation methodologies for diverse tasks, resulting in the creation of a broad, strong and multidisciplinary research community. Over the last ten years, CLEF has had a significant influence on the CLIR field by benchmarking various retrieval and annotation tasks and by making available very large-scale and standard test collections. While several CLIR research groups have participated over the years in the different evaluation tracks organized by the CLEF campaigns, even more IR researchers have acquired its datasets for experimentation.

4.4.2 CLIR Evaluation: Performance Measures and Criteria

According to [98], evaluation measures provide a way of quantifying retrieval effectiveness. Over the last few decades, different evaluation criteria and measures have been proposed to assess the performance of IR and CLIR systems. Most commonly used measures assume that every document is known to be either relevant or non-relevant to a particular query. The measures require a collection of documents and a query. To assess the performance of CLIR system, it is possible to study if the results returned by a certain query are related to it or not. This can be done by determining, given a query and a set of documents, the ones that are related (i.e., are relevant) and the ones that are not, and then comparing the number of relevant results returned by the retrieval system [99]. In this section, we focus on two set-based measures, namely, *precision* and *recall*.

As noted by [98], precision and recall have been the two simple but very popular IR performance measures for decades, (they were originally developed and used 1960s during the Cranfield IR evaluation experiments in the UK). Both are set-based measures: documents in the ranking are treated as unique and the ordering of results is ignored. While *precision* measures the fraction of retrieved documents that are relevant, *recall* measures the fraction of relevant documents that are retrieved. Precision and recall hold an approximate inverse relationship: higher precision is often coupled with lower recall. However, this is not always the case as it has been shown that precision is affected by the retrieval of non-relevant documents; recall is not [98]. Compared to other evaluation measures, precision is simple to compute because one only considers the set of retrieved documents (as long as relevance can be judged). However, to compute recall requires comparing the set of retrieved documents with the entire collection, which is impossible in many cases (e.g., for Web search). In this situation techniques, such as pooling, are used.

To formalize the notion of *precision* and *recall* and the relationship between them:

- Let *Rel* be the set of documents that are relevant, given a query *qry* in a reference collection.
- Also, let *Ret* be the set of documents retrieved by the retrieval system, when submitting the query *qry*, and,
- Let *RetRel* be the set of documents retrieved that were relevant (i.e., were in the set *Ret*).

Then, as shown in Figure 4.2, we can define ***recall*** as: *the ratio between the relevant retrieved documents, and the set of relevant documents.*

Similarly, ***precision*** can be defined recall as *the ratio between the relevant retrieved documents and the set of retrieved documents* (see Figure 4.2).

$$\boxed{\text{Recall} = \frac{RetRel}{Rel}}$$

$$\boxed{\text{Precision} = \frac{RetRel}{Ret}}$$

Figure 4.2 Formal definition of precision and recall

As noted by [98], other commonly used measures are based on evaluating ranked retrieval results, where importance is placed, not only on obtaining the maximum number of relevant documents, but also for returning relevant documents higher in the ranked list. A common way to evaluate ranked outputs is to compute precision at various levels of recall (e.g., 0.0, 0.1, 0.2, ... 1.0), or at the rank positions of all the relevant documents and the scores averaged (referred to as average precision). For example, to analyse graphically the quality of a retrieval system, we can plot the average precision-recall and observe the behaviour of the precision and recall of a system. These type of plots is useful also for comparing the retrieval of different CLIR systems [99]. This can be computed across multiple queries by taking the arithmetic mean of average precision values for individual topics. This single-figure measure of precision across relevant documents and multiple queries is referred to as mean average precision (or MAP). Another common measure is to calculate precision at certain document cut-offs, (i.e. at a given cut-off rank), considering only the topmost results returned by the IR system. This measure is called precision at n or P@n. For example, analysing the precision at the first 5 or 10 documents. Because the user is frequently presented with only the first ***n*** top documents retrieved, and not with the whole list of results, evaluating a CLIR system using this measure is very important. Basically, it represents the quality of the search results.

As mentioned before, to calculate precision and recall, it is necessary to analyse the entire document collection, and for each query determine the documents that are relevant. This judgment of whether a document is relevant or not, must be done by an expert or subject specialist in the field that can understand the need represented by the query. In some cases, this analysis is not feasible since the document collection is too large (for example, the whole Web) or maybe the user intention behind the query is not clear [99]. To address this problem, besides the document collection, the CLEF campaign organisers provide a set of queries (topics) and the corresponding set of relevant documents (relevance judgment). Using the document collection provided for the Ad-hoc track, they have defined a set of topics, with a description of the intention behind it, which can be used to query the CLIR system during the retrieval experiment and then compare the results obtained with the list of relevant documents.

5 Components and Architecture of OMEN-CLIR

As described in section 1.2 and section 4.1, although CLIR shares a number of IR features and techniques, it differs from it in many ways. Unlike in IR, the search queries and the documents do not share the same language in CLIR. Most of the earlier studies on the development of CLIR have assumed the availability of some reasonable linguistic resources such as MT system, multilingual dictionaries and parallel corpora. But, such core computational resources are either not available at all or available in a very limited quantity and poor quality for resource-scarce languages. This makes the task building CLIR for resource-scarce African languages exceedingly difficult and prohibitively expensive to perform. As noted by [11, 55], the task of developing CLIA for indigenous African languages has been severely constrained by the lack of sufficient linguistic resources and translation tools. Indeed, one of the essential prerequisites for development of CLIR is availability of adequate computational linguistic resources and translation tools related to the language pair being considered. In particular, the availability of basic language resources and translation tools such as bilingual dictionary, part-of-speech tagger, stemmer, and text corpora is very important for development of CLIR.

Hence, there is a compelling need for exploring and identifying basic linguistic resources and translation tools [55]. In an endeavour to explore and identify potentially useful translation tools and linguistic resources, we observed that, for a number of major African languages including Afaan Oromo, it is not difficult to find printed bilingual dictionaries or hard copies of published materials. We have been interested to study how such very limited translation resources and tools could be digitized and adopted for developing CLIR for African languages. To this end, we have proposed a dictionary-based CLIR that involves one of the severely under-resourced African languages (i.e., Afaan Oromo). In this chapter, we will describe the major architecture and components of OMEN-CLIR, which is designed and implemented without relying on very rich linguistic resources and translation tools. The general architecture of OMEN-CLIR is illustrated in figure 5.1.

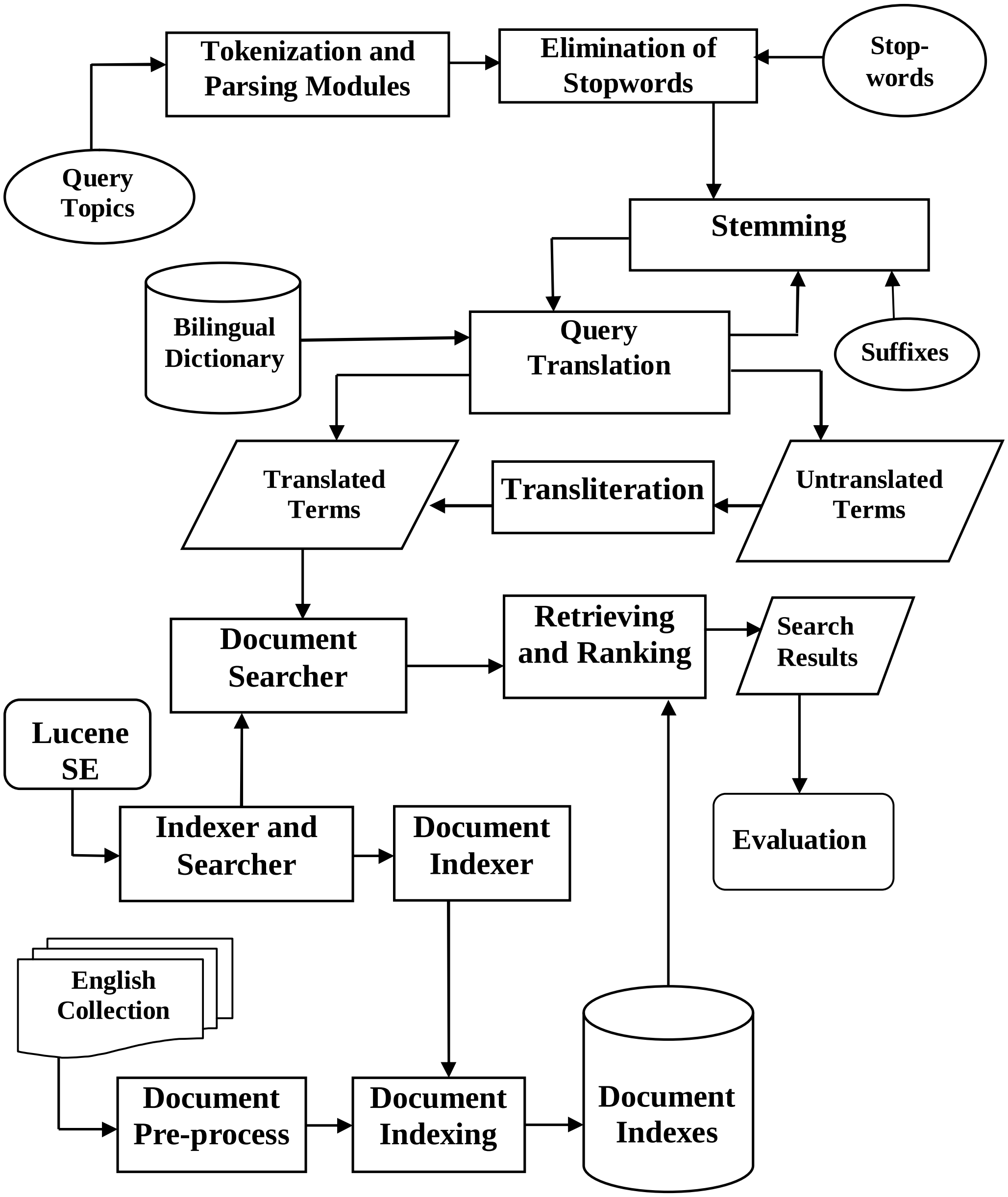

Figure 5.1 Overview of OMEN-CLIR Architecture

5.1 Construction of Oromo-English Bilingual Dictionary

As described in section 4.3, one of the most challenging problems that must be addressed in CLIR is language barriers, the linguistic disparity between the query language and documents language.

The main objective of CLIR is to find efficient translation resources and strategies that can be adopted to ensure that users can search and retrieve valuable information regardless of the language in which it is written or represented in. The availability of reliable translation resources such as multilingual and bilingual dictionaries is considered as an essential prerequisite for successful development of CLIR [95]. A bilingual dictionary provides list of vocabularies in a source language along with appropriate definitions and translations in a target language. Optionally, a bilingual dictionary may include translation probabilities assigned to the definition of each lexical entry in order to facilitate word sense disambiguation and weighting. Although a corpus-based query translation approach has recently received much attention in developing CLIR for well-resourced languages, a dictionary-based query translation technique remains one of the most popular approaches in developing CLIR for less-resourced languages [19, 18, 11].

As noted by [6, 17], bilingual dictionaries are crucial components of CLIR and MT systems as well as other multilingual natural language processing applications. As indicated in section 4.3.4, a number of natural language processing applications, including CLIR, require multilingual lexicons that provide detailed information about lexical, syntactic and semantic properties of words in different languages. In dictionary-based query translation approach, the performance of the CLIR largely depends on the size and quality of the translation lexicon.

One of the major challenges in developing CLIR for severely under-resourced African languages is unavailability of online lexical resources and translation tools. Due to the lack of suitable linguistic resources related to African languages, basic translation tools necessary for the development of CLIR need to be designed and developed from scratch. Consequently, the task of building a CLIR for under-resourced language is exceedingly difficult and time-consuming. Usually, published or printed bilingual dictionaries are the only lexical resources that can be easily obtained for severely under-resourced languages like Afaan Oromo. Although a human-readable printed bilingual dictionary can be scanned and converted to digital form, it needs a lot of editing, proofing and formatting. Since lexicographers often use their own individual style, there are many inconsistencies and differences in the way they treat and organize dictionary entries. Moreover, the scanned dictionary must be cleaned and formatted before it can be adopted for the development of CLIR.

In this study, we have focused on building and adopting a bilingual dictionary. As described in our earlier papers [25, 13, 26], we have constructed and adopted a machine-readable bilingual dictionary from a printed copy of Oromo-English dictionary [92]. However, as noted by [83], since printed bilingual dictionaries are typically designed for human users, their translations are often augmented with various examples and illustrations. Since such extraneous information provided for human users tends to confuse automated translation systems, human readable bilingual dictionaries must go through a series of pre-processes stages, including editing and formatting of the lexical entries. After scanning and digitizing the bilingual dictionary by using OCR technology, we have devoted a lot of efforts in editing and formatting the lexical entries. Besides removing unnecessary and inconsistent descriptions provided for human users, we have corrected various spelling errors and typos that have occurred during the scanning and conversion processes. Moreover, we have enhanced the coverage of the bilingual dictionary by incorporating additional vocabularies and technical terms from other lexical resources and glossaries. Currently, our machine-readable Oromo-English dictionary contains about 12,000 lexical entries. Although this general purpose and medium-size dictionary cannot be considered as a comprehensive and complete lexical database, it has a good broad coverage that makes it very useful for the purpose of query translation.

5.2 Construction of Afaan Oromo Stopwords

Stopwords can be defined as a set of commonly used function words or non-content-bearing terms such as postpositions, prepositions, pronouns and conjunction, which do not contribute to the content of textual documents [100, 101]. It refers to a list of very frequently used function words that carry information of minor relevance to the content or subject matter of documents. Due to their high frequency of occurrence, the presence of function words in the index of documents presents an obstacle to the identification and retrieval of relevant information. Hence, most IR systems, including CLIR, make use of stop words list to identify and eliminate non-content-bearing and less important words. While a few examples of the frequently used English stop-words include "a", "of", "the", "it", “in”, "you", and "and”, a few examples of Afaan Oromo stopwords include *“bira”, “fi”, “irra”, “of”, “irraa”, “itti”, “male”* and *“yookin”*. One of the most common characteristics of these stopwords is that they carry very little information about the contents of a document.

Since stopwords are mainly used for syntactic and grammatical purposes, they are frequently occurring throughout the corpus of the document collection. Moreover, stop-words are a finite set in a given language, i.e. new function words emerge less often than content words. In other words, the percentage of the function words in a given language is stable in a text sample as the size of the sample increases [100]. However, they are too common to be used as discriminator in searching and indexing documents. A search conducted by using stopwords like "*the*" "*of*" or "*and*" in English may return too many items, which are irrelevant to the user's query. Hence, it is useful to identify and remove function words. The assumption is that, when the contents of the document is analysed and determined, more relevant index terms or key words can be obtained by eliminating or ignoring common words that are frequently occurring throughout the text corpus [100].

There are two major techniques widely used to construct list of stop-words: statistical and linguistic [100, 101].

- **Statistical Approach:** The statistical technique is based on frequency count of words in a large text corpus. The basic idea is that function words are more frequent than content words. After the frequencies of all words found in a large text corpus are computed and ranked, the top most frequent *n* words could be identified and incorporated into stop-list. A threshold can be set, based on size of the text corpus and morphological features of the language being involved, to determine the number of words to be included into the stop-list. In general, words most frequently occurring in a text corpus are included into a stop-list and treated as stopwords. For instance, a word that frequently occurs in more than 80% of documents in the corpus is considered as a less important index term since it cannot help in discriminating between items. On the other hand, words with a very low inverse document frequency weight could also be considered as stop-words in some cases.

- **Linguistic Approach:** Stop-words can be identified and built by using certain linguistic features and morphological characteristics of language. For instance, words belonging to predetermined syntactical categories, such as determiners, conjunctions and prepositions can be easily identified and incorporated into sop-lists based on their part-of-speech tag. Inversely, it is also possible to choose index terms from words which belong to specific syntactic classes such as nouns and verbs. If a reliable and comprehensive bilingual dictionary is available, it is also possible to build a stopword list by translating existing list of stopwords in one language into another language(s).

As indicated in [61, 100], the performances of retrieval systems can be improved by removing or ignoring stopwords. Particularly, if done early in the indexing process, elimination of stop-words can make further processing of the candidate index terms more efficient and reduce the storage space. To this end, various stopword lists have been established for a number of European and Asian languages including English and French. Unfortunately, a standard list of stop-words is not readily available for most of indigenous African languages, including Afaan Oromo. Hence, it becomes necessary to identify and construct a list of stopwords for Afaan Oromo. In this study, we have identified and established a list of Afaan Oromo stopwords that can be used for eliminating function words and non-content bearing terms.

In order to establish the list of Afaan Oromo stopwords, we have adopted a combination of both statistical and linguistic approaches. Initially, we have created a list of about 350 words that were frequently occurred in a small-size Afaan Oromo text corpus. Then we have manually verified the list to remove some of the words that are considered as important index terms. Some function words are also incorporated into the list by consulting Afaan Oromo textbooks and reference materials. In addition, we have also translated relevant function words found in English stopwords and included them into Afaan Oromo stopwords. Accordingly, we have established and employed about 580 Afaan Oromo stopwords to eliminate non-content bearing words from Oromo queries. This stopword list can be easily adopted and used as an effective device to filter out function words. It can also serve as the basis for establishment of more complete and comprehensive Afaan Oromo stopwords. More detailed description about the impact of eliminating Afaan Oromo stopwords is given in section 5.4.

5.3 Construction of Afaan Oromo Stemmer

5.3.1 Rule-Based Stemmer: Definition and Justification

At the first glance, it seems easy to identify and store all possible inflected words in natural language in a dictionary along with their lemmas (or root-words). In that case, it is possible to search the dictionary database in order to find and interpret the word form variants without applying and using any morphological processing tools such as stemmer and lemmatizer.

Although this approach can be feasible for languages that are morphologically simple or analytic, it is not viable and sustainable approach for morphologically very rich and agglutinative languages like Afaan Oromo [23, 102]. In morphologically very rich and complex African languages, a single lemma or root-word can generate hundreds (or even thousands) of word form variants [103, 102]. Hence, it is crucial to devise an efficient morphological processing tool that can identify the stem (or base-word) from which word form variants are derived. In order to obtain the stem of inflected words, morphological affixes must be identified and analysed by appropriate morphological processing tools such as stemmer, lemmatizer and morphological analyser.

Stemming is the process of identifying and normalizing word form variants to their common stem or base form by removing possible morphological affixes. It is the process of verifying and reducing word form variant to a common stem by stripping off derivational and inflectional affixes [97, 68]. Normally, a stemmer takes an inflected word form as input, removes possible morphological affixes and returns the appropriate stem or base-form as an output. A stem can be defined as the basic lexical unit or base-form of a word that can add bound morphemes to make new words through the processes of derivational and inflectional morphology [56]. Unlike grammatical affixes or bound morphemes, which are mainly used for syntactic functions, the stem is the basic part of a word that conveys a particular concept, thought, or meaning.

