Chapter 2
Multilingual –Monolingual Information Retrieval

2.1 Introduction

The increasing volume of information available globally through the Internet places high demands on information systems that can handle multilingual documents in a unified manner. Also, the languages used for Web documents are expanded from English to various languages. However, there are many unsolved problems in order to realize an information system which can handle such multilingual documents in a unified manner.

2.2 Multilingual Information Access

Multilingual information access can be defined as the functionality allowing anyone to find information that is expressed in any language. Oard ([3], 1997) identifies it as a selection of useful documents from collections that may contain several languages. Another formulation refers to the capability for users to retrieve documents written in a language different from a query language (Lee, Kageura and Choi, 2004).

These requirements can be clarified by stating that in multilingual access environment information is searched, retrieved and presented effectively, without constraints due to the different languages and scripts used in documents and their metadata. This implies that in creating multilingual access services, both users’ native language and the multiplicity and richness of world-wide languages are to be accommodated, so that users can put queries expressed in any one language and retrieve information resources independently of the language of documents and indexing.

The requirement for multilingual access is based on the recognition that cultural diversity is vital to the maintenance of society and that languages are a strong element of the different cultural traditions. The role of the information professionals in this context is crucial, as clearly stated by Clews (1994), who points out that the naturally multilingual and multicultural position of libraries in
society means that they should lead the way in developing systems and services to foster cross-language retrieval. As the diversity of the world’s languages and cultures generates a wealth of knowledge and ideas, it is essential to develop research studies and tools to preserve and successfully use the variety of resources produced.

With the increasing moves towards an integrated Europe and the increasingly multicultural nature of modern society and its globalization, facilitated by the development of digital information and telecommunications networks, the need for multilingual information access has become more and more pressing and the issues connected with cross-language retrieval have increased in importance. Language barriers are critical to the effectiveness of resource sharing and world-wide common access, and their emergence as a problem is to be connected with the growing number of information databases now available on networks (Hudon, 1997; Oard, 1997; Michos, Stamatos and Fakotakis, 1999). Landry (2003) goes beyond, his main focus is on multilingual subject access, observing how users’ needs have expanded as the result of the Web, that has made OPACs available beyond local use, but he points out also that new technologies have opened up various possibilities and solutions to multilingualism.

This review concerns multilingual text retrieval, while image and speech retrieval, now coming into the scene (Gey, Kando and Peters, 2002) are very marginally addressed. The focus is on multilingual access through information systems, not on multilinguality in general, thus leaving aside functionality which is nevertheless the commitment of libraries, such as collection development in multi languages and reference assistance services to multilingual populations.

Multilingual access is a complex and multifaceted topic, embracing technical, functional and strategic issues which have been (and still are) under discussion in the information specialist community for many years. What is needed is functionality like thorough and proper handling of characters (their presentation, arrangement, transfer), putting queries in a preferred language and
script, retrieving resources irrespective of the language used in searching and indexing, having world-wide communication no matter what the language.

The extensive literature produced contains contributions encompassing these main themes: functional requirements of multilingual access, technical issues concerning character set standards, multiscripts manipulation, and various approaches to cross-language retrieval. These themes are often presented with a description of related projects, experiments and research studies.

Another important aspect of multilingual access concerns strategic and management issues refer to the need for general consensus and recommendations to achieve multilingual functionality. Emphasis is put on the need for a paradigm shift in information professional community to overcome language barriers in information retrieval. These themes are not as popular in the literature. As those concerning technical and functional aspects, are specifically addressed by some authors such as Borgman (1997) and Nardi-Ford (1998). They point out that the problem of English language dominance, initially encountered in the development of the character encoding systems, affects developments in CLIR and attention must expand beyond technical aspects. As the richness of human communication is extremely hard to tackle, the library world should become more aware of linguistic and strategic issues and be exposed more and more to the rest of the world. Similar opinions, yet with more emphasis on digital library and Internet related technologies’ issues, are expressed by Peters and Picchi (1997) who claim that, despite the technological developments which have occurred in the 1990s, in general digital library research and development until recently have somehow neglected the issues of multilingual presentation and access, and have concentrated developments and applications on monolingual environments, where English language employment has taken the lead. Development of tools and applications embracing different languages, including Asian ones, has nevertheless progressed in these last few years [15].
2.3 Multilingual Information Processing on the Internet

The increasing volume of information available globally through the Internet places high demands on information systems that can handle multilingual documents in a unified manner. Also, the languages used for Web documents are expanded from English to various languages. However, there are many unsolved problems in order to realize an information system which can handle such multilingual documents in a unified manner.

