Chapter 48
Building CLIA for Resource-Scarce African Languages:
A Case Study on Oromo—English CLIR

Kula Kekeba Tune
International Institute of Information Technology India

Vasudeva Varma
International Institute of Information Technology India

ABSTRACT

Since most of the existing major search engines and commercial Information Retrieval (IR) systems are primarily designed for well-resourced European and Asian languages, they have paid little attention to the development of Cross-Language Information Access (CLIA) technologies for resource-scarce African languages. This paper presents the authors' experience in building CLIA for indigenous African languages, with a special focus on the development and evaluation of Oromo-English-CLIR. The authors have adopted a knowledge-based query translation approach to design and implement their initial Oromo-English CLIR (OMEN-CLIR). Apart from designing and building the first OMEN-CLIR from scratch, another major contribution of this study is assessing the performance of the proposed retrieval system at one of the well-recognized international Cross-Language Evaluation Forums like the CLEF campaign. The overall performance of OMEN-CLIR was found to be very promising and encouraging, given the limited amount of linguistic resources available for severely under-resourced African languages like Afaan Oromo.

INTRODUCTION

As we move towards an increasingly globalized and knowledge-based economy, the ability to instantly access and share relevant information (Baeza-Yates & Ribeiro-Neto, 1999; Gey, Kando, & Peters, 2005; Nie, 2010) beyond language and cultural boundaries has become more and more crucial. The World Wide Web (WWW) contains massive volumes of multilingual and multimedia information resources that can be explored and exploited to address critical social and economic problems. Unfortunately, in

DOI: 10.4018/978-1-5225-5191-1.ch048

developing and culturally diverse regions like Africa and Asia, the accessibility and usability of online resources are severely constrained by formidable obstacles and challenges such as language barriers, linguistic digital divide and lack of robust CLIA systems (Adegbola, 2009; Gasser, 2006; Varma, Tune, & Pingali, 2007). As pointed out by (Georg & Hans, 2013; Oard & Diekema, 1998; Peters, Braschler, & Clough, 2012), language barriers and linguistic digital divide have continued to threaten and undermine the potential of the Internet to deliver universal and equitable access to online information resources and services. This is especially true in highly multicultural developing nations like Ethiopia and India.

Broadly speaking, *language barriers* can be defined as linguistic and cultural factors that impede the free flow of information across language boundaries. In this article, the term *language barriers* is more specifically used to describe linguistic and cultural obstacles that discourage or prevent users from seeking and sharing important information across different languages and cultures. Even though the term *linguistic digital divide* is closely associated with language barriers, it is often used to describe the disparity in technological development between different languages (Gasser, 2006; Scannell, 2007). While the term *digital divide* is generally used to describe the gap in accessing and using computing devices among various social groups, the term *linguistic digital divide* is more specifically used to describe the relative advantages of certain languages (or language communities) over the others with respect to modern language resources and information access technologies.

Since most of the existing commercial search engines and Information Retrieval (IR) systems have primarily focused on well-resourced European and Asian languages, they have not paid adequate attention to supporting under-resourced African languages (Adegbola, 2009; Gey, Kando, & Peters, 2005; Osborn, 2010; Pingali, Tune, & Varma, 2008). The need for exploring and developing multilingual information access technologies that permit African communities to search and discover information beyond linguistic and cultural barriers has, therefore, become more urgent today than ever before. In this regard, much attention has been paid to the development of *Cross-Language Information Retrieval* (CLIR), which is mainly concerned with searching and discovering information beyond language and cultural boundaries (Hedlund, et al., 2004; Nie, 2010). The main purpose of CLIR is to identify documents written in one or more language(s) in response to a query expressed in a different language (Nie, 2010; Peters, Braschler, & Clough, 2012). On the other hand, CLIA deals with much more general and broader issues. CLIA encompasses not only the academic domain of cross-language search or CLIR, but also many aspects of natural language processing and understanding, including text encoding, digitization, content analysis and visualization (Peters, Braschler, & Clough, 2012). In this paper, we use the term CLIA in its narrower sense to refer to the processes of querying, accessing and retrieving information across different languages.

Over the last two decades, CLIR has constantly evolved and emerged as one of the most challenging and dynamic subfields of IR (Chen & Bao, 2008; Gey, Karlgren, & Kando, 2009; Nie, 2010). A number of evaluation studies on CLIA systems in general and on CLIR in particular have been reported and discussed at well-recognized international conferences and workshops such as Text REtrieval Conference (TREC) and Cross-Language Evaluation Forum (CLEF) (Di Nunzio, Ferro, Mandl, & Peters, 2007; Peters, Braschler, & Clough, 2012). However, since most of the earlier studies were highly concentrated on a handful of resource-rich European and Asian languages, the need for exploring and building CLIR for severely under-resourced African languages has been left unaddressed for quite a long time (Kachale, 2008; Osborn, 2010; Tune, Varma, & Pingali, 2007). As a step towards filling in this research gap, this study seeks to explore and develop an experimental CLIR between one of the most resource-scarce