As noted by [68, 56], in most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. Hence, documents are often indexed and represented by the stem or lemmas rather than by the inflected word forms. Stemming is not only important to reduce different word variants to a common base or citation form, but also to reduce the size of the index, that is, the number of distinct terms needed for representing a set of documents. A smaller index size results in a saving of storage space and processing time [97]. Consequently, the task of building an algorithmic method for conflation of different inflected word forms is an important component of many IR and CLIR systems. It helps to improve the system's recall and can significantly reduce the index size. This is especially true for highly-inflectional languages like Afaan Oromo. According to [56], the benefits of stemming are two-fold:

- By conflating several forms to the same stem, the number of entries and the size of the search index are reduced.

- More importantly, stemming is very helpful for free text retrieval, where search terms can occur in various different forms in the document collection. The application of a stemmer can help the user by making the retrieval of documents independent from the specific word forms that are used in a search query. In particular, stemming is very essential for highly inflectional and agglutinative languages like Afaan Oromo.

In dictionary-based CLIR, where matching and translation operations are required between dictionary entries and query terms expressed in different languages, the role of stemming and normalizing word variants is very crucial. However, the output of the stemmer (i.e., stem) may not be a root-word or a valid lexical item. As noted by [103], the output of the stemmmer need not be identical to the morphological root of the given word. In IR or CLIR, it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root-word. Since morphological processing in IR does not demand detailed morphological information, identifying and recovering a common base-form (stem) of an inflected word is quite sufficient. In other words, although the output of a stemmer may not be a linguistically valid root-word, it is sufficient for the purpose of IR that related words map to the same stem or base form.

For example, in the context of IR and CLIR, it doesn't matter whether the stems generated are genuine words or not. Thus, Oromo plural noun "*adurroota*" (cats) might be stemmed and reduced to the stem "*adurr*" instead of the lemma "*adurree*" (cat), provided that different words with the same meaning are not conflated to the same base-form, and words with distinct meanings are kept separate [56]. Accordingly, a stemmer can remove morphological affixes and convert the inflected nouns and verbs as follows:

- A plural noun into a singular noun (e.g. *adurr**oota*** → *adurree* or *adurr,* (i.e., cat).
- A plural nominative noun into a singular noun (e.g. *adurr**ootiin*** → *adurree* or *adurr,* (i.e., cat).
- A plural dative noun into a singular noun (e.g. *adurr**ootaf*** → *adurree* or *adurr,* (i.e., cat).
- A conjugated verb into infinitive or verb-stem (e.g. *dhufera* → *dhufuu* or *dhuf* (i.e., come).

An algorithm which attempts to convert an inflected word to its linguistically correct root ("*adurree*" in the above example) is called a lemmatizer.

While a number of stemming algorithms have been developed and widely used for many European and Asian languages including English [68, 56], very little attention has been given to the development and application stemmers for indigenous African language. Although most of the indigenous African languages, including Afaan Oromo, are morphologically very rich and complex, they do not have reliable morphological processing tools such as stemmer, part-of-speech tagger and morphological analyser. A number of earlier studies have confirmed significant impacts of stemming in improving the performances of CLIR [103, 61].

Most of the existing stemming algorithms depend on morphological or grammar rules to identify and normalize inflectional and derivational affixes. Rule-based stemming algorithms have been widely used in various natural language processing applications including CLIR and MT systems. While rule-based approaches typically involve the use of grammatical rules to design and develop morphological processing tools, data-driven approaches use statistical techniques to model and develop morphological processing tools such as stemmer and morphological analyser. Unfortunately, statistical approaches require very large linguistic data or text corpora. Due to the lack reliable linguistic data for Afaan Oromo, we are focused on a rule-based approach. Although it is difficult and time consuming, a rule-based approach has the advantage of providing an insight into the kinds of linguistic phenomena encounter in Afaan Oromo.

Broadly, rule-based stemmers can be grouped into two classes: *root driven* and *affix stripping*. In the first approach, the stem of the word is found initially and then the affixes are determined. In the second approach, which used in our Afaan Oromo stemmer, the determination of the suffixes takes place first in contrast to the first one. After the removal of all possible suffixes, the remaining part of the word can be assumed to be a stem, or a lexicon database can be used to approve such assumption. In general, in contrast to morphological generator that uses a root or stem driven approach, a stemmer can be assumed as a procedure that verifies and removes morphological affixes by using the morphotactics and morphophonemic rules in a reverse order. A stemmer initially accepts the surface form of a word along with potential suffixes. Then, it checks required minimum length and identifies any possible bound morpheme or affix in the given word. Based on the morphotactics of the language, it removes possible inflectional affixes in each of its subsequent steps. At the end of the stemming process, a stem (or base-form) of the inflected word is expected to be identified and returned as an output of the stemmer.

5.3.2 Overview of Afaan Oromo Stemmer

Over the last few decades, a considerable amount of research effort has been put into the development of morphological processing tools such as stemmer, lemmatizer and morphological analysers for a number of well-resourced European and Asian languages. On the contrary, there has been very little work done on the development of computational models and algorithmic descriptions for resource-scarce African languages. Although certain morphological processing tools and algorithms have been recently developed for some of the African languages, they are not yet employed for natural language processing applications and CLIA systems like CLIR. As indicated in section 3.2, because of its complex morphological structures, Afaan Oromo poses a huge challenge for text processing and information retrieval tasks like CLIR. Afaan Oromo words can occur in base-form (or citation form), inflected form, derived form and reduplicated form as well as in a complex compound word forms. Oromo nouns and verbs can take multiple inflectional affixes, stacking one after the other. For example, a few of word form variants of the Oromo noun “*mana*” (i.e., house) include:

- “*man**oota**→ (mana +oota”),*
- “*man**ootatti** → (mana + oota + itti)*”,
- “*man**ichatti** → (mana + icha + itti)*”,
- “*mana**wwan** → (mana + wwan)*”,
- “*man**arratti** → (mana + irra + itti)*”,
- “*mana**tti** → (mana + itti)*”.
- “*man**ootawwanitti** → (mana +oota + wwan + itti”),*

As discussed in section 3.2 and 3.3, inflected words are formed by the affixation of grammatical features such a case, number, tense, aspect, mood etc. to the root-word or stem. Removing morphological affixes from an inflected word is very important in a dictionary-based query translation because it normalizes word form variants and provides the base-word or stem that could be found in a bilingual lexicon. Hence, the task of building and applying stemmer for Afaan Oromo has become one of the major components of our study. In this section we will briefly describe the basic features of two closely related Afaan Oromo stemmers, which are designed to deal with common inflectional affixes at different level of morphological layers. As indicated by [91], since prefixes are not very common in Afaan Oromo, we have mainly focused on identifying and removing inflectional suffixes and postpositions, which are frequently used in Afaan Oromo texts. No attempt has been made to identify and normalize Afaan Oromo prefixes in this study. Afaan Oromo prefixes often tend to be less frequent and derivational in nature. Unlike inflectional affixes, which are primarily used for grammatical functions, most Afaan Oromo prefixes are considered as lexically and semantically significant. In other words, since most of Afaan Oromo prefixes tend to change the lexical category and meaning of the word that they are attached to, removing them can affect the performance of a CLIR negatively. Consequently, we have focused on designing and building a rule-based Afaan Oromo stemmer that identifies and removes inflectional suffixes.

Towards this end, we have identified frequently used Afaan Oromo inflectional suffixes and constructed a suffix dictionary. The suffix dictionary along with Afaan Oromo morphophonemic rules and morphotactics has been adopted and used for developing two closely related Afaan Oromo stemmers: a rule-based *crude-stemmer* and *light-stemmer* (see Figure 5.4 and Figure 5.6). As discussed in detail in section 3.3, the major types of morphological suffixes that are commonly used in Afaan Oromo can broadly be classified into three categories, namely:

- *Derivational* or *constructive* suffixes,
- *Inflectional* suffixes and,
- *Enclitics* or *attached postpositions.*

In other words, based on morphotactics of Afaan Oromo it is possible to classify and deal with Afaan Oromo suffixes at three different morphological layers, with respect to: *derivational, inflectional,* and *enclitics or attached postpositions* [91, 62] (see Figure 5.2).

- **Enclitics or attached postpositions:** Afaan Oromo enclitics include various particles, postpositions or clitics that are appended or attached to a root-word or stem. Oromo enclitics include auxiliaries, pronouns and particles like "*irra*", "*irraa*", "*fi*", "*itti*" and "*dhaa*".

- **Inflectional suffixes:** As described in detail in section 3.2 and section 3.3, Afaan Oromo has several inflectional suffixes that are commonly used to indicate grammatical categories such as number, gender and case (e.g. "*-n*", "*-lee*", "*-een*", "*-icha*", "*-tu*", "*-oo*", "*-oota*" and "*–wwan*"). Inflectional suffixes are one of the most frequently used and morphologically productive affixes in Afaan Oromo.

- **Derivational suffixes:** Afaan Oromo has many derivational suffixes which include bound morphemes such as "*-achuu*", "*-eessuu*", "*-eenyaa*", "*-ina*" and "*-ummaa*".

These derivational suffixes are often used in formation of new words from the existing main class words.

Based on Afaan Oromo morphotactics described in detail in section 3.3 and additional observation of the syntactic structures of Afaan Oromo, the most common order of the above three types of morphological suffixes is given in Figure 5.2.

STEM* + [*derivational suffixes*] + [*inflectional suffixes*] + [*Enclitics or Postpositions]

Figure 5.2 Major morphological layers of Afaan Oromo suffixes

Based on this general pattern and order of Afaan Oromo suffixes, we have designed and implemented the two closely related rule-based stemmers, namely: crude stemmer and light stemmer. In the next two subsections we briefly describe the major procedures of these linguistically motivated Afaan Oromo stemming algorithms.

5.3.3 Overview of Afaan Oromo Crude Stemmer

As described in section 3.2, Afaan Oromo is a highly inflective and agglutinative language. An Oromo noun declines to mark number, definiteness and grammatical case. Similarly, an Oromo adjective must agree with the head-noun it modifies with respect to gender, number, definiteness and grammatical case. Consequently, the surface forms of Afaan Oromo nouns and adjectives greatly vary. As a first step towards handling this very frequent and common problem of word form variations, we have designed and implemented a crude stemmer that mainly focuses on identifying and removing the longest possible inflectional suffixes from Oromo nouns and adjectives in a single pass, (see Figure 5.3). As illustrated in Figure 5.3, the main focus of the crude stemmer is to remove the longest possible inflectional suffixes or concatenated morphemes, such as plural and case markers or definiteness and case markers (bound morphemes) of Afaan Oromo nouns and adjectives in a single pass. The insight behind this initial approach is to investigate and determine the extent to which our OMEN-CLIR performance can be improved by designing and implementing a simple stemmer that focuses on removing the most frequently used inflectional suffixes.

The suffix stripping procedure implemented in the crude stemmer first checks if the minimum length of the given stem is not less than three characters before removing any candidate inflectional morpheme associated with it in order to minimize or avoid over-stemming errors. If the candidate term is longer than 3 characters, the longest possible inflectional suffix is always removed before the shorter inflectional suffix. Some of most commonly used inflectional suffixes that are identified and removed by this crude stemmer include concatenated plural and case markers or concatenated definiteness and case markers. For example, suppose the input of the crude stemmer is the inflected word: "*fard**ooleen***" → *(fard + oolee + n)*, (i.e., "horses"), the longest possible inflectional suffix, which is "***-ooleen***" in this example, is removed and the remaining stem, i.e., "*fard,*" is returned by the stemmer. To be more specific, the task of removing the plural markers "*-oolee*" as well as the case marker "*n*" is executed in a single pass. The removal of the inflectional affixes is performed starting from the right end of the given word, by taking into consideration the orthographic and morphotactics of Afaan Oromo, which is described in detail in section 3.3.

5.3.4 Overview of Afaan Oromo Light Stemmer

Emphasizing the advantage of a light stemmer over morphological analyser, [103] stated that detailed linguistic analysis or full solution to the problem of morphology is not required for the purpose of most IR systems. Taking this fact into consideration, the main objective of our Oromo light stemmer is to identify and remove inflectional suffixes attached to a stem iteratively. Since a word can take multiple suffixes, the light stemmer should able to verify and remove agglutinated bound morphemes. As illustrated in Figure 5.4 and Figure 5.5, in contrast to the crude stemmer, our light stemmer iteratively removing different types of affixes at three different morphological layers (i.e. derivational, inflectional and postpositional suffixes). The insight behind this less aggressive and iterative stemming algorithm is to investigate and determine the extent to which the performance of OMEN-CLIR could be improved by classifying and removing different type suffixes step by step.

Based on the morphotactics of Afaan Oromo described earlier, the light stemmer start with removing longest possible *enclitics or attached postposition*, which is followed by the removal possible *inflectional suffixes*, if any (see figure 5.4 and figure 5.5). To facilitate the light stemming process, we have identified and constructed three different clusters or groups of suffix lists with respect to the three major types of morphological suffixes in Afaan Oromo. After the removal of enclitics or attached suffixes, each modified query term will go through the next stage of the light stemmer.

As indicated in Figure 5.4 and Figure 5.5, since most Afaan Oromo words can be associated with more than one suffixes, the light stemmer often needs to verify and remove stacked inflectional suffixes and postpositions. The three major stages of the light stemmer can be summarized as follows:

1. ***Removing enclitics or attached postpositions*****:** The first stage of the light stemmer deals with identifying and removing enclitics or attached postpositions. Initially, the stemmer computes and checks the minimum length of the input word or string. If the length of the string is greater than 3, the light stemmer tries to match the right end substrings of the word against the longest possible enclitics. It removes all possible attached postposition iteratively at this stage, if any. To illustrate this process with an example, suppose the input of the light stemmer is the word "*man**ootawwanirratti***", (which means "on the houses"). The light stemmer first starts with identifying and matching the longest possible attached postposition from the right end of the given string. Accordingly, it identifies and removes the attached postposition or enclitics *"-tti"* from the right end of the string and the modified and remaining part of the word becomes *"man**ootawwanirra.**"* The stemmer checks the length of the modified word and then tries to find additional attached postpositions. Accordingly, it identifies and removes the second attached postposition *"-irra"* and the modified and remaining part of the word becomes *"manootawwan."* If the remaining part of the word has any more attached postposition (which doesn't hold true here), the stemmer will iteratively removes all possible enclitics at this stage.