From the user’s point of view, three most fundamental text processing functions for the general use of the World Wide Web are display, input, and retrieval of the text. However, for languages such as Japanese, Chinese, Korean, Russian and Indian character fonts and input methods those are necessary for displaying and inputting texts, are not always installed on the client side.

From the system’s point of view, one of the most troublesome problems is that, many Web documents do not have Meta information of the character coding system and the language used for the document itself, although character coding systems used for Web documents vary according to the language. It may result in troubles such as incorrect display on Web browsers, and inaccurate indexing on Web search engines.

Also, other text processing applications such as categorization, summarization, and machine translation are dependent on knowing the language of the text to be processed. Moreover, there might be some cases where the user wants to retrieve documents in unfamiliar languages, especially for cases where information written in a language other than the user’s native language is rich. The needs for retrieving such information must not be small. Consequently, research on cross-language information retrieval (CLIR), which is a technique to retrieve documents written in one language using a query written in another language, are being paid much attention. However, it is difficult to achieve adequate retrieval effectiveness for Web documents in diverse languages and domains.
In this chapter, we introduce some basic information and current issues that are related to multilingual information processing on the Internet, with particular emphasis on the Web.

Table 2.1 shows the layers of multilingual information processing on the Web. The 1st layer is character coding system, which defines the characters sets and their encodings to be used in the upper layers. It can be further divided into two components; character encoding scheme and character set. These components will be described in detail in the next section. The 2nd layer is communication protocol, which defines how to transmit documents through a communication network, typically the Internet. HTTP (Hyper Text Transfer Protocol) [16] is an Internet protocol for communication between user agents (e.g. Web browsers) and Web servers. It has some features related to multilingual information processing, such as indicating the character encoding scheme of a page and indicating the language(s) of the specific bounds of a text, etc. MIME (Multipurpose Internet Mail Extensions) [17] is primarily defined for electronic mail messages. However, some features, especially the Content-Type header, are also used in HTTP.

The charset attribute of the Content-Type header, which will be described later in this chapter, is one of the most important features for multilingual information processing on the Internet and the Web.

The 3rd layer is text format, which defines the structure of a document. HTML (Hyper Text Markup Language) [18] is a fundamental text format for the Web. As described later in this chapter, it involves many features that are related to multilingual information processing. The 4th layer is user interface, which is typically a Web browser. Although a Web browser is an application in the sense of operating systems, it provides a user interface for Web applications that run on a browser. It also involves many features related to multilingual information processing, such as display and input. The 5th layer on the top is Web application, which runs on a Web browser. Typical Web applications include search engines, digital libraries, electronic commerce sites, etc. Since the Web itself is multilingual, every Web applications that manage Web documents, such Web search engines, must handle multilingual documents to some extent [19].
Table 2.1: IR Types: Cross Language IR and Monolingual IR

<table>
<thead>
<tr>
<th>Layer</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web application</td>
<td>Search engine, digital library, etc</td>
</tr>
<tr>
<td>user interface</td>
<td>Web browser</td>
</tr>
<tr>
<td>text format</td>
<td>HTML, XML, etc</td>
</tr>
<tr>
<td>communication protocol</td>
<td>HTTP, MIME, etc</td>
</tr>
<tr>
<td>character coding system</td>
<td>character encoding scheme UTF-8, ISO-2022, etc</td>
</tr>
</tbody>
</table>

2.3.1 Cross-language information retrieval

Defined as the retrieval of documents in another language than the language of the request. The language of the request is the source language and the language of the documents is the target language.