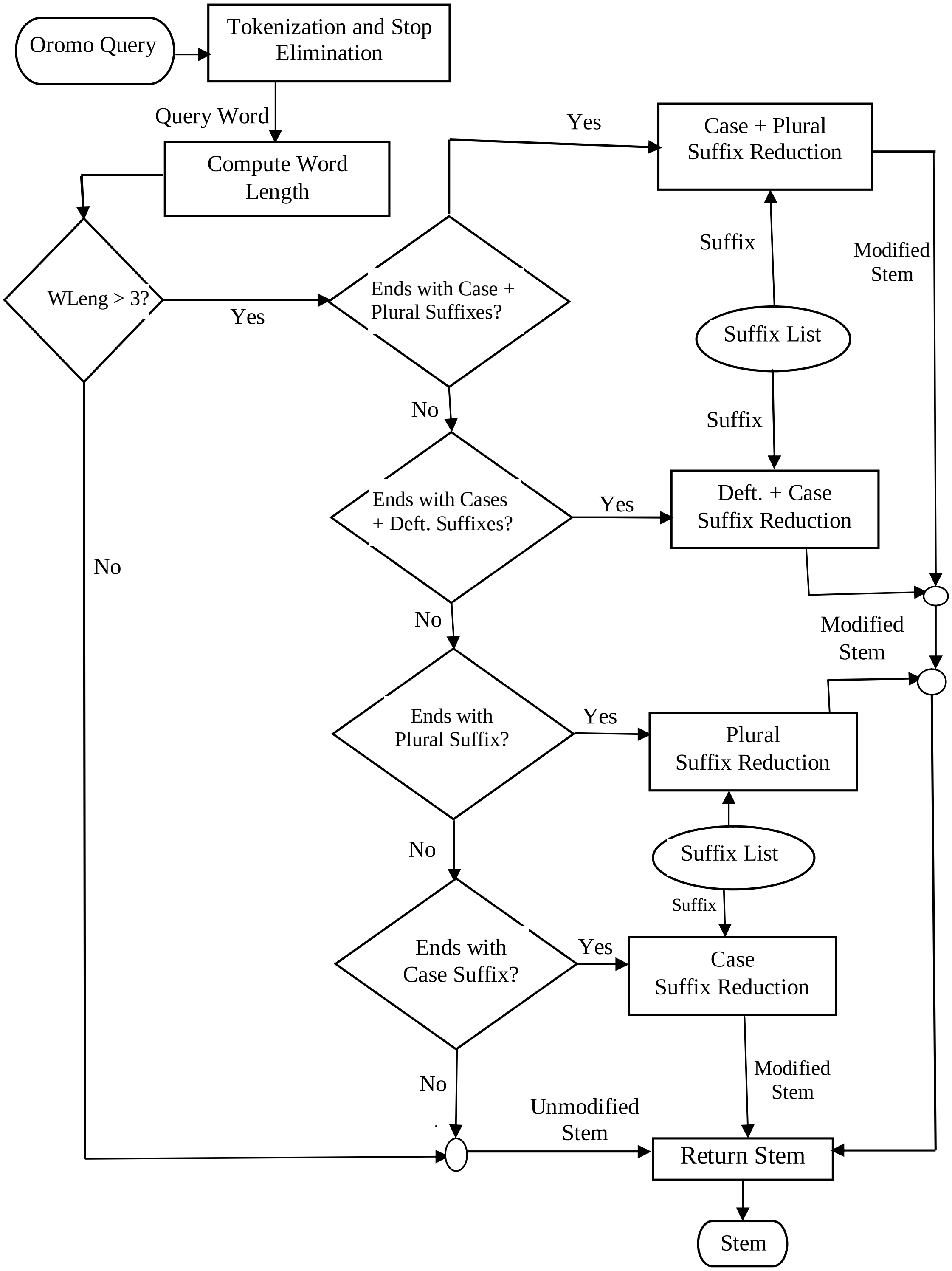

Figure 5.3 Flowchart of Afaan Oromo crude stemmer

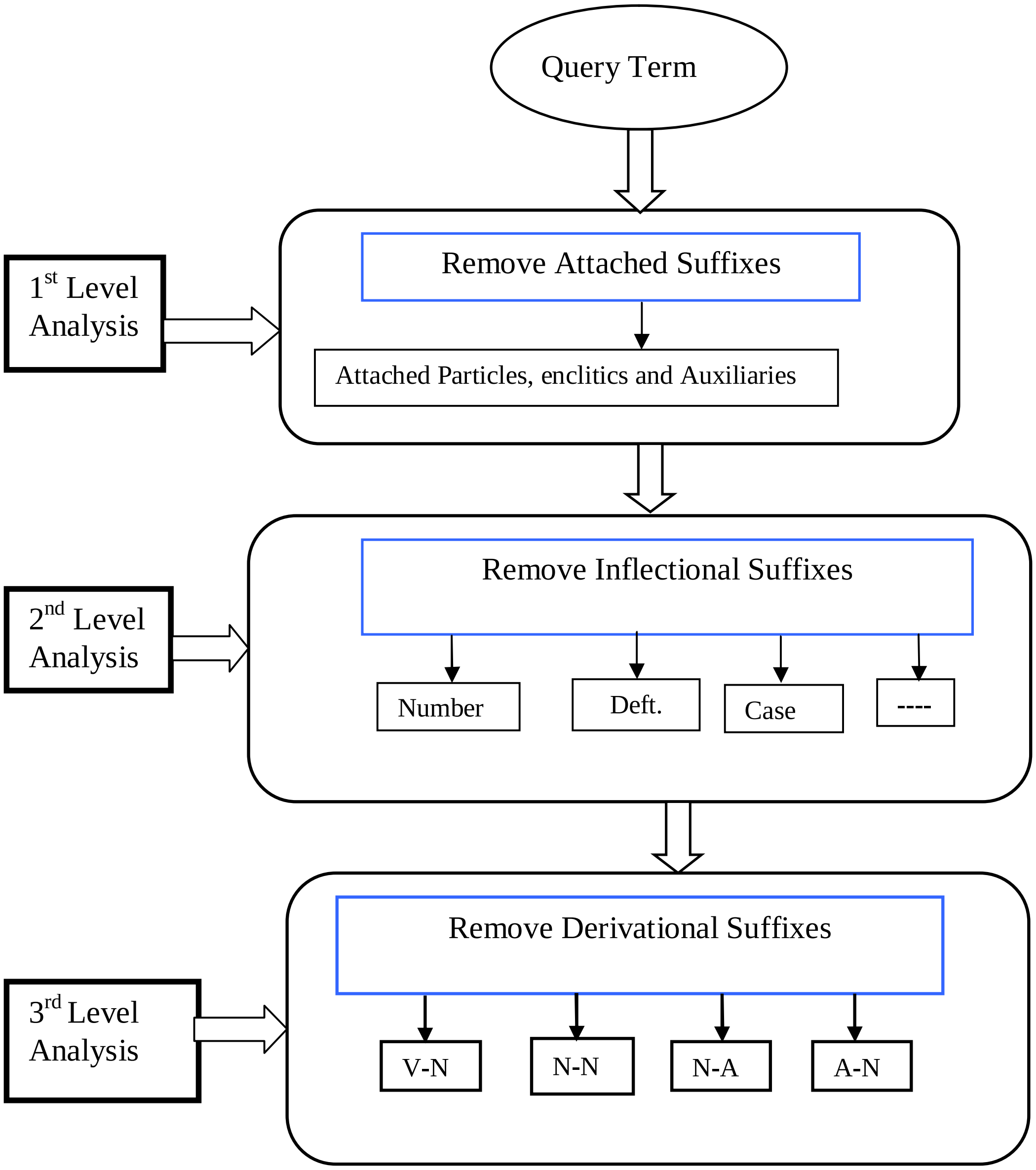

Figure 5.4 Major stages in Afaan Oromo light stemmer

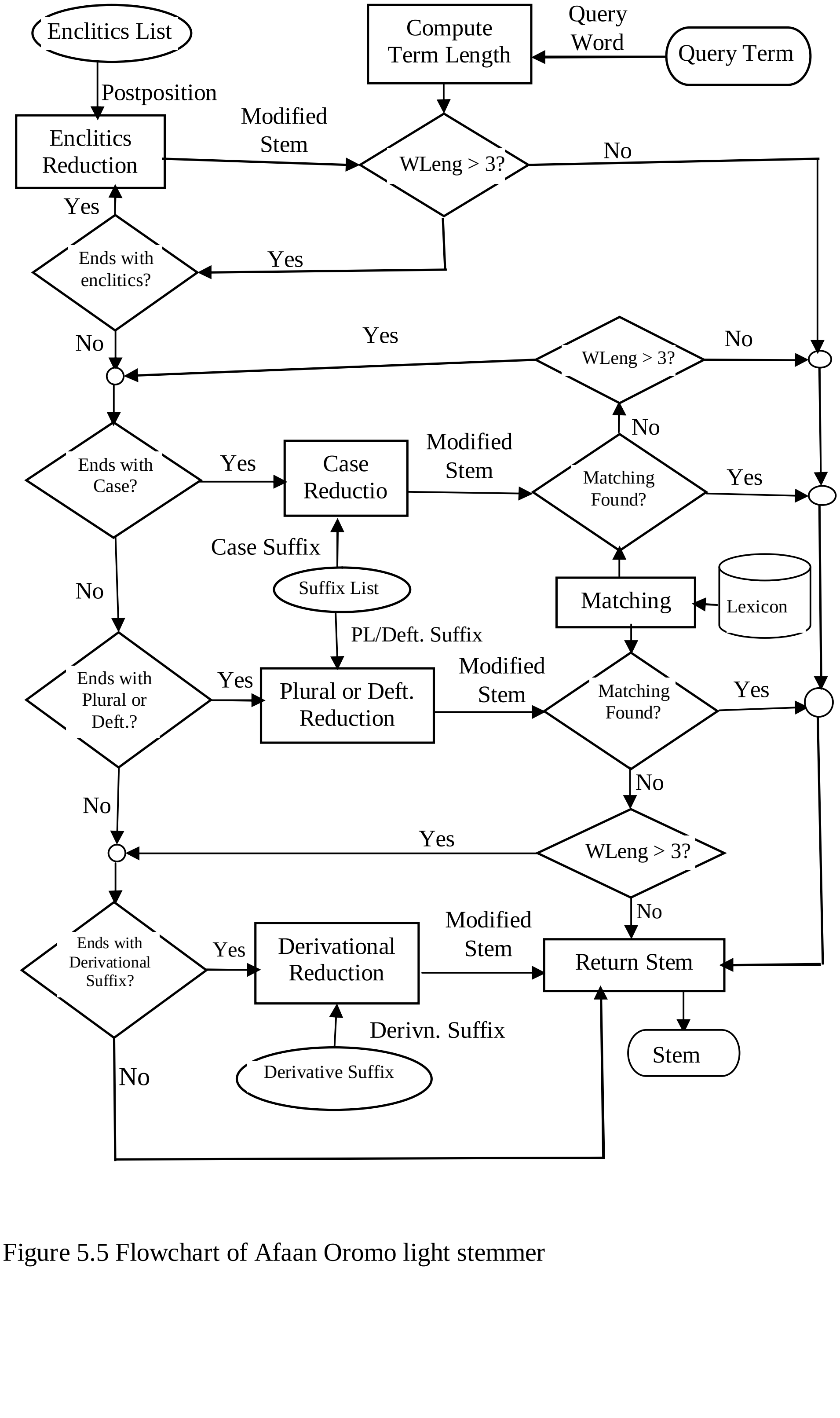

Figure 5.5 Flowchart of Afaan Oromo light stemmer

2. **Removing *inflectional suffixes*:** After removing possible attached postpositions or enclitics at first stage, Afaan Oromo *inflectional suffixes* are verified and removed by the light stemmer at the second stage. Accordingly, the light stemmer tries to find the longest possible inflectional suffix that can match with the right end of the query word, i.e., *"manootawwan."* The stemmer identifies the plural suffix*"-wwan"* and hence strip it off. Then, the light stemmer tries to identify and remove additional inflectional suffixes. Accordingly, it identifies and removes the plural suffix "*-oota"* and the modified word becomes *"man"* If the remaining stem is longer than three characters and has any more inflectional suffix, (which is not the case here), the light stemmer will continue to identify and remove possible inflectional suffixes step by step. After the removal of all possible inflectional suffixes, the identified stem (i.e., *"man"*) is normalized to "*mana*" and matched against the bilingual dictionary entries. If its translation is found in the dictionary, the corresponding English translation is identified and returned. Otherwise, the modified word may undergo the next final stage of the light stemming process, i.e. removing derivational suffixes.

3. **Removing *derivational suffixes*:** The task of removing *derivational suffixes* is applied at the final stage of the light stemmer. Unlike the above two major types of suffixes, the removal of derivational suffixes tends to change the lexical category of the stem. Hence, our light stemmer removes a derivational suffix only in case the candidate word is not found in the bilingual lexicon. If the modified word (stem) cannot be found in the dictionary after all of the stemming procedures have been applied, the candidate query term is marked and stored in a separate list for further analysis and transliteration process.

In summary, in contrast to morphological generator, which often uses a root driven approach, a light stemmer can be assumed as a procedure that verifies and removes morphological affixes in a reverse order. The surface form an input word is always verified and normalized from right to left, and the longest possible suffix is always removed before shorter bound morphemes. After a successful removal of inflectional suffixes, a dictionary lookup is performed with the remaining part of the stem in order to minimize over-stemming errors. At the end of the stemming process, an appropriate stem (or a base-form) of the morphological variants is expected to be returned as an output of the light stemmer.

5.3.5 A Preliminary Assessment of Afaan Oromo Light Stemmer

Assessing the performance a stemmer is important to determine the range of word form variants that it can handle properly. Since annotated text corpora and standard test datasets are not available for Afaan Oromo, we have conducted a preliminary evaluation of our light stemmer by using a small-size text data extracted from Afaan Oromo newspaper, (i.e., Barisa). The text was extracted from different portions of the newspaper including various domains such as business, economics, politics, entertainment and sports. A group of individuals who are familiar with Afaan Oromo morphology were given the outputs of light stemmer and instructed to classify the results into two categories: *correct stem* and *incorrect stem*. If the output of the stemmer was classified as incorrect, the evaluators were requested to specify whether the error is over-stemming or under-stemming. The preliminary evaluation results of our light-stemmer are summarized and presented in Table 5.1.

As described in section 5.3.1, it is important to remember that the output of a stemmer need not be a proper lexical item or a valid word in order to be treated as correct stem. The basic requirement is words mapped by the stemmer to a given stem should be genuine morphological variants that stand for the same basic concept. Accordingly, our light-stemmer is judged to be correct if it maps all possible inflected and derived forms of a word to a single stem after removing attached morphological suffixes. However [56], there are two major kinds of errors that often occur during the process of stemming.

1. **Under-stemming**: If the stemmer fails to map all genuine morphological variants, (including inflected and derived forms), to a single stem, then the phenomenon is called *under-stemming*.
2. **Over-stemming**: If a stemmer maps different words to a single stem that are not genuinely morphological variants or if it removes non-affix endings (or substrings) of a valid lexical item, then the phenomenon is called *over-stemming*.

More specifically, while the term *over-stemming* refers to incorrect stem that results from the removal of non-affix substrings of a word, *under-stemming* refers to incorrect stem that occurs when the stemmer fails to remove all possible morphological affixes or when it stops before reaching the appropriate stem or base-form. An example of under-stemming in Afaan Oromo is a stemmer conflating “*fardooleewwan*” to “*fardoolee*” and “*fardootan*” to “*fardoo*” instead of mapping both to “*farda*” (which means “horse” in English). An instance of over-stemming occur in Afaan Oromo if a stemmer removes *“–ttin*” from both “*mana**ttin***” and “*allatti**n***” instead of conflating “*mana**ttin***” to “*mana*” (house) and “*allatti**n***” to “*allatti*” (bird).

Number of Words	Correct Stem	Stemming Errors		Accuracy
		Over- stemming	Under-stemming	
2255	1879	254 (11.3%)	122 (5.4%)	83.3%

Table 5.1 Summary of Afaan Oromo light-stemmer performance

As shown in Table 5.1, we have assessed the performance of our light stemmer against a small-size text corpus containing about 2255 words. While about 254 (11.3%) words were found to be over-stemmed, about 122 (5.4%) words were found to be under-stemmed. Accordingly, the accuracy of our stemmer is found to be about 83.3%, which can be considered as a promising result given the morphological complexity of Afaan Oromo. Most of the errors occur because there are some exception words in Afaan Oromo that need special treatment as well as because of complex morphological properties of the language. As described in section 3.2 and section 5.3.4, Afaan Oromo has a very rich and complex morphological system where most words can be attached with more than one bound morphemes or affixes. Although most common inflectional and derivational suffixes were properly identified and removed by our light stemmer, we have observed certain under-stemmed and over-stemmed problems during our error analysis (see section 6.3). Currently, we are trying to address some of these problems by incorporating additional rules and list exceptional words into the modules of the light stemmer.

5.4 Query Processing and Translation

As indicated earlier, a query is a short statement that is often expressed in a natural language as a representation of the user’s information needs. A document is a source of data or information that is analysed and represented by the IR system in its index database. In a classical IR system, since both search queries and target documents are represented in the same language, they can be directly matched against one another in order to retrieve relevant documents. In CLIR, unlike monolingual IR, since search queries and documents do not share the same language, the task of bridging the language gap between the search query and the target document collection is one of the major challenges that should be addressed. A search request or query expressed in a source language must be pre-processed and translated into the target language before it can be used for searching and retrieving relevant documents. In this study, the source language is Afaan Oromo, the language in which search quires (topics) are represented. The target language is English, the language in which the documents are written in.

Query Topics Feature	CLEF-2007 Oromo Topics	CLEF-2006 Oromo Topics
Number of queries	50	50
Number of query tokens	2378	2355
Average query length	47.56	47.10

Table 5.2 Statistics of Oromo query datasets

```
</top>
<top lang="OM">
<num>425</num>
<title> Sanyiiwwan Baduutti Argaman</title>
<desc>Dokumantoota waa'ee ajjeechaa, ajjeesaanii gurguruu ykn  seeraan ala
qurxummii kiyyeessuu sanyiiwwan badiitti jiranii ykn akka egaman godhaman
kamiyyu (harma hoosistoota, allaatti ykn qurxummii) haasawwan
barbaadi.</desc>
<narr>Dokumantoonni dhimma kanaan walitti dhufeenya qaban sanyiiwwan
baduutti jiran irraa gochawwan seeraan alaan  balaa gahaa jiru ibsu qabu.
Dokumantoonni bineensoota baduutti hin jirre faarsan fudhatama hin
qaban.</narr>
</top>
```

Figure 5.6 Example of Oromo query topic

As shown in Table 5.1 and Figure 5.6, in the ad-hoc tracks of the CLEF campaign, the test data often contains a set of query topics that represents a typical user’s information needs [98]. These search topics could be expressed as short queries, questions or relatively longer descriptive statements. As shown in Figure 5.6, the ad-hoc tracks of the CLEF campaign uses the notion of a topic, which typically consists of three fields, namely: *title, description and narration.* These three related fields can be used as representation of an alternative or complementary search request that could be issued or asked by a user. A *title* is a phrase or a very short statement that indicates the basic subject matter of desired documents. A *description* is a sentence that provides a brief explanation or definition about the search topic. A *narration* field provides more elaborative statements or descriptions about the search topic. Apart from covering a number of important national and global issues, the query topics used in the ad-hoc tracks of the CLEF campaign are created by taking into account a wide range of IR and CLIR issues that could be encountered in the search requests of users in a real-world setting, including lexical ambiguities, named entities, multiword expressions and complex phrasal terms. The original English search topics were translated into Oromo query topics by native speakers Afaan Oromo who are fluent in English, but with no background in IR and CLIR in order to avoid a bias towards a particular task during the translation process. A sample of Oromo topics from the ad-hoc tracks of the CLEF-2007 is given in Figure 5.6. Normally, the searching process begins when a query is given as an input to the CLIR system. In the context of the ad-hoc tracks of the CLEF campaign, a search query does not uniquely identify a single object in the collection. Instead, several documents or data objects are expected to match the query, usually with different degrees of relevancy [99].

In CLIR, query translation refers to a technique or strategy in which a search request expressed in a given source language is processed and converted into a semantically equivalent search query in the target language(s). In this study we have focused on translating Afaan Oromo query topics into English search queries. Various important pre-processing tasks are required in order to identify and extract the appropriate key terms from Afaan Oromo topics (see Figure 5.7). For example, query topics prepared in Afaan Oromo should be segmented, tokenized, morphologically examined and stemmed before matching them with the bilingual dictionary entries for translation. Figure 5.7 and Figure 5.8 illustrates some of the major steps involved in query translation.

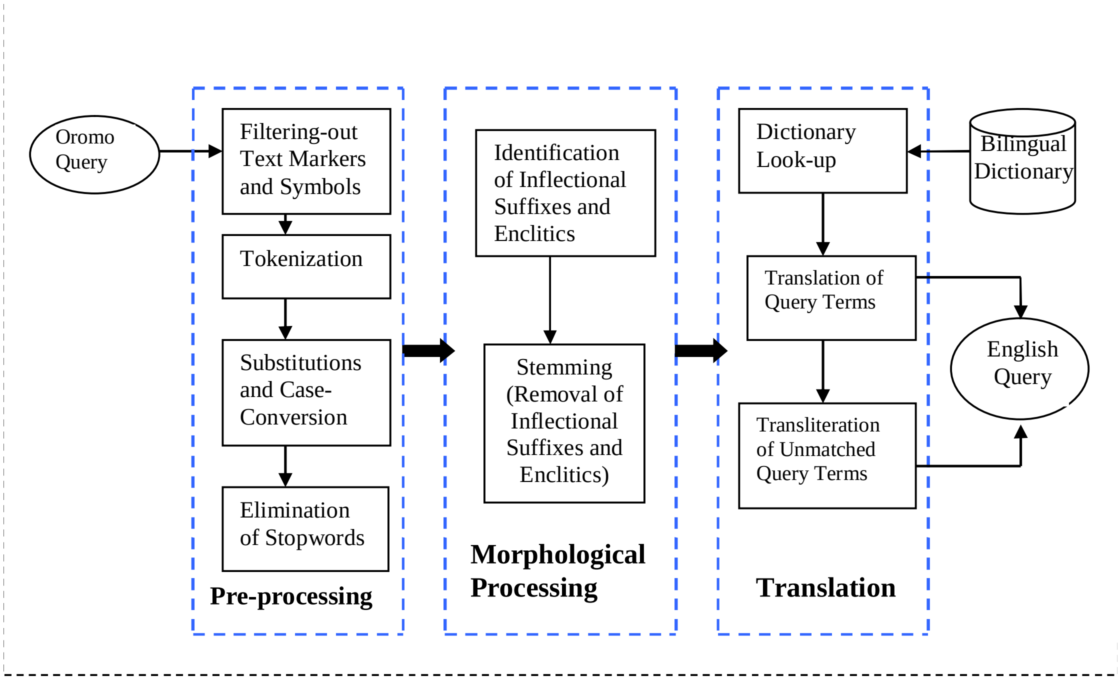
Oromo Query
Filtering-out Text Markers and Symbols
Tokenization
Substitutions and Case-Conversion
Elimination of Stopwords
Pre-processing
Identification of Inflectional Suffixes and Enclitics
Stemming (Removal of Inflectional Suffixes and Enclitics)
Morphological Processing
Dictionary Look-up
Bilingual Dictionary
Translation of Query Terms
Transliteration of Unmatched Query Terms
English Query
Translation