The term "cross-language information retrieval" has many synonyms, of which the following are perhaps the most frequent: cross-lingual information retrieval, translingual information retrieval, multilingual information retrieval. The term "multilingual information retrieval" refers to CLIR in general, but it also has a specific meaning of cross-language information retrieval where a document collection is multilingual.

The vast increase of multilingual content both on the Internet and corporate intranets has created the need for information access across languages and cultures. While a large proportion of users of information retrieval systems may possess varying levels of multilingual skills that enable them to input queries and read and understand documents in more than one language, there is often demand for interfaces that allow the input of queries in the languages the users know best and feel most comfortable with. CLIR aims to overcome the cross-lingual access problem by enabling the users to retrieve documents written in one language (often called the target language) based on queries typed in another (often called the source or query language). [20]
There are two types of translations namely Query translation and Document Translation. In Query Translation, the given query will be converted from Native language to English and will search the database to get the documents in English. Later the retrieved documents in English language can be converted to Native language.

In Document Translation, all the documents are translated from English to Native language. It allows the user to ask query in Native language and now the searching will take place to obtain the resultant documents in Native language. Among the two, the former is easier compared with later, because of the size of translation. The efficiency of the query translation depends on the best translation words and weight for the given query. But, the drawback with Query Translation is the given query normally will be short and hence ambiguity problem may arise. Since, Document Translation is not feasible, in most of the research works, Query Translation will be carried out instead of Document Translation. [21]

Cross-language information retrieval is based on translation – either queries are translated into the document language(s), or document(s) are translated into the query language. The latter alternative would be comfortable for the user, but it is expensive and hard to implement. The query translation approach is more common in CLIR, and it is applied in the present research as well. There are three main approaches in CLIR: a dictionary based approach, a corpus based approach and a machine translation based approach (Gachot & al. 2000).

2.3.1.1 Corpus-based CLIR

The corpus-based approach utilizes parallel or comparable corpora. The parallel corpora consist of a collection of pairs of documents in two languages which are translations of each other. Document alignment (sentence alignment, segment alignment, word alignment), which means finding relations between a pair of parallel documents, is a crucial part of the corpus-based approach. (Yang & Kar Li 2004.) There are two main approaches for sentence alignment: length-based and text-based alignment. The former approach is based on the total number of
words or characters in a sentence, while the latter utilizes lexical information of sentences. Sentence alignment is based on the assumption of one-to-one translation of sentences. If the number of sentences differs between parallel documents, it is possible to perform segment alignment before sentence alignment. Segment alignment takes into account insertion and deletion of paragraphs or sentences. Word alignment can be performed in sentence-aligned corpora. (Fung & McKeown 1997.)

2.3.1.2 Machine translation-based CLIR

Machine translation (MT) systems analyze the source text, including morphological, syntactic and semantic analysis utilizing special lexicons. The aim of machine translation is to translate complete sentences, and it is the only translation approach applicable for document translation. MT systems return only one translation variant for a word, which may cause loss of recall in retrieval (Yamabana & al. 2000). In addition, MT-based query translation may not produce very good results with short source queries which are typically not complete sentences and thus do not provide sufficient contextual information for translation (Chen & Gey 2004; Kishida 2005).

Despite the possible drawbacks mentioned above, MT-based query translation has performed quite well in IR tests, when the MT system has been of good quality, and source queries have been complete descriptions of information needs, e.g. TREC topics (see Oard 1998; Rosemblat & al. 2003; Huang & al. 2007). On the other hand, the performance of a poorer MT system can be boosted by combining other methods with translation, for example pseudo relevance feedback. It is also possible to combine translations of two or more MT systems in order to achieve a better query. (See Jones & Lam-Adesina 2002; Chen & Gey 2004.)

Document translation would be beneficial for users of a retrieval system, but translating a large document collection into numerous languages is exorbitant.
Fujii and Ishikawa proposed in 2000 a lighter version of document translation: only retrieved documents were translated.

### 2.3.1.3 Dictionary-based CLIR

The dictionary-based approach relies on standard machine-readable bi- or multilingual dictionaries. In dictionary-based CLIR, each query word is translated into the target language. The translation process produces none, one or more translation equivalents for each source word. (Hedlund 2003, 26-27.)