Figure 5.7 Major steps in Afaan Oromo query processing and translation

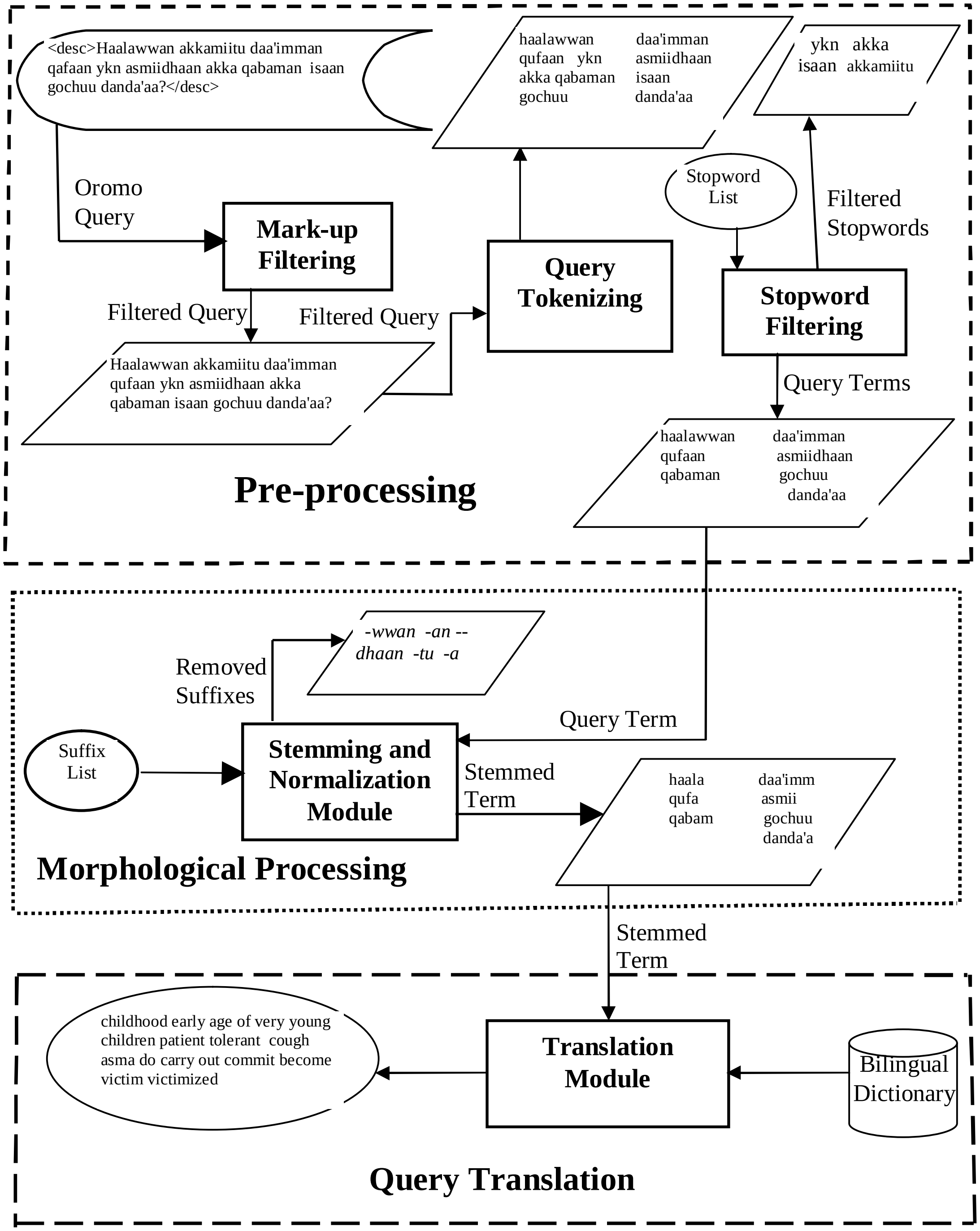
Pre-processing
<desc>Haalawwan akkamiitu daa'imman qafaan ykn asmiidhaan akka qabaman isaan gochuu danda'aa?</desc>
Oromo Query
Mark-up Filtering
Filtered Query
Haalawwan akkamiitu daa'imman qufaan ykn asmiidhaan akka qabaman isaan gochuu danda'aa?
Filtered Query
Query Tokenizing
haalawwan daa'imman qufaan ykn asmiidhaan akka qabaman isaan gochuu danda'aa
Stopword List
Stopword Filtering
Filtered Stopwords
ykn akka isaan akkamiitu
Query Terms
haalawwan daa'imman qufaan asmiidhaan qabaman gochuu danda'aa
Morphological Processing
Removed Suffixes
-wwan -an -- dhaan -tu -a
Query Term
Suffix List
Stemming and Normalization Module
Stemmed Term
haala daa'imm qufa asmii qabam gochuu danda'a
Stemmed Term
Query Translation
Translation Module
Bilingual Dictionary
childhood early age of very young children patient tolerant cough asma do carry out commit become victim victimized

Figure 5.8 Major steps in Afaan Oromo query processing and translation,

Some of the major steps involved in automatically processing and translating Oromo queries into English queries are summarized as follows.

1. **Tokenization:** Given sequence of character within a document or a query, tokenization refers to the task of chopping it up into pieces, called tokens. It is the process of identifying and recognizing appropriate linguistic units that can be treated as distinct query words. It usually involves separation and segmentation of words as well as isolating them from punctuation marks or other related formatting and mark-up symbols. Many text processing programs including query translation require that the input text should be formatted in a certain prescribed way. For instance, Oromo query terms should be separated from control characters and mark-up symbols like XML tags. Accordingly, each the Afaan Oromo query topics obtained from the CLEF campaign have been segmented and tokenized after punctuation marks and query formatting notations were filtered out (see Figure 5.7 and Figure 5.8). A summarized statistics of tokenized Afaan Oromo queries is shown in Table 5.3.

CLEF-2007 Oromo Topics			CLEF-2006 Oromo Topics	
Query Tokens		Unique Words	Query Tokens	Unique Words
Total Number	2378	1428 (60%)	2355	1555 (66%)
Average Length	47.56	28.56 (60%)	47.10	31.10 (66%)

Table 5.3 Statistics of tokenized Afaan Oromo topics

2. **Removal of Stop-words:** As indicated in section 5.3, the identification and removal of stopwords is useful in IR. To this end, we have constructed a list of Afaan Oromo stopwords that can be easily used for removing function words (see Appendix A). We used this list of stopwords to eliminate non-content bearing terms from Oromo query topics. Table 5.4 gives a statistical summary of unique query terms before and after elimination of stopwords from Oromo topics, which had been used in the ad-hoc tracks of the CLEF 2006 and CLEF 2007. The CLEF 2006 topics contains about 1555 unique query words out which about 610 words, which includes frequently used query terms or phrases such "*barbaadi*", "*dokumantooni*" "*ibsu*", and "*qaban*" in Afaan Oromo and "*find*", "*documents*" "*items*" and "*explain*" in English. After eliminating the stopwords, the remaining Oromo query terms, about 945 unique words, were considered for further morphological processing and translation. Likewise, the Oromo topic dataset for the CLEF 2007 contains 1428 unique query words out which about 450, which includes frequently used phrases or elaboration terms, were found to be stopwords. After

eliminating the stopwords, the remaining CLEF-2007 Oromo query terms, about 978 unique words, were considered for further morphological processing and translation (see Table 5.4, Figure 5.7 and Figure 5.8).

Query Filtering Stages	Oromo Topics (CLEF-2007)		Oromo Topics (CLEF-2006)	
	Number	Av. Length	Number	Av. Length
Before Stopword Elimination	1428	28.56	1555	31.10
After Stopword Elimination	978 (68.50%)	19.56	945 (60.50%)	18.90

Table 5.4 Statistics of Oromo topics before and after elimination of stopwords

3. **Stemming:** As described in detail in section 5.3, one of the common problems in the process of query translation is associated with word form variations. We have used our rule-based stemmers to remove frequent inflectional suffixes found in Afaan Oromo query topics. As shown in Figure 5.8, inflected query terms were morphologically analysed and conflated to their common stem or base form through the application of our rule-based stemmers that have been designed and developed during the course of this study. The impacts of our Afaan Oromo crude and light stemmers on OMEN-CLIR are described in section 6.2.

4. **Dictionary Look-up and Query Translation**: Oromo query terms were translated into English (target language) using bilingual dictionary. We have used automatic query translation technique in conducting our CLIR experiments. As indicated in section 5.1, our machine-readable bilingual dictionary is used as the main source of knowledge for query translation. Each of the normalized Afaan Oromo query term is automatically looked-up in the bilingual dictionary and converted into equivalent English search terms whenever a matching lexical entry is found. Since advanced word sense disambiguation tools are not available for Afaan Oromo, we opted to keep all possible translations as a candidate English search query. Since our bilingual dictionary is medium-size general dictionary, its coverage for proper names and technical terms is very limited. For instance, as shown in Table 5.5, the 50 different Oromo query topics prepared for the ad-hoc task of the CLEF 2007 campaign contains 978 query terms or key words out which only about 835 were found in our bilingual dictionary. Since the remaining 143 query terms, most which are proper names and technical terms, did not occur in the bilingual

dictionary, we have used a transliteration technique in order to convert them into potentially equivalent English search terms.

	CLEF-2007		CLEF-2006	
	Number	Percent	Number	Percent
Translated Terms	835	85.4%	791	83.7%
Untranslated Terms	143	14.6%	154	16.3%
Total	978	100%	945	100%

Table 5.5 Summary of Query Translation

5. **Transliteration:** As mentioned earlier, although bilingual dictionary can provide a reasonably accurate translation for commonly used words, it has limited vocabulary coverage. A bilingual dictionary has limited coverage of named entities, technical words and phrasal terms. In particular, a general bilingual dictionary often does not contain or cover proper names, compound words and technical terms. Since such query terms are important for the effectiveness of CLIR, the need for handling out-of-vocabulary (OOV) words becomes very critical for accurate query translation. Most of the unmatched or untranslated query terms in Afaan Oromo topics were found to be proper names, technical terms and loanwords from foreign languages, (e.g., *Iraaqi, Kurdi, Buush, fiilmi, "kilooniingiin"*). We have tried to address this problem using a phonetic-based transliteration technique, which modifies and converts out-of-vocabulary words into their potentially equivalent spellings in English. To this end, we have adopted a transliteration scheme that has been used by some Afaan Oromo lexicographers and translators. It converts unmatched or untranslated Afaan Oromo query terms into potentially equivalent English words based on phonetic and orthographic rules designed for this purpose. In this regard, advantage has been taken of the similarity in the writing system used in Afaan Oromo and in English (i.e., Latin alphabets).

6 Evaluation

Evaluating the performance of retrieval system is one of the fundamental factors for technological advancement in IR and CLIR. Evaluation is crucial for the development and maintenance of a CLIR system. It allows the developers to test and determine the effectiveness of the retrieval system. Apart from designing and building the basic components of our OMEN-CLIR, for which most of the required linguistic resources and translation tools have been constructed and developed from scratch, another major contribution of this study is to assess the performance of the proposed retrieval system at a well-recognized international evaluation forum. To this end, we had participated in one of the well-recognized international cross-language evaluation competitions (i.e., CLEF-2006 and CLEF-2007) over the past couple of years [25, 26, 27]. CLEF has been supporting and promoting comparative evaluations of various cross-lingual and multilingual information retrieval systems over the last decade [104, 7]. The participation of researchers in such well-recognized international evaluation campaigns plays a key role not only in allowing them to test and measure the performance of their CLIR system against well-established test-collections and evaluation techniques, but also in enabling them to compare the performance of their retrieval system with the performances of other CLIR systems tested on the same data sets. The results of such comparative evaluation campaigns provide some clues on the advantages and drawbacks of adopting different approaches and resources in developing CLIR for resource-scarce languages. In this chapter, we present and discuss some of the results of evaluation experiments that have been conducted to assess the performance of OMEN-CLIR.

6.1 Experimental Settings

6.1.1 Test Collections

We have conducted a series evaluation experiments on OMEN-CLIR using very large-scale test collections obtained from the ad-hoc track of the CLEF campaign. Table 6.1 presents a summarized statistics of the test collections. The English test collections that we have used for evaluating the performance of OMEN-CLIR contain more than 300,000 documents, which were published in different national newspapers during 1994, 1995 and 2002. These test collections were originally obtained from Los Angeles Times (i.e. LAtimes94, and LAtimes2002) and

Glasgow Herald (GHA95). Table 6.1 gives more details information about the test collections including the size and number of documents as well as number of query topics we have employed in conducting the bilingual retrieval experiments.

Category of the Test Datasets	Label of the Test Collection	Documents Size (in MB)	Number of Documents	Number of Query Topics
CLEF-2006	LAtimes94	425	113005	50
	GHA95	154	56472	
CLEF-2007	LAtimes2002	434	135,153	50
Total	NA	1,013	304,630	100

Table 6.1 Statistics of test collections and query topics

6.1.2 Query Topics

As described in section 5.4, we have assessed and evaluated the performances of OMEN-CLIR by using two different sets of Afaan Oromo query topics that were distributed by the ad-hoc tracks of the CLEF-2006 and the CLEF-2007 campaigns. The original sets of English topics, which were supplied by the organizers of CLEF campaign were translated into Afaan Oromo query sets by a group of translators who are native speakers of Afaan Oromo and fluent in English. It is important to note that the task of translating some English query topics, especially scientific and technical terms, into Afaan Oromo queries has been a huge challenge because of the cultural and linguistic disparities between the two languages. It is, indeed, hard to find the translations of some English technological terms in Afaan Oromo dictionaries. This is simply because most technological terms are as foreign as digital technologies to indigenous African languages.

Normally, query topics designed for the Ad-hoc tracks of the CLEF campaigns are structured statements similar to an essay question (see Figure 6.1). In other words, unlike ordinary search queries, CLEF query topics are expressed in relatively longer statements, which often observe proper syntactic and grammatical structures. As indicated in figure 6.1, each topic consists of three major fields: *title, description, and narration.* While a *title* is a phrase or a very short statement that indicates the basic subject matter of desired documents, a *description* is a sentence that provides a brief description or definition about the search topic. Likewise, a *narration* provides more elaborative statements or descriptions about the search query. Statements in the narration fields can also be used to restrict the search results as they often provide certain criteria on how the relevancy or irrelevancy of the search results could be judged.

According to the general guidelines for the ad-hoc track of the CLEF campaign, a query topic presented in a given source language could be translated into the target language by using different approaches or techniques. Either manual or automatic approaches could be employed for construction and formulation of the target language search queries. We have adopted the latter approach. Afaan Oromo query topics were automatically processed and translated into English search queries using our Oromo-English bilingual dictionary. The query translation is carried out based on word-by-word matching. Experimenting with different CLIR approaches, [94] had found that keeping all possible translations as candidate search terms of the target language is a good strategy in the context of a dictionary-based CLIR. The researchers had argued that one should not try to select a particular translation unless he/she is very sure about the correctness of the term(s) that is going to be chosen among the other alternative translations. The effectiveness of CLIR system may not be damaged nearly as much by adding extra incorrect translations as it is by leaving out some of the correct ones [94]. Taking these factors into account and due to lack NLP tools for disambiguation, all possible translations of Afaan Oromo query words were used as candidate English search terms. Apart from a short (title) query, we have automatically constructed several search queries by combining different fields in Afaan Oromo topics. Each of the translated English search queries were fed into Lucene text search engine, which was adopted for searching and retrieving English documents. Accordingly, we had used several search queries constructed from 100 different Afaan Oromo topics (50 topics from the CLEF-2006 and another 50 topics from the CLEF-2007) in conducting evaluation experiments on OMEN-CLIR.

```
<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>
```

Figure 6.1 Sample Afaan Oromo query topic from the CLEF-2006

6.1.3 Relevance Judgment

The main purpose of any functional IR system including CLIR is to identify and retrieve relevant documents from the target collections. Hence, availability of appropriate set of relevance assessment is vital in evaluating the performances of a CLIR. A set of relevance judgment is required in order to determine the documents that are relevant to a given query topic. Once the set of relevance assessment is established, the performance of a CLIR system can be easily assessed and evaluated in terms of its ability to retrieve the documents that are known to be relevant to a specific query topic. However, due to the vastness and very large number of documents in the ad-hoc track of the CLEF test collections, it is usually impractical to judge each and every document for relevance. Thus, the CLEF campaign uses various methods to ensure a high degree of consistency in determining the sets of relevance judgments [7, 104]. Generally, in very large-scale evaluation competitions like the TREC and CLEF, it is assumed that the effectiveness of IR systems can be objectively assessed by analysing a representative set of sample search results. Approximate recall values are calculated using pooling techniques, in which the results submitted by group participants in specific tasks are used to form a pool of documents for each topic. This is accomplished by selecting and collecting the top ranked documents or search results from the various runs that are submitted by different participants based on certain sets of predefined criteria. Traditionally, the top 100 ranked documents from each of the selected runs are included into the pool [7]. This pool is then examined and used for subsequent relevance judgments by human experts. After searching and retrieving the best matching documents from the target collection for each of the query topics, the performance of the CLIR system can be evaluated by comparing the search results of each topic against the corresponding set of documents in the relevance assessment.

6.1.4 Metrics for Performance Evaluation of CLIR

Building a successful and sustainable CLIR requires suitable methodologies and metrics for assessing its effectiveness. In the context of system-oriented evaluation techniques, the effectiveness of a CLIR system is measured by the extent to which it is able to identify and rank more relevant documents on the top of the search results. Chief among the important evaluation measures that are widely used for determining the effectiveness CLIR system include *precision* and *recall*, which can be defined as [98]:

- *Precision*: refers to the percentage of retrieved documents that are relevant to the search topic. It is defined by the amount of relevant documents retrieved compared to all documents retrieved.
- *Recall*: refers to the percentage of relevant documents which have been retrieved in response to a given search query. In other words, recall is defined by the amount of relevant documents retrieved compared to total number of relevant documents available in the target collection.

Some of the retrieval effectiveness measures that are commonly used include: *average precision* (i.e. the average of precision after each relevant document is retrieved), *mean average precision* or *MAP* (i.e. non-interpolated average precision over all relevant documents), *and precision@n* (i.e. the precision after *n* documents have been retrieved). Since most of the users of CLIR systems are less likely to go through large numbers of retrieved documents, it is very important to rank relevant items at the top of the search results. Based on the standard experimental settings within the ad-hoc track of the CLEF campaign, the performances of our OMEN-CLIR system have been evaluated in terms of the extent to which it ranks the relevant documents on the top of about 1000 search results for each of the query.