Because all translation variants are included, there is no fear of losing the right one (supposing that the dictionary is good enough), which might happen in machine translation. There is even the possibility that translation acts like a query expansion, because translation dictionaries often include synonyms. On the other hand, there is also a possibility of retrieving noise in the case of an ambiguous source word. The dictionary-based approach is the most common CLIR approach, because translation dictionaries are often relatively cheap and easy to use. [22]

### 2.3.2 Monolingual Information Retrieval System

Refers to the Information Retrieval system that can identify the relevant documents in the same language as the query was expressed.

While information retrieval (IR) has been an active field of research for decades, for much of its history it has had a very strong bias towards English as the language of choice for research and evaluation purposes. Whatever they may have been, over the years, many of the motivations for an almost exclusive focus on English as the language of choice in IR have lost their validity. The Internet is no longer monolingual, and non-English content is growing rapidly. Today, less than a third of all domain names is registered in the US, and by 2005 two-thirds of all Internet users will be non-English speaking. Multilingual information access has become a key issue. The availability of cross-language retrieval systems that match information needs in one language against documents in multiple
languages is recognized as a major contributing factor in the global sharing of information.

Multilingual IR implies a good understanding of the issues involved in monolingual retrieval. And there are other important factors that motivate monolingual IR system development. Even in relatively multilingual countries, users continue to feel the need to access information and services in their native languages. For small languages, the costs of developing and maintaining a language technology infrastructure are relatively high. But languages with inferior computational tools are bound to suffer in an increasingly global society, for both cultural and economic reasons. What are the issues involved in monolingual retrieval other than English? One common opinion is that the basic IR techniques are language-independent; only the auxiliary techniques, such as stop-wordlists, stemmers, lemmatizes, and other morphological normalization tools need to be language dependent (Harman, 1995a). But different languages present different problems. Methods that may be effective for certain languages may not be so for others; issues to be addressed include word order, morphology, diacritic characters, languages variants, etc [23].

2.4 Information Retrieval with Asian Languages

Asia is the largest and the most culturally and linguistically diverse continent. It covers 39 million square kilometers, about 60% of land area of the world, and has an estimated 3.8 billion population, which is approximately 60% of the world’s population. There are more than 50 countries and roughly 2200 languages spoken in Asia. Being the largest, most populous and most diverse, the challenge of development of Asian community is equally important, urgent and formidable. Utilization of Information and Communication Technology (ICT) to store, process and communicate information promises an effective and efficient remedy to socio-economic problems of poverty, health, education, gender parity, governance, etc. across this continent. This technology is increasingly being leveraged in the developed and developing countries across the world, and is bound to play significant role in Asia’s future.
Most of Asia still lags in effectively gaining the promised benefits of ICT. As a measure, Asia has only 34.5% of total Internet users in the world. 90% of these are in seven Asian countries. There are a variety of reasons why Asia is still behind in leveraging ICTs. One of the key factors has been the limited ICT infrastructure. However, significant investment has been made over past decade to improve this infrastructure in Asia. This has had significant impact. As infrastructure has improved and information has started to flow, it has increasingly been realized that the information is not usable unless it is generated or converted in languages that Asian populations can understand. About 10-15% of Asians can communicate in non-Asian languages, and only 11% of content on the Internet is available in Asian languages, most of which is in Chinese, Japanese and Korean. This indicates a significant barrier for Asians to access information, and therefore to synthesize this information for their development.

The solution is to empower Asian people to generate and access culturally relevant information content. But, before the problem of content can be addressed, it is an essential precursor to enable ICTs in Asian languages. Developing ICT “software framework,” including standards, terminology, utilities and applications, to enable information processing in local language is called localization. Clearly, the foremost task is to develop this software framework for Asian languages. Once ICTs are enabled in local languages, they can be more effectively used towards generating and accessing the much needed local language content. Unfortunately, large population in Asia is also deprived from information due to high illiteracy.