6.2 Analysis of Results and Discussion

6.2.1 Results

This section presents the results and analysis of the findings of the study. We used the English test collections and Afaan Oromo topics described in the foregoing sections to evaluate the performance of OMEN-CLIR. As shown in Table 6.2, we had conducted and submitted various official runs to the CLEF-2006 and CLEF-2007 annual campaign. Apart from the official runs that were submitted to the ad-hoc track of the CLEF campaign, we have also conducted various additional retrieval experiments in order to improve the performance of OMEN-CLIR. The evaluation experiments that we had conducted could be distinguished from one another in terms of the data sets (i.e., the CLEF-2006 vs. the CLEF-2007), and the topic fields from which the search queries were automatically constructed and translated. Although most of our evaluation experiments were conducted with the application of Afaan Oromo stemmer, a few of them were carried out without using the stemmer in order to identify its impacts on the performances of OMEN-CLIR. Table 6.2 gives the descriptions of the retrieval experiments including both official and unofficial runs. While the "*Run-Ids*" prefixed with "*NOST_*" were conducted without applying a stemmer, the rest of retrieval experiments were conducted with the application of our rule-based stemmers, i.e., either crude stemmer or light stemmer. While the "*Run-Ids*" prefixed with "*LTST_*", were carried out by applying Afaan Oromo light stemmer, the "*Run-Ids*" prefixed with "*CRST_*" were performed with the application of Afaan Oromo crude stemmer. Moreover, while the "*Run-Ids*" suffixed with "*-06*" shows the evaluation experiments that were performed with the CLEF-2006 test data sets, the "*Run-Ids*" suffixed with "*-07*" shows experiments that were carried out using the CLEF CLEF-2007 test data sets. We used three different metrics to measure the performance of OMEN-CLIR. First we provide the overall performance of OMEN-CLIR using mean average of precision (MAP) of each run. In a multilingual information retrieval environment like CLIR, users are often interested in browsing the top few retrieved search results. Hence, we provide measures for the *top n* documents retrieved (or precision@n) in the last two columns of Table 6.3. We provide the recall-precision scores at 11 standard points in the fifth column of Table 6.3.

Run-Id	Official Run?	Run Description
CRST_OMT06	No	***Title*** run with the application of crude stemmer
CRST_OMT07	No	
LTST_OMT06	Yes	***Title*** run with the application of light stemmer
LTST_OMT07	Yes	
CRST_OMTD06	No	A combination of ***Title*** and ***Description*** run with the application of crude stemmer
CRST_OMTD07	No	
LTST_OMTD06	Yes	A combination of ***Title*** and ***Description*** run with the application of light stemmer
LTST_OMTD07	Yes	
CRST_OMTDN06	No	A combination of ***Title, Description*** and ***Narrative*** run with the application of crude stemmer
CRST_OMTDN07	No	
LTST_OMTDN06	Yes	A combination of ***Title***, ***Description*** and ***Narrative*** run with the application of crude stemmer
LTST_OMTDN07	Yes	
NOST_OMT06	No	*Title* run without stemming
NOST_OMT07	No	
NOST_OMTD06	No	A combination of ***Title*** and ***Description*** run without stemming
NOST_OMTD07	No	
NOST_OMTDN06	No	A combination of ***Title***, ***Description*** and ***Narrative*** run without stemming
NOST_OMTDN07	Yes	

Table 6.2 Description of official and unofficial runs

Table 6.3 shows summarised statistics of the evaluation results in terms of Mean Average Precision (*MAP*) and R-Precision (*R-Prec.*) values. The total number of *Relevant* documents (*#Rel*) and number of *Relevant* documents *Retrieved* (*#Rel-Ret*) are given in the second and third columns of the table. As indicated earlier, in the context of multilingual retrieval settings like the WWW, users are less likely to go through a large number of search results to identify relevant items because they may not be proficient in the target language. Hence we had conducted and presented the average precision scores of the top ranked documents, called documents cut-off levels or precision@n, (see P@10 and P@20) in the last two columns of Table 6.3. The official runs and their corresponding results are indicated in boldface. In our discussion, more emphasis is given to the MAP score presented in the fourth column of the table since it is one the most commonly used metrics for effectiveness of CLIR.

As shown in Table 6.3, the retrieval experiments conducted with the application of Afaan Oromo stemmers have performed much better than the experiments carried out without stemming. This is due to the fact that Afaan Oromo is a highly synthetic and agglutinative language. By identifying and removing inflectional affixes, our stemming algorithms were able to normalize and group different word form variants into smaller conflation classes or base forms that could be found in the bilingual dictionary.

Run-Id	#Rel.	#Rel-Ret.	MAP	R-Prec.	P@10	P@20
CRST_OMT06	1,258	638	0.1895	0.1768	0.1823	0.1452
CRST_OMT07	2,247	1354	0.1993	0.2036	0.2465	0.2237
LTST_OMT06	**1,258**	**870**	**0.2200**	**0.2433**	**0.2380**	**0.1950**
LTST_OMT07	**2,247**	**1,554**	**0.2420**	**0.2624**	**0.3380**	**0.2880**
CRST_OMTD06	1,258	673	0.2054	0.1735	0.2137	0.1758
CRST_OMTD07	2,247	1557	0.2456	0.2534	0.3624	0.2736
LTST_OMTD06	**1,258**	**848**	**0.2504**	**0.2624**	**0.2660**	**0.2310**
LTST_OMTD07	**2,247**	**1707**	**0.2991**	**0.3063**	**0.4200**	**0.3470**
CRST_OMTDN06	1,258	643	0.1953	0.1735	0.2382	0.1669
CRST_OMTDN07	2,247	1587	0.2282	0.2314	0.2836	0.2543
LTST_OMTDN06	**1258**	**892**	**0.2450**	**0.2572**	**0.2780**	**0.2220**
LTST_OMTDN07	**2247**	**1693**	**0.2894**	**0.2973**	**0.4320**	**0.3380**
NOST_OMT06	1258	605	0.1687	0.1241	0.1476	0.1165
NOST_OMT07	2,247	1224	0.1736	0.1808	0.2104	0.1639
NOST_OMTD06	1,258	582	0.1820	0.1571	0.1643	0.1387
NOST_OMTD07	2,247	1333	0.2010	0.1955	0.3467	0.2853
NOST_OMTDN06	1,258	568	0.1421	0.1436	0.2687	0.2376
NOST_OMTDN07	**2,247**	**1439**	**0.2038**	**0.2217**	**0.3140**	**0.2600**

Table 6.3 Statistical summary of evaluation results

The highest MAP, (i.e. 29.91%) has been scored by a LTST_OMTD07 run, which was conducted with application of our light stemmer by using queries constructed from title and description fields of Oromo topics. This best performance was closely followed by the LTST_OMTDN07 run, which yielded about 28.94% MAP. As indicated in Table 6.2, the LTST_OMTDN07 run, which was conducted with a relatively longer search queries constructed from the combination of "*Title*", "*Description*" and "*Narration*" fields. The experiments that were carried out with very short queries, (i.e., the "*Title*" field, e.g. LTST_OMT06 and LTST_OMT07), were scored a relatively lower MAP when compared with the other runs, because longer queries were able to provide better description of user's information needs than shorter queries. However, even though longer queries have increased the effectiveness of OMEN-CLIR, this improvement is not linearly proportional to a query length. As shown in Table 6.3, a query with medium length (e.g., LTST_OMTD07) has scored a better MAP than a very long query (e.g., LTST_OMTDN07). Although the retrieval experiments conducted with the CLEF-2007 dataset had performed better than the experiments conducted with the CLEF-2006 dataset, the scores of our top runs in both annual competitions have been found to be almost equivalent by achieving about 77% of the best bilingual run in the corresponding bilingual retrieval tasks. On the other hand, while our top official run (experiment) at the CLEF-2006 (i.e. 25.04%) was achieved about 60.40 of the English monolingual baseline, our best official run (experiment) at the CLEF-2007 was achieved about 67.95% of the English monolingual baseline. In spite of the fact that the retrieval performance achieved by OMEN-CLIR is lower than the state-of-the-art CLIR performances, (which usually ranges from 80% to 95% of the monolingual baseline for closely related European languages), we found most of the evaluation results to be very promising and encouraging, given the limited amount of linguistic resources and translation tools that have been employed in developing OMEN-CLIR.

In addition to assessing the overall performances of OMEN-CLIR, we have been also interested in assessing the effects of Afaan Oromo stemmers on the performances of the OMEN-CLIR. To this end, we had designed and carried out various retrieval experiments on three different types of Afaan Oromo search queries, (i.e. short queries constructed from the "*Title*" field; medium queries constructed from the "*Title*" and "*Description*" fields, and long queries constructed from all of the three topic fields). Table 6.4 shows the results of these experimental results in terms of mean average precision (MAP). Best scores are indicated in boldface. The results of the base runs that were conducted without the application of Afaan Oromo stemmer as well as the rate of changes or gains caused by the application of Afaan Oromo crude stemmer and light stemmer over the corresponding base runs are given in Table 6.4.

Category of Test Data	Query Type	Using Crude Stemmer	Using Light Stemmer	Without Stemmer	Change (in %)	
					Crude Stemmer	Light Stemmer
CLEF-2006	Short query	0.1895	0.2200	0.1687	12.33	30.40
	Med. query	0.2054	**0.2504**	0.1820	12.86	37.58
	Long query	0.1953	0.2450	0.1721	13.48	**42.36**
CLEF-2007	Short query	0.1993	0.2420	0.1736	14.80	39.40
	Med. query	0.2456	**0.2991**	0.2010	22.19	**48.80**
	Long query	0.2450	0.2894	0.2038	20.22	42.00

Table 6.4 Impacts of Afaan Oromo stemmers

The evaluation results given in Table 6.4 show that a considerable improvement of OMEN-CLIR performance as result of the applications of Afaan Oromo stemmers. As clearly shown in Table 6.4, the retrieval experiments conducted with the application of Afaan Oromo light stemmer had considerably outperformed the base runs. More specifically, while the application of Afaan Oromo crude stemmer had improved the performances of OMEN-CLIR by average precisions of 12% to 25% over the base runs, the application of the light stemmer wad increased the performances of OMEN-CLIR by 30% to 50% in comparison with the base run.

A number of Afaan Oromo query terms that were used in the base runs (without applying the stemmer) yielded lower scores due to the problem of morphological variations. Most inflected words found in Afaan Oromo queries were not translated since they did not match with lexical entries provided in the bilingual dictionary. Inflectional suffixes associated with Afaan Oromo query terms had severely degraded the performances of the base runs. On the other hand, the application of Afaan Oromo light stemmer had achieved much better performance because it has reduced the problem of word form variants. This implies that the development and application of Afaan Oromo stemmer is very important for improving the performance of our CLIR system. Since Afaan Oromo is highly synthetic and agglutinative language, the development of the stemmer should be further explored and fine-tuned. A number of earlier studies on the impacts of stemming algorithms for morphologically very complex languages like German and Dutch had shown significant improvements in retrieval effectiveness. For instance, [105] had applied linguistically motivated stemming algorithms which have improved the retrieval performance by 43% and 30% for German and Dutch respectively. The researcher had also reported significant contributions and benefits using stemming algorithms in other Romance languages such as French, Italian, and Spanish.

Figure 6.2 and Figure 6.3 shows the recall-precision curves for some of the major retrieval experiments that we have conducted using the CLEF-2006 and CLEF-2007 datasets. The performances of each of the retrieval experiments was measured and represented by an interpolated average precision curve over standard 11 recall levels. Both figures contain precision-recall plots for retrieval experiments conducted without the application of stemming (as base runs e.g. NOST_OMTD06 and NOST_OMTD07) as well as the retrieval experiments conducted with the application of Afaan Oromo light stemmer.

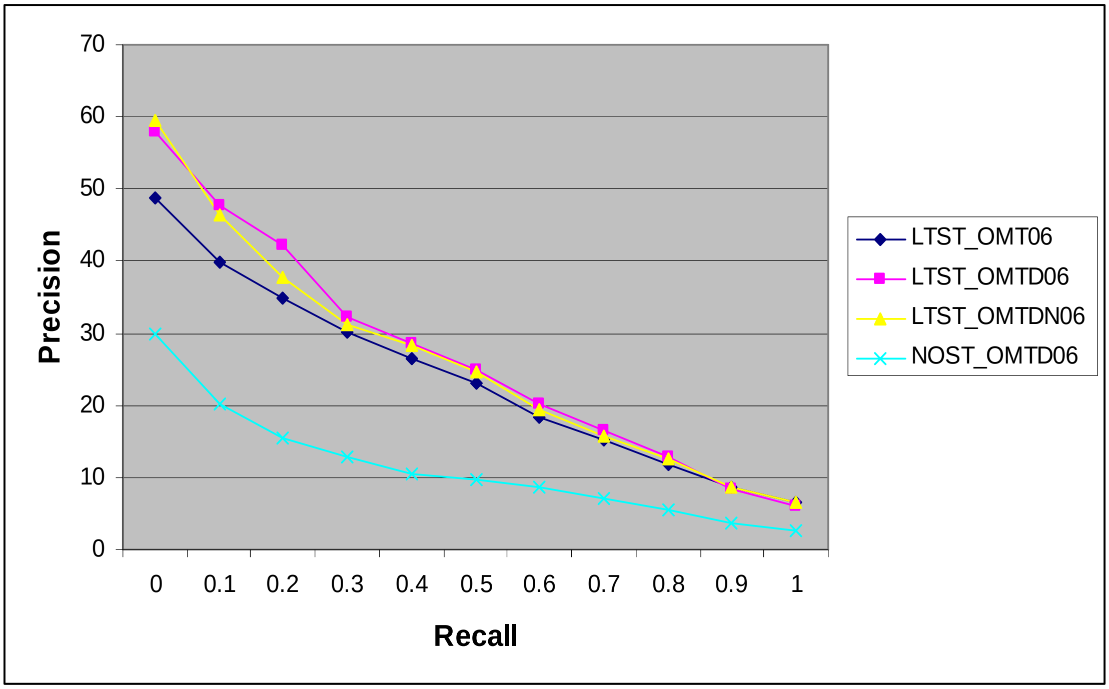

Figure 6.2 Recall-Precision curve for the CLEF-2006

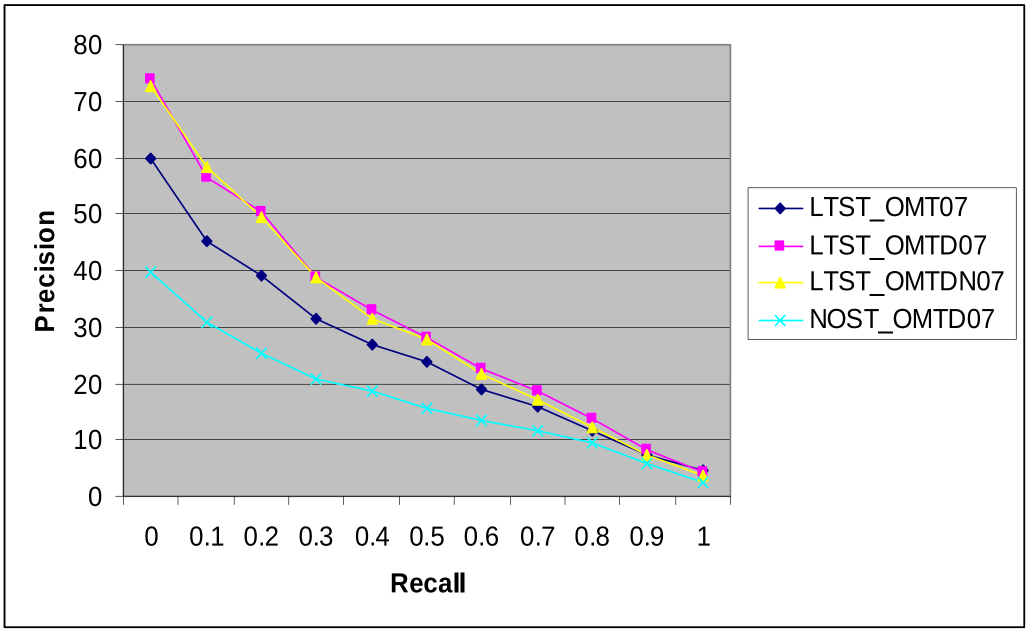

Figure 6.3 Recall-Precision curve for the CLEF-2007

Similar to the evaluation results discussed above, the retrieval experiments conducted with the application of Afaan Oromo light stemmer had outperformed the base runs that were conducted without the application of the stemmers. NOST_OMTD06 and NOST_OMTD07 runs, which were conducted without the application of Afaan Oromo stemmer, had very low precision-recall curves. Among the different runs that have been conducted with the application of Afaan Oromo light stemmer, LTST_OMTD07 has yielded the highest precision-recall curve followed by LTST_OMTDN07. Thus, the application of our light stemmers has resulted in substantial improvements in the performances of OMEN-CLIR. Even though Afaan Oromo is very agglutinative language, and its morphological phenomena still needs more detailed linguistic studies and computational models, our experimental results demonstrates the viability of developing and employing light-stemming algorithm for identifying and normalizing Afaan Oromo word form variants.

6.2.2 Significance Testing

One of the objectives of this study is to identify linguistic resources and IR methods that can significantly improve the effectiveness of OMEN-CLIR. Given two or more different sets of search results produced by using different linguistic resources or tools such as stemmer and lemmatizer, determining the significance of applying such resources is a non-trivial task. As described in the preceding section, in common batch-style CLIR experiments like the CLEF ad-hoc retrieval evaluations, the effectiveness of each retrieval method and system is often measured based on common IR metrics such as precision and recall, especially mean average precision (MAP). However [106, 107], in some instances, it is possible that the average measurements may not be enough to indicate and describe the overall performance change or gain between different retrieval methods. Hence, performing a significance test to determine whether the differences in average measures between various retrieval methods can be considered statistically significant is necessary. Such statistical testing can play a key role in helping the researcher to judge the impacts of different retrieval tools and methods. According to [107], statistical tests can be very useful because they provide information about whether observed differences in evaluation scores are really meaningful or simply due to a chance.

Over the last few decades, various statistical testing techniques have been suggested and used to judge or determine whether the performance differences scored by different retrieval methods are statistically significant [107]. In the context of a paired difference test, the IR researcher often wishes to decide if one retrieval procedure significantly better than the other. Since the performance differences are much greater between queries than between retrieval methods, measurements should be considered as matched pairs, meaning that it is the difference between the scores for each query which is analysed [107]. One of the most common significance testing techniques that is widely used in IR evaluation is the *paired t-test.* Paired t-test is often used to determine if two sets of data are significantly different from each other. It compares the magnitude of the difference between the performances of two retrieval methods for each of the query topic. There is also a non-parametric alternative to the t-test, *the paired Wilcoxon signed-rank test*, which is also used to determine whether two matched groups of data are different. The Wilcoxon test replaces each difference with the rank of its absolute value. These ranks are then multiplied by the sign of the difference, and the sum of the ranks for each group is compared to its expected value under the assumption that the two groups are equivalent [107].

In order to determine the significance of using Afaan Oromo light stemmer, our preliminary assumption or null hypothesis is that the application of the light stemmer has no effect (or does not make any difference) on the performances of OMEN-CLIR. In other words, the performance of the retrieval methods conducted with the application of Afaan Oromo light stemmer and the base runs, which were conducted without the application the stemmer, are equivalent in terms of MAP. Based on the statistic calculated from each query result, we determine whether to accept or reject this null hypothesis. Our significance test disproved this null hypothesis by determining a *p-value*, a measurement of the probability that the observed difference could have occurred by chance [107]. The p-value is the probability of obtaining either the observed difference or a more extreme value of the difference between the two retrieval methods, purely based on chance. If the p-value is very low (say less than the threshold value of 0.05), we reject the null hypothesis and the contribution of our light stemmer is considered statistically sound or significant.