However, with today’s technology it is also possible to overcome this barrier by employing more innovative forms of ICT interface for accessing and generating information. This includes speech interface, visual interface using touch-screens, and usage of increasing pervasive mobile technology. After basic local language computing support has been achieved, the second step is to provide these higher-end user-centric tools which catalyze generation and access of content and overcome illiteracy and similar barriers. Advanced speech and language processing applications like Machine Translation, Text-to-Speech, and
2.4.1 Background: The Challenge of Asian Language Processing

Asian language processing presents formidable challenges to achieving multilingualism and multiculturalism in our society. One of the first and most obvious challenges is the multitude and diversity of languages: more than 2,000 languages are listed as languages in Asia by Ethnologue (Gordon, 2005), representing four major language families: Austronesian, Trans-New Guinea, Indo-European, and Sino-Tibetan. The challenge is made more formidable by the fact that as a whole, Asian languages range from the language with most speakers in the world (Mandarin Chinese, close to 900 million native speakers) to the more than 70 nearly extinct languages (e.g., Pazeh in Taiwan, one speaker). As a result, there are vast differences in the level of language processing capability and the number of sharable resources available for individual languages. Major Asian languages such as Mandarin Chinese, Hindi, Japanese, Korean, and Thai have benefited from several years of intense language processing research, and fast-developing languages (e.g., Filipino, Urdu, and Vietnamese) are gaining ground. However, for many near extinct languages, research and resources are scarce, and computerization represents the last resort for preservation after extinction. A comprehensive overview of the current state of Asian language processing must necessarily address the range of issues that arise due to the diversity of Asian languages and must reflect the vastly different state-of-the-art for specific languages. Therefore, special issues on Asian language technology have been divided into two parts. The first is a double issue entitled Asian Language Processing: State of the Art Resources and Processing, which focuses on state-of-the-art research issues given the diversity of Asian languages. Although the majority of papers in this double issue deal with major languages and familiar topics, such as spell-checking and tree-banking,

They are distinguished by the innovations and adaptations motivated by the need to account for the linguistic characteristics of their target languages. For
instance, Dasgupta and Ng’s morphological processing of Bengali has an innovative way to deal with multiple stems while Ohno et al.’s parsing of monologues makes crucial use of bunsetsu2 and utterance-final particles, two important characteristics of Japanese. A subsequent issue entitled New Frontiers in Asian Language Resources will focus on both under-computerized languages and new research issues, such as the processing of non-standard language found on the web. Overall, these special issues on Asian language processing assess the state-of-the-art for more than thirteen languages from six of the eight major Asian language families. As such, they provide a snapshot of the state of Asian language processing as well as an indication of the research and development issues that pose a major challenge to the accommodation of Asian languages in the future.

### 2.4.2 Language Processing in Asia

Research on Asian language technology has thrived in the past few years. The Asian Language Resources Workshops, initiated in 2001, have had over sixty papers presented in five workshops so far (http://www.cl.cs.titech.ac.jp/alr/). Interest in Asian language processing among researchers throughout the world was made evident in a panel entitled Challenges in NLP: Some New Perspectives from the East at the COLING/ACL 2006 joint conference. At the same conference, fifteen papers were accepted in the Asian language track, while many other accepted papers also dealt with processing Asian languages. The growing literature on Asian language processing attests to the robustness of current paradigms. For instance, corpus-based stochastic models have been widely adopted in processing of various Asian languages with results comparable to that of European languages. Studies on less computerized languages in Asia, however, do not have the luxury of simple adaptation of accepted paradigms and benchmarks. They are burdened by the dual expectations of infrastructure building and language engineering applications. On one hand, early stages of computerization mean that many types of language resources must be built from scratch. On the other hand, the maturing field of computational linguistics expects
attested and quantifiable results not tenable without substantial language resources. It is remarkable that this delicate balancing act has been performed successfully, as attested by many papers appearing in this and the subsequent issues that deal with Bengali, Filipina, Hindi, Marathi, Thai, Urdu, and Vietnamese, among others. A particularly striking example of how infrastructure building can go hand in hand with technological innovation is Collier et al.’s work on multilingual medical information extraction for Asian languages. Japanese scholars were the pioneers in Asian language processing. The Information Processing Society of Japan (IPSJ) was formed in 1960 with a significant number of members interested in Machine Translation and related areas. Natural language processing (NLP) activities in Japan were extensive in the 1980’s, starting with the first international conference on computational linguistics held in Asia: the 1980 Tokyo COLING. In 1982, the Fifth Generation Computer Project contained significant segments on NLP. One of the most visible products of this project was the EDR dictionary from the Electronic Dictionary Research Center founded in 1986. Lastly, the Association for Natural Language Processing was formally formed by the Japanese in 1994. The development of NLP research in Japan is atypical of Asian languages, largely because Japan leads Asian countries in terms of technology development.