Using the official runs that had been submitted to CLEF-2007 ad-hoc retrieval evaluation, we computed the significance values for the paired t-test and Wilcoxon signed rank. Table 6.5 gives the statistical significance test interpretation of our experiments. For each of these two statistical tests and for each pair of runs, we measured the statistical significance of the difference in their mean average precision (MAP).

Categories of queries	**Paired t-test (p-value)**	**Wilcoxon sign test (p-value)**
LTST_OMTDN07 vs. NOST_OMTDN07 (Medium query vs. Non-Stemmed query)	0.000039	0.000080
LTST_OMTDN07 vs. NOST_OMTDN07 (Long query vs. Non-Stemmed query)	0.000046	0.000098

Table 6.5 Statistical significance test

As shown in Table 6.5, the *p-values* obtained for t-test, which are < 0.001 for all tests, demonstrate that the observed performance differences of the stemmed search queries (i.e., OMTD07 and OMTDN07) over the surface form or base run (i.e., NOST_OMTDN07) is significant at a 99% confidence interval for both medium and long queries in the paired t-test. Similarity, the observed difference is significant at the 99% level, (p-value < 0.001), using the non-parametric Wilcoxon test. Accordingly, the application of Afaan Oromo light stemmer has, indeed, made a big difference in improving the performance of OMEN-CLIR. The results demonstrate that the contribution of Afaan Oromo stemmer for the effectiveness of OMEN-CLIR is statistically significant.

6.3 Error Analysis

In addition to assessing and determining the overall performance of OMEN-CLIR, another major goal of this study is to make a survey of error analysis with a view of identifying the major challenges that need to be tackled to improve the effectiveness of our bilingual retrieval system. As indicated earlier, the effectiveness of CLIR system is heavily dependent on the quality of translation resources and the linguistic disparity of the language pairs involved. When the language pairs being considered are genetically and typologically related (e.g., French and English), the accuracy of query translation is very high because they often share many vocabularies as well as have similar morphological and syntactic structures. Cognates or similar words in two closely related languages are likely to have the same meaning and translation. On the other hand, for disparate language pairs with nothing or little in common (e.g., Afaan Oromo and English), the effectiveness of CLIR can be extremely low [95]. As the relationship between the language pairs differs more and more, the quality of the query translation decreases, whereas the chance of mistranslation increases. Like other translation resources and approaches, dictionary-based query translation has various limitations. While it is relatively more efficient and less expensive to build than parallel corpora and MT system, bilingual lexicon may not deliver very accurate query translation. A number of earlier studies had shown that the coverage and quality of a dictionary play a critical role for the effectiveness of a CLIR system [95, 81]. In fact, one of the major bottlenecks in developing CLIR for resource-poor indigenous languages, including Afaan Oromo, has been lack of reliable online bilingual and multilingual lexicons. As noted by [2], limited coverage of bilingual lexicon is one the most serious problems in improving the performance of a dictionary-based CLIR. If important query terms such as phrasal terms or scientific words cannot be accurately translated, they are not only missed from contributing to the retrieval of relevant items, but might be mistranslated and increases the retrieval of irrelevant items. In other words, a dictionary-based CLIR cannot be expected to return less irrelevant search results than the missing translations or misinterpretations occurred during the process of the query translation. According to [95], some of the major reasons for poor performance of CLIR in comparison with monolingual IR are: mismatching of search terms, lexical ambiguity and out-of-vocabulary (OOV) words. Query translation errors may occur because of various reasons including:

- Limited coverage of dictionary,
- Ambiguity of query terms,
- Lack of reliable morphological, semantic and syntactic processing tools, and
- Lack of efficient NE and MWEs recognizers.

Error Category	Number	Percent	Examples
Named Entities and Technical Terms	95	66.40%	“Buuleent Ekeviitii” (418), “Kosteliikaa” (421), Fortuyinii (447), “kosmootiiksii” (430), “Baaliitti” (409), “kilooniingiin” (408)
Phrasal Terms and Compound Words	26	18.20%	“harma hosistoota” “425”, “turistoota samii” (441), “olka'iinsa” (401), “waliigaltee” (410), “seeraan ala” (425)
Stemming Errors	14	9.82%	“baaliitti” (402), “qoricha” (415)
Acronyms and Abbreviations	2	1.20%	“naton” (404), “vipdha” (436)
Need Further Analysis	6	4.20%	“ofbakka” (402), “daareektarichaas” (411)
Total	143	100%	NA

Table 6.6 Categories of Errors in CLEF-2007 Topics

In order to identify the major problems in improving the performance of OMEN-CLIR, we have conducted a survey of error analysis with special focus on query topics with very low performance. As shown in Table 6.6, a number of important query words including technical terms, compound words, proper names, acronyms and abbreviations are not properly covered in our current Oromo-English dictionary. Obviously, missing important named entities and technical terms has degraded the retrieval effectiveness of OMEN-CLIR, especially when the search query is very short. Search topics with named entities have produced substantially worse results than other queries. We have also observed that the performance of domain specific queries (or technical terms) is much lower than the average performance of OMEN-CLIR. As shown in Table 6.6, the main problem of query translation is related to lack of adequate coverage for named entities, technical words, and phrasal terms or multiword expressions, which together accounted for more than 80% of the OOV words. While the translation of commonly used query words is quite accurate, the translation of most technical terms and named entities is not accurate enough to enable the search engine to retrieve and rank relevant documents on the top. Some of our major observations during the error analysis are summarized and presented in the subsequent subsections.

6.3.1 Mistranslation of Multiword Expressions (MWEs)

Multiword expressions (MWEs) or phrasal terms are lexical units which consist of two or more lexemes and whose meaning may be not derivable, or is only partially derivable, from the semantics of their constituents. As shown in Table 6.5, a few examples of Oromo MWEs are:

- "*of bakka buusanii", (*which is literally translated as "self-place" or "self-represent", but semantically refers to "self-renewable"),
- "*harma-hosistoota*", (which is literally translated as "breast feeding", but semantically refers to "mammals"),
- "*baduutti argaman"*, (which is literally translated as "found at loss" or "disappearing", but semantically refers to "endangered" or "at the risk of extinction"),
- "*seeraan ala*", (which is literally translated as "outside law", but semantically refers to "illegal" or "unauthorized").

The problem of translating MWEs poses a serious challenge in Afaan Oromo not only because they are not properly covered in bilingual dictionaries, but also because most English scientific terminologies are translated to Afaan Oromo as phrasal terms or compound words. As noted by [2], retrieving relevant documents in a language different from the query language is often affected by inaccurate translations of phrasal terms. A highly relevant document could be judged as less relevant or even irrelevant due to the semantic gap between the source language query words and target language search terms [95]. Indeed, due to word-for-word translation of Afaan Oromo phrasal terms to English, some of translated topics do not preserve the semantics of the original Afaan Oromo topics, and thus lead to a very low retrieval performance.

For example, the translation of Afaan Oromo phrasal term ''*harma-hosistoota*'' into English, (from CLEF-2007 topic no. "425"), produces "breast feeding", which may not be absolutely wrong. Yet the appropriate translation is supposed to be "mammals." Obviously, most users are less likely satisfied with the retrieval documents about "breast feeding" if they are looking for articles on “mammals”. In another topic (from CLEF-2007 topic no. “415”), the phrasal term “*seeran ala*” is translated into English as “outside law” missing the appropriate terms such as “illegal” or “unlawful”. Similarly, the translation of phrasal term “*tapha kubbaa miilaa*” (from the CLEF-2006 topic no. “344”) into English produces (“play” + “ball” + “foot or leg”), missing any of the more appropriate terms such as “soccer” or “football”. Because words in phrasal search term are treated as independent, they are translated into English separately rather than as a unit or part of one MWEs.

Moreover, some of the low performing Oromo query topics also contain ambiguous phrasal terms, which are very difficult to translate into English using a bilingual dictionary alone. For example, the phrasal term “*Afriikaa Kibbaatti*” (i.e., “South Africa” from the CLEF-2006 no. “343”) is not clear whether it refers to the country, (i.e. the “Republic of South Africa”) or the “southern parts (regions) of the continent”. Even if the coverage of the bilingual dictionary can be expanded to incorporate phrasal terms or named entities, such ambiguous query terms are likely to remain a problem for any dictionary-based CLIR. Although there are online dictionaries that can provide definition of phrasal terms and their translation probabilities for resource-rich European and Asian languages, such advanced and modern multilingual dictionaries do not exist for most resource-scarce African languages including Afaan Oromo.

One of the solutions that suggested by many CLIR researchers to reduce the problem of (MWEs) is to use special dictionaries such as bilingual or multilingual phrasal dictionaries. Indeed, such multilingual phrasal dictionaries and glossaries can provide more accurate translation of compound words and technical terms. Unfortunately, they are scarce for most indigenous African languages and very expensive to design and produce. A more promising and alternative approach to reduce the problem of MWEs terms is web mining, which will have a great potential for many African languages in the near future. As discussed in section 4.3.3, annotated comparable and parallel corpora are crucial resources for obtaining more accurate translation of MWEs and phrasal terms. The establishment annotated parallel corpora and development of Afaan Oromo MWEs recognition system is essential for further improvement of OMEN-CLIR. Sadly, although parallel corpora-based approaches dominate the development of CLIR and MT systems for many well-resourced European and Asian languages, most resource-scarce African languages are not yet in a position to enjoy the privilege of having large-scale parallel corpora.

6.3.2 Limited coverage of Named Entities and Lack of NER

Although the term *named entity* (NE) can be used to refer to the name of any object in the real world, in IR it is more specifically used to describe proper nouns, times and amounts. Named-Entity Recognition (NER) is a subfield of IR and Information Extraction (IE) that seeks to identify and classify named entities in a document into certain pre-defined categories such as the names of persons, organizations, locations, products, services, expressions of times, quantities, monetary values, percentages. One of the major sources of errors in our query translation is the limited coverage of named entities and technical terms. Typically, general concepts and commonly used words are selected for bilingual dictionaries. Unfortunately, from IR point of view, named entities and domain-specific concepts or special terms, which are called out-of-vocabulary (OOV) words, are often the most interesting ones. Since named entities and special terms are frequently used to identify relevant information across language boundaries, their accurate detection and translation is very important to improve the performance of CLIR. In this study, due to lack of Afaan Oromo NER system, we have faced a serious problem in identifying and translating or transliterating named entities such as names of persons, organizations, products and locations. In Afaan Oromo, named entities do not only appeared in different spelling form English, but also take inflectional suffixes to convey syntactic and grammatical functions like any other Oromo nouns. A few of such examples from CLEF0-2006 Oromo topics include proper names such as "*Chaarlas Mooti**chaatiifi** Diyaanaa Mooti**ttii***" (CLEF-2006 topic no. "329"), "*Afriikaa Kibbaa**tti***" (CLEF-2006 topic no. "343) and temporal expressions such as "*1995**tti**.*"

Although we have used our Afaan Oromo stemmer to remove inflectional suffixes before converting them into English using our transliteration scheme, we have observed that the transliteration of certain named entities are inaccurate, mainly owing to their foreign origin, especially differences in pronunciation, spelling and culture. For example, since the proper name "*Kosteliikaa*" (CLEF-2007 topic no. "421") was not found in our bilingual dictionary, it was wrongly transliterated into English as "Kostelik" instead of "Kostelic." Surprisingly, the inflected location name "*Baaliitti*" (CLEF-2007 topic no. "409") was wrongly translated into English as "leaf." After the process of stemming and normalizing, the inflected location name "*Baaliitti*" becomes "*baala*", an Oromo noun whose dictionary translation is "leaf". Accordingly, limited coverage of named entities in our bilingual dictionary and lack of appropriate NER system are found to be the major bottlenecks that hinder the effectiveness of OMEN-CLIR. Clearly, there is a compelling need for the development and application of Afaan Oromo named-entity recognizer in order to improve the performance of OMEN-CLIR. We also feel that incorporating the descriptions of frequently used acronyms and abbreviations into the dictionary may improve the coverage and translation quality of the bilingual lexicon.

In summary, the major types of query terms that are missing from the bilingual dictionary include proper names, technical terms, loan words and complex expressions such as “9/11”. One of the solutions that suggested by many CLIR researchers, including [94, 19], to overcome the problem of the limited coverage in a general bilingual dictionary is to identify and use certain domain-specific (special) dictionaries such as medical dictionaries, legal dictionaries, scientific glossaries and gazetteers. Indeed, such special bilingual dictionaries and directories can provide access to many named entities, uncommon vocabularies and technical terms [95]. Unfortunately, they are scarce for indigenous African languages and extremely expensive to prepare and produce. As described in section 5.5, transliteration, a method that we have adopted in this study, is another important supplemental technique that can be used for converting untranslated query terms into the target language, especially when the linguistic resources available for the language pair are very limited or incomplete. A more recent alternative approach to reduce the problem of OOV terms is web mining, which will have a great potential and promise for many African languages in the near future. As discussed in section 4.3.3, online comparable and parallel corpora are crucial resources for obtaining more accurate translation of OOV terms. The establishment annotated parallel corpora and development of Afaan Oromo named entity recognition (NER) system is essential for further improvement of OMEN-CLIR.

6.3.3 Stemming Errors

As discussed earlier, most indigenous African languages, including Afaan Oromo, pose a huge challenge for the development and effectiveness of CLIR, not only because they are severely under-resourced, but because they have very complex morphological systems. Although most frequent inflectional suffixes were successfully identified and removed based on the rules that have been implemented in our light stemmer, we have observed some under-stemmed and over-stemmed problems during our error analysis (see section 5.3.5 for more detail description). For example, the stems of some Oromo query terms such as “*alla**tti**n*”→ “*alla**tti*** + *-n*” (i.e., “bird”), “*bul**tii**n*”→ “*bul**tii*** + *-n*” and “*maa**ti**tti*” → “*maa**ti*** + *-tti*” (i.e., “family”), have endings similar to

Oromo inflectional suffixes or postpositions (i.e., false positives). As a result, our stemming algorithm has wrongly removed the endings of valid lexical items (indicated in boldface above). Consequently, appropriate lexical items (or semantically equivalent entries) found in the lexicon were not able to match with the query terms. Accordingly, there some instances where our stemmer has contributed to the translation errors.

Moreover, since we have not considered Afaan Oromo prefixes, reduplication and compound words in our current rule-based stemmer, some complex query terms such as "*itti-fayyadama*" (CLEF-2007 topic no. "402"), "*wal-laaqiinsa*" and "*gu-gguuruu*" (CLEF-2007 topic no. "410") were not properly stemmed and translated into English. As indicated in the preceding section, the proper name "*Baaliitti*" → ("*Baalii* + *-tti*"), (i.e., "at Bali", an island in Indonesia from CLEF-2007 topic no. "409") was wrongly stemmed and incorrectly translated. While the first step stemming, i.e., the removal of the attached postposition "*-tti*" was appropriate, the second step stemming, i.e., the removal of the long final vowel "*ii*" was inappropriate, and thus leads to the mistranslation of "*Baalii"* to "leaf" in English. Even though we have recently tried to identify and store some of Afaan Oromo exceptional words in a separate list, it is not yet comprehensive enough to avoid the problem of over-stemming. Hence, a more comprehensive list of exceptional words needs be prepared and consulted during morphological processing and stemming of Afaan Oromo queries in order to prevent the stemmer from removing valid endings.

7 Conclusions and Future Directions

7.1 Conclusions

The WWW is a crucial sources information and knowledge that needs to be universally and efficiently accessible to all users. In the digital age we live in today, knowledge has increasingly become a critical engine for social and economic development. Even though physical and geographical barriers have been drastically reduced or virtually eliminated by the rapid development of the Internet, language barriers and linguistic digital divide have emerged as the major obstacles to the accessibility and usability of online information resources and services. The main focus of CLIA is to unlock language and cultural barriers to make sure that online resources are efficiently and equitably accessible to all nations and citizens. This study is intended to contribute to making CLIA systems extendable and useable for resource-scarce African languages. We have presented the major approaches and resources that we have developed and adopted for building and evaluating OMEN-CLIR, which has achieved a promising performance by using very limited linguistic resources. Most of the earlier studies were focused on building CLIR using very rich linguistic resources and translation tools such as machine translation systems and parallel corpora. But such advanced language resources are not readily available for most indigenous African languages, making it exceedingly difficult to perform CLIR for severely under-resourced language like Afaan Oromo. Although closing the linguistic digital divide is a major challenge not only for resource-scarce African languages but also for technologically more advanced ones, African languages need a special research attention to reduce the widening gap in computational linguistic resources and information access technologies. In this study, we have explored and demonstrated the viability of building for CLIR for one of the most severely under-resourced African languages, (i.e., Afaan Oromo).

Most resource-scarce African languages present immense challenges to the development of CLIA systems. From the very outset of our study we faced chronic lack of electronic data and translation resources for Afaan Oromo. Indeed, building a CLIR for Afaan Oromo has proved to be time-consuming and labour intensive because automated linguistic resources and translation tools such as machine-readable bilingual dictionaries have been designed and developed from scratch. Exploring and constructing basic linguistic resources have been proven to be the biggest challenges to getting our research project off the ground. As described in section 2.3 and section 2.4, indigenous African languages suffer from not only the chronic shortage of online linguistic and translation resources but of the fragmentation of research efforts in building these basic resources. This problem often leads to very small, uncoordinated and individual research works that have been only usable for limited purposes. Most existing African language resources and technologies have been developed and evaluated in isolation, without any benchmarking and standard test datasets. Currently, the availability of large-scale monolingual and multilingual text corpora is out of the reach of most resource-scarce African languages. This study suggests that there is a critical need for more coordinated and collaborative research efforts in Africa to create the missing computational linguistic resources and CLIA technologies. Appropriate actions and steps should be taken by concerned government and non-government organizations to prevent the linguistic digital divide from even getting wider. Building CLIA systems should be encouraged to accelerate the development and application of indigenous African language on the Internet.

In this study, we have investigated and demonstrated the viability of building CLIR for one of the major indigenous African languages without relying on very rich linguistic resources. In addition to achieving a very promising and reasonable performance level, OMEN-CLIR is notable for its success using only very limited amount of linguistic and translation resources. As described in section 6.2, we have conducted different evaluation experiments by using very large test collections and two different query sets obtained from the ad-hoc tracks of the CLEF campaigns. We have assessed the performances of OMEN-CLIR by using more than 300,000 documents, i.e. English test collections. We have used several search queries constructed from different fields of Afaan Oromo topics.

However, from our initial evaluation experiments with surface forms of Afaan Oromo queries, we had observed that the performance of OMEN CLIR is very low mainly owing to the mismatch between the citation forms of the bilingual dictionary and the inflected forms of query terms. To reduce the problem of morphological variations and ambiguities, we proposed and constructed two closely related rule-based Afaan Oromo stemmers. Besides assessing the overall performances of OMEN-CLIR by formulating and using different types of queries, we have evaluated the effects of Afaan Oromo stemmer on the performances of OMN-CLIR. The application of Afaan Oromo light stemmer has substantially improved the performances of the cross-language retrieval system. In fact, most of our experimental results have shown that the application of our Afaan Oromo stemmers has consistently improved the performance of OMEN-CLIR, regardless of the lengths or types of the search queries that were employed in the retrieval experiments.