In most other Asian countries, research on NLP is relatively new or in its infancy: interest in Chinese has increased dramatically over the past ten years due to China’s emergence as a world power, but many other countries are only now initiating work on NLP for their languages. In general, the history of the development of language processing capabilities for Chinese is more similar to that of other Asian languages than to Japanese. T'sou (2004) summarizes the developments of Chinese language processing. Even though the earliest efforts on Chinese language processing can be traced back to the 1960’s, more concerted efforts started in the late 1980’s, marked by the first computational linguistics conferences in both China and Taiwan in 1988 and followed by increased research activity in the 1990s (T’sou,2004). Related research became more visible in the 1990’s. Based on a chronology provided by Chu-Ren Huang, T'sou (2004)
showed that the maturing of the field was marked by the arrival of sharable resources in the early 1990’s, which were developed independently at the Academia Sinica and at Peking University. The quantity and quality of NLP research increased through the years, and finally reached the milestone of the formation of SigHAN, the special interest group on Chinese language processing, within the Association for Computational Linguistics in 2002. One may observe that in this chronology, the availability of language resources has served as both a foundation for research activity and a landmark of its maturity. This observation underlines the design feature of this special issue on Asian language processing. The dual foci on both language resources and language technology allow us to capture the dynamic, multi-dimensional state of Asian language processing, a research sub-field in its early development stage yet already producing exciting and challenging results [25].

2.4.3 Monolingual information retrieval for Asian languages

In order to develop IR systems for Asian languages, many of the underlying assumptions made about European morphology must be revised, and new indexing and retrieval strategies must be developed. While only one byte was used to code one character in European languages, now one to four bytes are needed for the Asian languages (Lunde, 1998). Moreover, Chinese documents may be written either in the traditional writing system (usually encoded in BIG5) or in simplified Chinese characters (encoded using a GB standard character set). On the other hand the vocabularies used are not always the same, due to the existence of various dialects (e.g., Mandarin, Wu, Hsiang, Min). In the Japanese language, documents may be written using Kanji ideograms (originating in China) together with the Hiragana and Katakana syllabic character sets and may possibly include some ASCII characters (used to express, for example, numbers or company names such as Honda). Finally, in the Korean language, both the Hanja and Hangul writing systems are found, although currently Hangul characters are clearly the ones most often used [26].
2.5 Asian languages and Localization

Localization is the process of enabling computing experience in local culture and language. This would require developing solutions to input process and output information in local language. For oral cultures, which do not have written languages, this would also mean ability to input, process and output speech instead of text. Also, it is important that the input, processing and output are agreeable with culturally acceptable norms, e.g. writing direction (left to right, right to left, top to bottom, etc.), formatting (e.g. Arabic script does not have italics form of text), color (red color represents friendship in China but danger in North America), etc. This is not easily possible, as current computing has evolved out of western cultural traditions and languages and also because Asian languages and conventions are not always as well defined as required for computational modeling. This chapter explains the scope of localization for a language. A greater part of localization is dependent on modeling linguistic details of languages. In order for proper computational modeling, very precise definitions are required for all the relevant linguistic phenomena. For many languages spoken in developing countries, these linguistic details are either not studied or at best partially and imprecisely defined. This poses a significant obstacle to localization. Therefore many times, a significant linguistic analysis is required before taking the localization process forward. Similar challenges also exist in cultural conventions, which are known but normally not documented. Thus, it becomes very important to involve native experts in the process. As localization involves definition and standardization of linguistic phenomena for computers, the process requires technical experts and technical organizations (e.g. Ministries of Communication or IT) to work with linguists and related organizations (e.g. National Language Authorities and/or Cultural Ministries). This poses another challenge because in most of the developing countries there is little cooperation between these two disciplines and hardly any people who have cross disciplinary expertise. In fact, many developing Asian countries have very limited number of formally trained and practicing computational linguists. Listed below are some of the linguistic requirements and the corresponding modeling for localization.
2.5.1 Character Set and Encoding

The most fundamental and foremost requirement of localization is the definition of the character set or alphabet of a language. This includes the basic characters, digits, punctuation marks, currency symbol, special symbols (e.g. honorifics, etc.), diacritical marks, and any other symbols conventionally used in dictionary making and publishing. Though the basic repository is normally known, it has been the experience of the authors that when more precise definition is required, especially for standardization, there are always a few ambiguities. Some common linguistic level challenges faced during standardization process are listed below, to illustrate the kind of decision standardization bodies may need to make.

- It is not always clear what is part of basic character set and what is to be included in auxiliary characters
- It is sometimes ambiguous if diacritics should be independently included or extra characters need to be defined which the diacritics have fused within them
- Though basic character set is known, larger set used for dictionary making and publishing is not known or well documented
- Some characters are not defined, e.g. currency symbols

2.5.2 Fonts and Rendering

Defining an encoding is not sufficient for supporting a language in computers. The internal codes must be displayed on the screen in terms of textual characters for it to be put to any significant use. This is done through fonts and rendering. Fonts represent the shapes of characters (also called glyphs) corresponding to each code for the language and also rules to indicate how these characters may alter shape or position on the screen in context of other characters. Font files store this information. Software (called a rendering engine) is required to take the input from user and corresponding shapes and rules from a font file to generate the actual shape and position for display on the screen. Initially fonts were “simple” as they were designed for Latin script in which character shapes or
positions are not context dependent. For example, an ‘a’ always looks the same where ever it occurs and is always on the baseline. These fonts only stored the basic shape and position of each letter, e.g. True Type fonts (TTF). However, as more scripts were computerized, it was realized that they were context-sensitive, cursive and required multiple shapes and variable positioning for their characters. For example, in Arabic script, letters have different shape in isolation, and in word-initial, word-medial and word-final positions. So font formalisms were extended and improved to store multiple shapes for each character and positioning and contextual rules for them, e.g. Open Type fonts (OTF, open standard by Microsoft and Adobe) and Apple Advanced Typography (AAT by Apple) . As explained earlier, displaying output requires a rendering engine, which can read a font file and create appropriate output against the input. There are a few rendering engines being used. Microsoft has developed Uniscribe rendering engine (shipped as USP10.dll file), which allows Open Type fonts to be displayed on Windows platform. Similarly, Apple has a rendering engine associated with its AAT fonts. Graphite engine by Summer Institute of Linguistics (SIL) is available for both Microsoft and Linux platforms. Pango is another engine available for GNOME (GTK+) platform on Linux. These engines support Unicode but provide varying degree of support for different scripts and languages. Level of support by some of these engines is discussed for each language later in this report.

2.5.3 Keyboard Layout and Input Method Engines

After character set is finalized, the next step is to place the characters across the keyboard to allow users to key-in the text. For keyboards lack of standards is normally not the problem; the problem is that there are multiple standards. These standards can be categorized in the following manner.

- Most of these standards are inherited from layout for typewriters, teletypewriters and other such devices
- Due to easy to configure utilities, which enable users to define their own on-screen keyboard layouts for most languages, there are “phonetic” versions of
keyboard layouts. These are defined by users who are used to English layout and map English letters to the similar sounding characters in their language.

- Many vendors also offer their own keyboard layouts, based on their own encoding schemes. These may be arbitrarily different from others.

The existing standards may be adopted and adapted for newer standards. The decision could be based on a variety of (not always scientific) reasons. Some of the problems associated with keyboards are listed below, which would need rectification.