As indicated in Table 6.4, although the application of both crude stemmer and light stemmer has improved the performance of OMEN-CLIR, the contribution of the latter was found to be considerably superior. While the applications of the crude stemmer have increased the performances of OMEN-CLIR by the average precisions of 10% to 25%, the application of the light stemmer yields the performance gains of 30% up to 50% in comparison with the base runs, which was found to be statistically significant (see section 6.2.2). Like many other major indigenous African languages, since Afaan Oromo has very rich and complex morphological structures, we have been expecting a significant improvement in the performances of OMEN-CLIR as a result of the application of the stemmer. Generally, we found the performances of our retrieval system very promising and encouraging, given the linguistic disparity of the language pair and the limited amount of translation resources we have used for developing and implementing OMEN-CLIR.

However, it is important to note that most of the performances of OMEN-CLIR are lower than the average precisions of the current state-of-the-art performances of CLIR systems, (which may range from 80% to 95% of the monolingual baseline). There were also many discrepancies in the performance levels of our individual query topics. As indicated in section 6.3, in certain instances, the performance level obtained by an individual query topic was much worse than the mean average performance level that was obtained for all of the queries. Incorrect transliteration of important proper names, mistranslations of phrasal terms and scientific words are found to be the major reasons for lower performance of OMEN-CLIR. In some cases, many important key words and search terms such as proper names, technical terms and acronyms found in Oromo query topics were not properly translated into English search queries due to the limited coverage of the bilingual dictionary. We have tried to address this problem by using a transliteration scheme adopted for this purpose, which modifies the out-of-vocabulary words and then matches their phonetic equivalents in the target language.

Moreover, we have discovered that some of the translation errors were resulted from the word-by-word translation and keeping all of those translations as equally candidate English search terms. As discussed in section 6.3 in detail, the translation of Afaan Oromo phrasal terms such as "*Intarneetiin Tajaajila Baankii*", (CLEF-2007 Oromo topic no. "424") and "*Turistoota Samii*", (CLEF-2007 Oromo topic no. "441"), were not able to capture the semantic meanings of the concepts in the original Oromo query topics. Consequently, some of our translated search queries were not able to rank the relevant documents on the top of the search outputs and thus reduced the overall performance of OMEN-CLIR. In general, lack of MWEs and NE recognition systems for handling Afaan Oromo have been found to be the major causes of low performances for some Oromo query topics. Hence, as indicated in section 6.3, the need for the development and application of Afaan Oromo MWEs and NE recognition system is critical for further improvement of OMEN-CLIR.

In summary, some of the major contributions of this study include pioneering the development of Oromo-English CLIR, designing an FST-based computational model for Afaan Oromo inflectional morphology and construction of a machine-readable bilingual dictionary that has been adopted as a main source of knowledge for query translation. In addition to building a general architecture of OMEN-CLIR that can work with minimal linguistic resources and translation tools, designing Afaan Oromo stemmer and evaluating its impacts on the performance of OMEN-CLIR are chief among the other contributions of this study. Apart from designing and building the first Oromo-English CLIR, for which most of the basic linguistic resources and translation tools that have been designed and developed from scratch, another major contribution of this study is assessing the performance of OMEN-CLIR at a well-recognized and international cross-language evaluation forum like the CLEF campaign. In one of our official retrieval experiments in which we have used our Afaan Oromo stemmer, our CLIR system has achieved an average mean precision (AMP) of 29.90%, which is about 67.95% of a monolingual baseline. Overall, we found the evaluation results very promising and encouraging, given the disparity of the languages involved and the limited amount linguistic resources employed for developing OMEN-CLIR.

7.2 Future Directions

One of the most pressing research issues in an increasingly networked and knowledge-based society is the need to provide universal and equitable access to online information resources and services. As noted by [7], if the vast wealth of information available on the Web is efficiently made accessible to all users, it has the power to transform a society in a profound way. In multilingual developing nations like Ethiopia, it is crucial to exploit the unprecedented opportunities presented by the rapid development the WWW. CLIR is a multidisciplinary field of study that stands at the intersection of several complementary research fields including MT, IR, computational linguistics, and natural language processing, which are increasingly becoming more challenging and interesting research areas. Nowadays, there is an increasing demand for the development of multilingual information access systems that can transcend linguistic and cultural barriers. However, meeting these requirements presents several challenges to information professionals and IR researchers. For instance, filling the gap in language resources and technologies, (i.e., bridging the linguistic digital divide), is one of the major challenges that hinders the development CLIA for African languages. Core linguistic resources and technologies including multilingual lexicons, parallel corpora, machine translation systems, MWEs and NE recognition systems must be investigated and established for resource-scarce languages. Our experience in building and evaluating OMEN-CLIR points not only to the need for establishment reliable multilingual resources and translation tools, but also the need for development of robust MWEs and NE recognition systems, which can significantly improve the effectiveness of the proposed cross-language retrieval system.

Besides the critical need for establishment core linguistic resources and translation tools, another major prerequisite for the development CLIA for African languages is highly qualified computational linguists and language technology developers. Africa’s academic programmes in Computational Linguistics and Language Technologies need to be further strengthened and promoted. In recent years, many research projects on language technology development have turned to crowdsourcing. We see such online collaborative models as viable approach for developing CLIA resources and systems for severely under-resourced African languages like Afaan Oromo. Since the concept of crowdsourcing arose from the evidence that some language resource development tasks could be handled by ordinary citizens and native speakers of a given language online, exploring and adopting such collaborative approach is crucial to accelerate the development of African languages technologies.

Currently, the performance of OMRN-CLIR largely depends on limited linguistic resources such as bilingual lexicon, list of stopwords and light stemming. Since we feel the priority should be given to the overall performance improvement of OMEM-CLIR, it is important to start with enhancing the coverage of the bilingual dictionary. Towards this end, a lot of effort has been made not only to correct many errors and inconsistencies in the bilingual dictionary entries, but also to incorporate additional vocabularies and technical terms from various online and offline lexical resources. The coverage of the bilingual dictionary should be more comprehensive and up-to-date in order to achieve better query translation and search results. Nevertheless, lexical entries provided in a dictionary are always limited. Since language is dynamic and evolving, there are new words being created from time to time. In other words, new technical terms, abbreviation, names of persons, organizations, and events may not be fully covered in general bilingual dictionaries. Hence, supplementing the coverage of the bilingual dictionary with reliable comparable and parallel corpora is very crucial to improve the performance of OMEN CLIR. A number of researchers have recently reported that CLIR tends to benefit from query expansion and relevance feedback techniques. We are planning to incorporate query expansion techniques such as blind relevance feedback into our retrieval experiments. Although our current study has been primarily focused on the development of Oromo-English CLIR, the general architecture of OMEN-CLIR can be easily extended and adopted for other resource-scarce African languages. As part of our future work, we would like to conduct similar studies for closely related Cushitic languages such as Somali and Sidama.

In summary, some of the research issues that we will consider in our short term plan include:

- To conduct more detailed errors analysis on the results of our retrieval experiments with special focus on individual query topics with very low level of performance;
- Enhancing and enriching the coverage of our bilingual dictionary with additional vocabulary entries including named entities, technical words and phrasal terms;
- To undertake more comprehensive evaluation and detailed analysis on the performance of Afaan Oromo stemmer;
- To conduct additional evaluation experiments to improve the performances of OMEN-CLIR further;
- To refine and improve the computational model for Afaan Oromo morphology;
- To implement and evaluate Afaan Oromo morphological analyser;
- To explore and incorporate comparable corpora into the major components of OMEN-CLIR.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, New York: Addison Wesley and ACM Press, 1999.

[2] J. Nie, Cross-Language Information Retrieval, San Rafael, California: Morgan & Claypool, 2010.

[3] M. Gasser, “Machine translation and the future of indigenous languages,” in *I Congreso Internacional de Lenguas y Literaturas Indoamericanas,* Temuco, Chile, 2006.

[4] D. Z. Osborn, African Languages in A Digital Age: Challenges and Opportunities for Indigenous Language Computing, Cape Town: HSRC Press, 2010.

[5] T. Adegbola, “Building capacities in human language technology for African languages,” in *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages,* Athens, Greece , 2009.

[6] D. W. Oard, “Serving users in many languages: Cross-language information retrieval for digital libraries,” *D-Lib Magazine,* December 1997.

[7] C. Peters, M. Braschler and P. Clough, Multilingual Information Retrieval: From Research to Practice, Berlin: Springer, 2012.

[8] R. Georg and U. Hans, “Multilingual Europe: Facts, Challenges, Opportunities,” in *META-NET Strategic Research Agenda for Multilingual Europe 2020*, Berlin., Springer, 2013, pp. 12-18.

[9] R. Lehtokangas, E. Airio and K. Jrvelin, “Transitive dictionary translation challenges direct dictionary translation in CLIR,” *Information Processing & Management,* vol. 40, no. 6, p. 973–988, 2004.

[10] J. Wang, “Matching Meaning for Cross-Language Information Retrieval,” (PhD Thesis), Graduate School of the University of Maryland, College Park, Maryland, 2005.

[11] F. C. Gey, N. Kando and C. Peters, “Cross-language information retrieval: The way ahead,” *Information Processing & Managment,* vol. 41, no. 3, p. 415–431, 2005.

[12] M. Braschler, G. Di Nunzio, J. Gonzalo, C. Peters and M. Sanderson, “From CLEF to TrebleCLEF: Promoting technology transfer for multilingual information retrieval,” in *Working Notes of the Second DELOS Conference on Digital Libraries*, Pisa, Italy, 2007.

[13] K. K. Tune, V. Varma and P. Pingali, “Evaluation of Oromo-English Cross-Language Information Retrieval,” in *Twentieth International Joint Conference on Artificial Intelligence: Workshop on Cross-Language Information Access (CLIA),* Hyderabad, India, 2007.

[14] K. P. Scannell, “The Cr´ubad´an Project: Corpus building for under-resourced languages,” in *In Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop,* Louvain-la-Neuve, Belgium, 2007.

[15] O. Vikas, “Multilingualism for cultural diversity and universal access in cyberspace: An Asian perspective,” in *UNESCO WSIS Thematic on Multilingual in*

Cyberspace, Bamako, 2005.

[16] M. Gasser, S. Hockema and M. Kane, "Information is power: intelligent tools for information access and evaluation," in *World Forum on Information Society,* Tunis, Tunisia, 2006.

[17] T. Hedlund, E. Airio, H. Keskustalo, K. Järvelin, A. Pirkola and R. Lehtokangas, "Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002," *Information Retrieval,* vol. 7, no. 1-2, p. 99–119, 2004.

[18] T. Helund, "Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation," (PhD Thesis), Department of Information Studies, University of Tampere, Tampere, Finland, 2003.

[19] A. R. Diekema, "Translation Events in Cross-Language Information Retrieval: Lexical Ambiguity, Lexical Holes, Vocabulary Mismatch, and Correct Translations," (PhD Thesis), School of Information Studies, Syracuse University, New York, 2003.

[20] D. Z. Osborn, "African languages and information and communication technologies: Literacy, access, and the future," in *Selected Proceedings of the 35th Annual Conference on African Linguistics: African Languages and Linguistics in Broad Perspectives*, Somerville, MA, 2006.

[21] B. Yimam, "The Phrase Structure of Ethiopian Oromo," (PhD Thesis), School of Oriental and African Studies, University of London, London, 1986.

[22] A. Nafa, "Long vowels in Afaan Oromo: A Generative Approach," (M.A. Thesis), Institute of Language Studies, Addis Ababa University, Addis Ababa, 1988.

[23] T. Korenius, J. Laurikkala, K. Järvelin and M. Juhola1, "Stemming and lemmatization in the clustering of Finnish text documents," in *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, New York, 2004.

[24] W. Kraaij, "Embedding web-based statistical translation models in cross-language information retrieval," *Computational Linguistics,* vol. 29, no. 3, p. 381–419, 2003.

[25] K. K. Tune and V. Varma, "Oromo-English Cross-Language Information Retrieval Experiments," in *Working Notes for the CLEF 2006 Workshop*, 2006.

[26] K. K. Tune, V. Varma and P. Pingali, "Pingali. Improving recall for Hindi, Telugu, Oromo to English CLIR," in *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers*, Berlin, Springer, 2008, p. 103–110.

[27] K. K. Tune, V. Varma and P. Pingali, "Hindi, Telugu, Oromo, English CLIR evaluation," in *Evaluation of Multilingual and Multimodal Information Retrieval, 7thWorkshop of the Cross-Language Evaluation Forum, Revised Selected Papers,* Berlin, Springer, 2007, p. 35–42.

[28] C. Peters and P. Sheridan, "Multilingual information access," in *Lectures on Information Retrieval, Third European Summer School, (ESSIR 2000)*, Varenna, Springer, 2001, p. 51–80.

[29] R. R. Korfhage, Information Storage and Retrieval, New York: Wiley Computer Pub., 1997.

[30] E. M. Elnahrawy, N. M. Ghanem and M. A. Youssef, “Cross-language information retrieval: Layout strategies for gloss translation,” in *Workshop on Evaluation of Interactive Cross-Language Information Retrieval*, College Park, Maryland, 2001.

[31] I. Maja, “Towards the human rights protection of minority languages in Africa,” Globalex, New York, 2008.

[32] G. Negash, “Globalization and the role of African languages for development,” in *Language Communities or Cultural Empire*, Institute of European Studies, University of California, Berkeley, 2005.

[33] A. Mroz, “From Gutenberg to ICT: The new lingua dxoctus - threat or opportunity?,” in *The Eight ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies, (The ETHICOMP Decade 1995-2005)*, Linköping, Sweden, 2005.

[34] D. Jurafsky and J. H. Martin, *Speech and language processing,* New Jersey: Prentice Hall, 2000.

[35] V. Berment, “Methodes pour Informatiser des Langues et des groups de Langues peu Dotees,” (PhD Thesis), L'université Joseph Fourier, Grenoble, 2004.

[36] W. D. Lewis and P. Yang, “Building MT for a severely under-resourced language: White Hmong,” in *Association for Machine Translation in the Americas*, San Diego, CA, 2012.

[37] J. Muhirwe, “Towards Human Language Technologies for Under-resourced languages,” in *Series in Computing and ICT Research*, Kampala, Fountain Publishers, 2007, pp. 123-128.

[38] A. Barbaresi, “Challenges in web corpus construction for low-resource languages in a post-BootCaT world,” in *6th Language & Technology Conference, Less Resourced Languages special track*, Poznan, Poland, 2013.

[39] Y. W. Bradshaw, J. Britz, T. Bothma and C. Bester, “Using Information Technology to Create Global Classrooms: Benefits and Ethical Dilemmas,” *International Review of Information Ethics (IRIE),* vol. 7, no. (09/2007) , pp. 1-9, 2007.

[40] E. C. Kachale, “Accessibility of ICT to speakers of indigenous African languages,” *Opening Societies through Advocacy,* vol. 2, no. 3, p. 88–94, 2008.

[41] J. J. Britz, “ To know or not to know: A moral reflection on information poverty. In Journal of Information Science,” *Journal of Information Science,* vol. 30, no. 3 , pp. 192-204, 2004.

[42] J. J. Britz, “Information poverty: The development of a global moral agenda,” in *International Symposium on the Transformation and Innovation of Library and Information Science,* Taipei, Taiwan, 2010.

[43] Deloitte, “Value of connectivity: Economic and social benefits of expanding internet access,” 2014. [Online]. Available: https://fbcdn-dragon-a.akamaihd.net/hphotos-ak-ash3/t39.2365/851546_1398036020459876_1878998841_n.pdf. [Accessed 5 April 2014].

[44] P. M. Musau, “Constraints on the acquisition planning of Indigenous African

Languages: The case of Kiswahili in Kenya," *Language, Culture and Curriculum,* vol. 12, no. 2, pp. 117-127, 1999.

[45] International Telecommunication Union (ITU), "World Summit on the Information Society: Declaration of Principles," in *The World Summit on the Information Society (WSIS)*, Geneva, 2003.

[46] International Telecommunication Union (ITU), "World Summit on the Information Society: Tunis Commotment," in *The World Summit on the Information Society (WSIS II)*, Tunis, 2005.

[47] United Nations Educational, Scientific and Cultural Organization (UNESCO), Building inclusive knowledge societies: A review of UNESCO's action in implementing the WSIS outcomes, Paris: UNESCO, 2014.

[48] United Nations Educational, Scientific and Cultural Organization (UNESCO), Towards knowledge societies: UNESCO world report, Paris: UNESCO, 2005.

[49] A. Large, "The new babel: Language barriers on the World Wide Web," *Journal of Universal Language*, vol. 3, no. 1, p. 77–95, 2002.

[50] Summer Institute of Linguistics (SIL International), Ethnologue: Languages of the world, Dallas: SIL International, 2005.

[51] Y. Lodhi, "The language situation in Africa today," *Nordic Journal of African Studies*, vol. 2, no. 1, p. 79–86, 2003..

[52] S. J. Shim, "Using Cross-Language Information Retrieval methods for bilingual search of the Web," in *Proceedings of the International Conference on Intelligent Agents, Web Technology and Internet Commerce*, Vienna, Austria, 2005.

[53] Deloitte , "Sub-Saharan Africa Mobile Observatory," 21 November 2012. [Online]. Available: http://www.gsma.com/spectrum/sub-saharan-africa-mobile-observatory-2012/. [Accessed 12 November 2013].

[54] G. T. Childs, An Introduction to African Languages, Amsterdam, the Netherlands: John Benjamin's Publishing Company, 2003.

[55] B. Gamback and G. Eriksson, *Natural language processing at the School of Information Studies for Africa (SISA),* Addis Ababa, Ethiopia, 2005.

[56] M. Porter , "An algorithm for suffix stripping," *Program,* vol. 14, no. 3, pp. 130-137., 1980.

[57] R. Schäler, "Enabling the gobal conversation in communities," in *W3C Workshop: Making the Multilingual Web Work*, Rome, 2013.

[58] M. M. Ali and H. Suleman, "Multilingual Querying," in *Proceedings of Arabic Language Technology International Conference (ALTIC),* Alexandria, Egypt, 2011.

[59] G. Peruginelli, "Accessing legal information across boundaries: A new challenge," *In International Journal of Legal Information,* vol. 37, no. 3, pp. 276-304, 2009.

[60] M. K. Chinnakotla, S. Ranadive, P. Bhattacharyya and O. P. Damani, "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007," in *Advances in Multilingual and Multimodal Information Retrieval,* Berlin, Springer, 2008, p. 103–110.

[61] I. Al-Sughaiyer and I. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *Journal of the American Society for Information Science*

and Technology, vol. 55, no. 3, p. 189–213, 2004.

[62] Gumii Qormaata Afaan Oromoo, *Caasluga Afaan Oromoo,* Finfinnee: Komishinii Aadaaf Turizmii Oromiyaa, 1995 E.C.

[63] H. Stroomer, Comparative Study of Southern Oromo Dialects in Kenya: Phonology, Morphology and Vocabulary, Hamburg: Burke, 1987.

[64] O. Streiter, . K. P. Scannell and M. Stuflesser, "Implementing NLP projects for noncentral languages: Instructions for funding bodies, strategies for developers," *Machine Translation* , vol. 20, no. 4, p. 267–289, 2006.

[65] E. Cosijn, H. Keskustalo, A. Pirkola, K. De Wet and K. Jrvelin, "Afrikaans-English cross-language information retrieval," in *Progress in Library and Information Science in Southern Africa, Proceedings of the third biennial DISSAnet Conference,* Pretoria, 2004.

[66] E. Cosijn, A. Pirkola, T. Bothma and J. Nel, "Cross-lingual information access in indigenous languages: A case study in Zulu.," in *Cross-Language Information Retrieval: A Research Roadmap, a Workshop at SIGIR'02: 22nd International Conference on Research and Development in Information Retrieval,* Tampere, Finland, 2002.

[67] E. Cosijn, A. Pirkola, T. Bothma and K. Järvelin, "Information access in indigenous languages: a case study in Zulu," *South African Journal of Libraries and Information Science,* vol. 68, no. 2, pp. 94-103, 2002.

[68] J. Lovins. , "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics,* vol. 11, no. 1-2, p. 97–111, 1968..

[69] V. Steinbiss, "Human language technologies for Europe," in *Work Commissioned by TC-STAR Project*, Trento, Italy, 2006.

[70] G. A. Sharma, G. B. Van Huyssteen and M. W. Pretorius, "A technology audit: Human language technologies research and development in South Africa," in *Proceedings of the Portland International Conference on Management of Engineering and Technology,* Portland, Oregon, 2011.

[71] G. De Pauw and G.-M. De Schryver, "African Language Technology: The Data-Driven Perspective," in *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics,* Bozen-Bolzano, Italy, 2008.

[72] R. Rodmant, "Why Learn About Language?," in *The 5 Minute Linguist*, London, Equinox, 2006, pp. 7-10.

[73] B. Sands, "Africa's Linguistic Diversity," *Language and Linguistics Compass,* vol. 3 , no. 2, p. 559–580, 2009.

[74] H. Eifring and R. Theil, Linguistics for Students of Asian and African Languages, Oslo, Norway: Institutt for østeuropeiske og orientalske studier, 2004.

[75] J. Good, "African languages and linguistic typology," in *A Handbook of Contemporary Linguistics,* Uyo, Nigeria, University of Uyo Department of Linguistics and Nigerian Languages, 2009, pp. 1-56.

[76] M. S. Dryer, "Relationship between the order of object and verb and the order of adposition and noun phrase," in *The world atlas of language structures*, Oxford, Oxford University Press, 20013, p. 386–389.

[77] J. Garland, “Morphological typology and the complexity of nominal morphology,” in *Santa Barbara Papers in Linguistics: Proceeding from the Workshop on Sinhala Linguistics*, Santa Barbara, CA, 2006.

[78] J. Zeller, *The syntax of African languages: A review,* Durban, South Africa: (Unpublished), 2011.

[79] M. S. Dryer, “Order of subject, object and verb,” in *The World Atlas of Language Structures Online*, Leipzig, Germany, Max Planck Institute for Evolutionary Anthropology, 2013.

[80] A. Pirkola, T. Hedlund , H. Keskustalo and K. Jarvelin, “Dictionary-based cross-language information retrieval: Problems, methods, and research findings,” *Information Retrieval,* vol. 4, no. 3-4, p. 209–230, 2001.

[81] M. Abusalah , M. Oakes and J. Tait , “Literature review of cross language information retrieval,” *World Academy Of Science, Engineering And Technology,* vol. 4, p. 175–177, 2005.

[82] J. Chen and Y. Bao, “Information access across languages on the Web: From search engines to digital libraries,” in *Proceedings of the American Society for Information Science and Technology*, Columbus, Ohio, 2008.

[83] D. W. Oard and A. R. Diekema, “Cross-language information retrieval,” *Annual Review of Information Science and Technology (ARIST),* vol. 33, pp. 223–256,, 1998.

[84] M. Diki-Kidiri, “How to include less resourced languages into the Internet,” in *Global Symposium on Promoting the Multilingual Internet*, Geneva, 2006.

[85] K. K. Tune, “An Assessment of the Existing Information Retrieval Practices and Facilities in the Institute of Ethiopian Studies Library with a View to Developing Mechanisms and Tools for Improvement,” (M.Sc. Thesis) Addis Ababa University, Addis Ababa, 1998.

[86] A. Alemu and A. Lars, “An Amharic stemmer: Reducing words to their citation forms,” in *Proceedings of the 5th Workshop on Important Unresolved Matters,* Prague, Czech Republic, 2007.

[87] A. Alemu and A. Lars, “Dictionary based Amharic-English information retrieval,” in *CLEF 2004 Working Notes,* 2004.

[88] A. Alemu and A. Lars, “Dictionary based Amharic-English information retrieval,” in *CLEF 2006 Working Notes.*, 2006.

[89] S. Adugna and A. Eisele, “English – Oromo machine translation: An experiment using a statistical approach,” in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, European Language Resources Association (ELRA), 2010, pp. 2196-2199.

[90] Central Statistical Agency Population Census Commission, “Summary and Statistical Report of the 2007 Population and Housing Census,” Central Statistical Agency, Addis Ababa, 2008.

[91] M. Abdulsamed, Seerlugaa Afaan Oromoo, Finfinnee: Caffee Oromiyaa, 1994.

[92] T. Gamta, Oromo-English Dictionary, Addis Ababa: Addis Ababa University Press, 1989.

[93] J. Kamps, C. Monz and M. De Rijke, “Combining evidence for cross-language information retrieval,” in *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum (CLEF 2002) Revised Papers*, vol. 2785, C. Peters, M. Braschler, J. Gonzalo and M. Kluck, Eds., Berlin, Springer, 2003, pp. 111-126.

[94] W. Kraaij and R. Pohlmann, “Different approaches to cross language cross language information retrieval,” in *Computational Linguistics in the Netherlands 2000,* Amsterdam, Rodopi, 2001, p. 97–111.

[95] D. Zhou, M. Truran, T. Brailsford, V. Wade and H. Ashman, “Translation techniques in cross-language information retrieval,” *ACM Computing Surveys (CSUR),* vol. 45, no. 1, 2012.

[96] A. Shakery and C. Zhai, “Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs,” *Information Retrieval* , vol. 16, no. 1, pp. 1-29, 2013 .

[97] J. Xu and W. B. Croft, “Corpus-based stemming using co-occurrence of word variants,” *ACM Transactions on Information Systems,* vol. 16, no. 1, p. 61–81, 1998.

[98] P. Clough and M. Sanderson, “Evaluating the Performance of Information Retrieval Systems Using Test Collections,” *Information Research,* vol. 18, no. 2, p. paper 582, 2013.

[99] C. Middleton and R. Baeza-Yates, “A Comparison of Open Source Search Engines,” 2007. [Online]. Available: http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf. [Accessed 23 April 2013].

[100] F. Zou, P. F. Wang, X. Deng and S. Han, “Evaluation of stop word lists in Chinese language,” in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[101] C. Fox, “A stop list for general text,” *SIGIR Forum,* vol. 24, no. 1-2, pp. 19-21, 1989.

[102] K. Kettunen, “Managing word Form variation of text retrieval in practice – why five character truncation takes it all?,” in *The Fifth International Conference Human Language Technologies: The Baltic Perspective,* Tartu, Estonia, 2012.

[103] L. S. Larkey, “Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis,” in *SIGIR’02 Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Tampere, Finald, 2002.

[104] P. Forner, R. Paredes, P. Rosso, B. Stein and H. Miller, “Information Access Evaluation Multilinguality, Multimodality, and Visualization,” in *4th International Conference of the CLEF Initiative Proceedings, CLEF-2013,* Valencia, Spain, 2013.

[105] M. Braschler and B. Ripplinger, “How effective is stemming and decompounding for German text retrieval,” *Information Retrieval,* vol. 7, no. 3-4, pp. 291316, , 2004.

[106] M. Aljlayl and F. Ophir, “On Arabic Search: Improving the Retrieval Effectiveness

via a Light Stemming Approach," in *ACM Eleventh Conference on Information and Knowledge Management*, New York, 2002, pp. 340--347.

[107] D. Hull, "Using Statistical Testing in the Evaluation of Retrieval Performance," in *Proceedings of the 16th ACM/SIGIR Conference*, New York, 1993, pp. 329-338.

[108] F. Gey, J. Karlgren and N. Kando, "Information access in a multilingual world: transitioning from research to real-world applications," *ACM SIGIR Forum*, vol. 43, no. 2, pp. 24-28, 2009.

[109] R. Goswami, B. Datta and S. K. De, "Linguistic diversity and information poverty in South Asia and Sub-Saharan Africa," *Universal Access in the Information Society*, vol. 3, no. 8, pp. 219-238, 2009.

Appendixes

Appendix A: Sample Afaan Oromo Stopwords

aga
achi
achuma
afoo
al
ala
akka
akkam
akkamii
akkamiitu
akkana
akkataa
akkataan
akkas
akkasitti
akkasumas
akkuma
akksumas
amma
ammaa
ammoo
ana
ani
asham
asitti
attam
ati
awu
bira
biraa
biratti
biroo
biroon
biroos
biyyam
booda
booddee
bukkee
cinaa
dhaa
dhaan
duuba
dura
duraa
eega
ennaa
eenuu
eenyu
eennu
eenyufaa
eenyuun
eennuun
eenyuuf
eennuuf
eenyufaa
eenyufaadhaan
eenyufaaf
eenyufaarraa
eenyufatti
eenyurra
eennurraa
eenyurraa
eennurraa
eenyurratti
eennuratti
eenyuuf
eenyuree
eennuree
eenyutti
eennutti
eenyuma
eessa
ega
eessatti
eessarraa
eessaaf
eessaan
erga
faa
faati
faatidhaa
faatiidhaa
fakkaatanitti
fakkaatti
fakkaatu
fakkaattu
fakkeenyaa

fakkeenyaaf
fedhetuu
fi
fk
fkf
fkn
fknf
fuuldura
fundura
fuullee
gaa
gad
gadi
gaditti
gahaa
gala
galan
galani
galuu
gama
gamam
gar
gara
garam
garamiin
garamitti
garana
garas
gararraa
gararree
garii
garjalee
garuu
godhe
gootee
gubbaa
haa
haga
hagam
hamma
hammam
hammamiif
hammamiin
hammamtu
hanga
hennaa
himantu
hin
hinjira
hinjiru
hinjirtu
hinjirtan
hinjiru
hoggaa
hunda
hunduma
idda
iddoo
if
illee
immoo
inni
irra
irraa
irratti
irrattii
isa
isaaf
isaan
isaan
isaanii
isaaniif
isaaniis
isaanniis
isaaniirratti
ishe
ishee
isee
ishii
ishiif
isiidhaa
ishiidhaa
isiin
ishiifaa
ishiin
isin
isiniif
isiniin
isiniis
itti
ittillee
ittiin
jala
jechuu
jedhu
jedhus
jira
jiraadha
jiraadhe
jiraanne
jiraate
iraattee

jiraatta
jiraatti
jiraanna
jiraatan
jiraattan
jiran
jiranii
jiranirra
jiranu
jirre
jirta
jirti
jirra
jirtan
jirtanu
jirtu
jiru
kaa
ka'e
ka'en
kam
kamfaa
kami
kamidha
kamifaadha
kamiin
kamiinu
kamiinuu
kamiif
kaminiyyuu
kamirraa
kamitti
kamttuu
kamitu

kamuma
kamuu
kamirrattu
kamiyyuu
kan
kana
kanaa
kanaan
kanaaf
kanaafuu
kanaafi
kanaafiis
kanarra
kanarratti
kanisaanii
kanishii
kankee
kankeenya
kankoo
kanneen
kanneeni
kanarraa
kee
keenya
keessan
keeysa
kiyya
kkf
koo
kun
qunnama
kuni
kunoo
kunneen

kanneeni
laata
maal
maalfaa
maali
maalif
maaliin
maalirraa
maalirratti
maaliree
maalittuu
maaltu
maaluma
maalumaaf
malee
mee
meeqa
meeqaan
meeqaaf
meeqarraa
meeqatti
meeqatu
meerre
meerreree
miti
moo
na
naaf
nan
ni
nuhi
nui
nuu
nuuf

nuun
nuti
nuyi
obboo
odoo
of
ofii
ofirraa
ofiif
ofiin
oftiin
ofitti
ofuma
ofumaa
ofumaaf
ofumaanirrattii
ofumatti
ol
oli
olkaa'i
osoo
otoo
san
sani
sana
saanii
sanaa
sanaan
sanaas
saniin
si
sii
siif
silaa

sirri
sirritti
sirrumatti
suma
sun
suni
suniin
sunneen
ta'a
takka
takkaan
ta'an
ta'anii
ta'aniifi
ta'ani
ta'e
ta'etti
ta’uun
taane
ta’anii
ta’aniif
taa'ee
taa'een
taa'eetti
taata
taati
taana
taatan
ta'u
ta'uu
ta'uun
ta'uuf
taatu
taanu

taatan
tansaa
tam
tamiif
tamiin
tamuma
tan
tana
tanisii
tanneen
tanneeni
tantee
tanteenya
tasaanii
ta'ullee
tee
teenya
teessan
teeysa
tiyya
tuqa
tuqan
tuquu
tokko
tokkoo
tokkootti
tokkos
too
tun
tuni
tunoo
turan
ture
turee

turre
turte
turtan
utuu
waan
wa'ee
waa'ee
wajjin
wal
walakkaa
walii
waliif
waliin
waliis
walitti
walirratti
walirra
walirraa
Waljidduu
walleenuu
walqabatan
walqabatanis
walqunnaman
walumaaf
walumaan
wahii
waanta
waantoota
waa’ee
waayee
wojiin
wolbira
xiqqaa
xiqqaatuu
yemmuu
yennaa
ykn
yoo
yookaan
yookiin
yoom
yoomiif
yomiree
yoomuma
yoomittuu
yoos

Appendix B: Sample Oromo Topics

<topics>
<top lang="OM">
<top lang="OM">
<num>401</num>
<title> Olka'iinsa Gatii Yuroodhaa </title>
<desc>Erga Yuroon hojiirra oolee asitti dokumantoota waa'ee olka'iinsa gatii ibsan barbaadi.</desc>
<narr> Dokumantii biyya maallaqa Awurooppaa fayyadamuu hojiirra oolche irratti waa'ee odeeffannoo olkaa'insa gatiidhaa qaban kamiyyu fudhatama qaba.</narr>
</top>
<top lang="OM">
<num>402</num>
<title> Maddawwan Humna of Bakka Buusu </title>

<desc>Dokumantoota waa'ee maddisiisa humnaatiif hojiirra oolma maddawwan human of bakka buusan gabaasan barbaadi.</desc>

<narr>Dokumantoonni dhimma kanaan walitti dhiyeenya qaban waa'ee ittifayyadama humna maddawwan human ofbakka buusanii yookin of haaromsanii kana akka humna baawoomaasii (humna tortora gataarraa uummamuu), bishaan, aduu, hurka lafa keessaa bahuu/ji'ootarmaalii, ykn bubbee/qilleensa irratti ni gabaasu. Odeeffannoon waa'ee konkolattoota boba'aa qusatanii kana waliin hin deeman. </narr>

</top>

<top lang="OM">

<num>403</num>

<title> Akka Qondaala Poolisiitti Taatessuu </title>

<desc>Dokumantoota waa'ee taatoota gaheewwan poolisii fiilmii ykn televizyiinii keessatti taphatanii odeeffannoo kenaan barbaadi.</desc>

<narr>Dokumantoonni dhimma kana walitti dhiyeenya qaban waa'ee mata duree fiilmichaa, sagantichi ittifufiinsaan kan darbuu ta'uu fi dhiisu isaa, akkasumas maqaa taatichaa ykn taatittii akka poolisii dhiiraa ykn poliisii dubartiitti ta'uun taphate ykn taphatte ni ibsu.</narr>

</top>

<num>404</num>

<title> Eeginsa Nageenyaa Yaa'ii Dhaaba Waliigaltee Atilaantika Kaabaa (NATO) </title>

<desc> Gabaasota tarkaanfiiwwan nageenyaa bakkee yaa'iin Dhaaba Waliigaltee Atilaantika Kaabaa ykn NATO gaggeeffame kaminittuu fudhataman ibsan barbaadi.</desc>

<narr> Adunyaa kanarratti of eeggannoowwan nageenyaa bakkee yaa'iwwan Dhaaba Waliigaltee Atilaantika Kaabaan ykn NATO gaggeeffamanitti godhaman ilaalchisee odeeffannoo ni barbaanna. Yaa'ii kanaan haala walqabateen odeeffannoon waa'ee walitti bu'iinsa, hiriira nagayaa, fi gochoota mormii gaggeeffamanii hin ta'an ykn firooma hin qaban.</narr>

</top>

<top lang="OM">

<num>405</num>

<title> Xiixaa ykn Boyidaa Daa'imummaa </title>

<desc> Haalawwan akkamiitu daa'imman xiixaan akka qabaman isaan gochuu danda'aa?</desc>

<narr> Dokumantoonni barreeffamoonni dhimma kanaan walitti dhiyeenya qaban karaa xiixaan daa'ima gamisaanis ta'u guutuman guutuun waantoota ni qabsiisu jedhamanii amanaman ni tuqu.</narr>

</top>

<top lang="OM">

<num>406</num>

<title> 'Kaartuunota Animeetidii'</title>

<desc> Dokumantoota waa'ee 'kaartuunota animeetiidii' badhaasaaf yaadamanii barbaadi.</desc>

<narr> Dokumantoonni dhimma kanaan walitti dhiyeenya qaban maqaa 'Kartuunii Animeetiidii' feestivaala fiilmii irratti badhaasaaf yaaduun dhiyaatee ykn immoo feestivaala fiilmiin kamiiniyyu alatti dhunfaan badhaasaaf yaadame tokko ni tuquu.</narr>

</top>

<top lang="OM">

<num>407</num>

<title> Muummicha Ministeera Awustiraaliyaa</title>

<desc> Bara 2002 keessa Muummicha Ministeera Awustiraaliyaa eenyu dha? <desc>

<narr> Dokumantoonni dhimma kanaan walitti dhiyeenya qaban maqaa Muummicha Ministeera Awustiraaliyaa bara 2002 keessa aangoo irra turee tuquu qabu.</narr>

 </top>

<top lang="OM">

<num>408</num>

<title> Sanyii Ilma Namaa Kilooniingiin Uumuu </title>

<desc> Nama duraa kilooniingiin uumuu irratti mormiiwwan jiran dabalatee ilma namaa kilooniingiin uumuu irratti odeeffannoowwan jiran kamiyyuu ni barbaanna.<desc>

<narr> Odeeffannoo waa'ee kilooniingiin jiisa ykn ulfa ilma namaa fooyyeessuu fudhatama qaba. Dhaabbileen kilooniingii gaggeessan, malawwan kilooniingii, yaalii fayyaaf abdii inni qabu fi kilooniingiin walqabatee rakkoolee jiran to'achuuf seerawwan bahan hundumtuu fudhatama qabu. Dokumantoonni rakkoolee naamusaa haasawan kana wajjin hin deeman.</narr>

</top>

</topics>