- A keyboard layout may not include all characters in a language encoded by current computing standards, e.g. Unicode, because character set inventory has been expanded or altered from the earlier definition, e.g. tele-printers had different requirements from publishing industry so layouts for them may not have all the characters. Also, many countries are now introducing currency symbols, which did not exist earlier.

- Due to mechanical limitations, earlier layout was not intuitive for writing system of a language; those mechanical limitations are not applicable to computing paradigm any more. For example, single vowels which surround a consonant from left and right in Thai, Lao, Khmer, etc. had to be broken into two parts, one typed before the consonant and other after the consonant due to mechanical limitations. This is not a limitation in computing paradigm.

- Sometimes encoding has implications on keyboards. For example, Unicode has redundancies due to some design decisions, e.g. backward compatibility. It has been a compromise between practical and academic challenges. So it has to be decided which letter(s) within the encoding need to be placed on the keyboard.

Faced with these challenges, the countries need to reach a consensus on a formal layout which can serve their languages as comprehensively as possible and is intuitive for the users.
Chapter 2: Multilingual – Monolingual Information Retrieval

2.5.4 Collation

For applications which go beyond basic word processing, one of the most significant standards required for processing of any language is the definition of collation or sorting sequence, also sometimes called lexicographic sequence. Given different words in any language, collation determines the order in which they would be arranged, as is expected by the users. This is defined by their arrangement in the dictionaries. This standard is required for indexing in databases and any significant textual processing, e.g., making voter lists. Encoding standards are normally implicitly based on character order, but often do not determine collation completely. This is especially true for Unicode standard, which defines an arbitrary collation order (based on default character collation weights given in DUCET) which does not sort languages properly. Unicode standard requires language specific collation weights specified and standardized independently by relevant organizations for each language. These weights can be used with Unicode Collation Algorithm (UCA, available at Unicode website) for sorting. This algorithm orders words based on collation weights provided to it for a language. Languages use a variety of mechanisms to collate strings. This may be based on stroke count or phonetically equivalent Latin strings (e.g., in Chinese, Japanese, and Korean), letter sequence along with diacritics and/or capitalization (e.g., in Latin-based scripts), consonantal root (e.g., in Arabic language), dictionary order (e.g., in Khmer on Choun Nat Dictionary) or syllabic content (e.g., in Lao), etc. For many languages in developing countries, this sequence is not very precisely defined. In authors’ own experience, analyses have shown that different dictionaries in at least some languages do not agree in collation especially in finer details. However, for the computer, these orders must be defined to last detail. First step, again, is to involve language and cultural authorities and other relevant organizations to finalize the linguistic level standards for collation very precisely for all the characters encoded. This has to be done at the level of each language, for at least a country or a region. Second step would then involve developing effective algorithms or collation weights to realize that order. Many times lexicographic order for existing words may be determined based on dictionaries.
However, in these cases mechanisms still have to be devised for the introduction of new words and proper names not present in the dictionaries.

### 2.5.5 Locale

Locale is used to define some basic language and cultural conventions for the user interface of computers and other ICT devices. It includes definition of date, time, number and other formats preferred by different countries. For example, fractional part in a number is separated by a dot in US and UK but by a comma in some European countries. It also specifies day, month and other common strings, currency symbols and calendars used by different cultures. Locales need to be defined in standard repositories so that same information can be used by everybody for consistency. One such repository, recently established to eliminate any variations, is Common Locale Data Repository (CLDR), available through Unicode website. IBM ICU also has locale definitions. Locales are also maintained by other vendors. Locales are defined for every language for every country. Therefore, a combined language and country identification is used, e.g. ur_PK indicates Urdu as spoken in Pakistan and ur_IN indicates Urdu as spoken in India. These language and country codes are standardized through ISO 639.2 and ISO 3166 standards respectively.

Many developing countries are still not decided on standard conventions and therefore it becomes difficult to define these locales. For example, in Urdu in Pakistan both Latin and Arabic script digits are used and people disagree on which conventions should be used in the future. Once the conventions are defined by a country for a language they are submitted for standardization.

### 2.6 Asian Languages with reference to localization brief overview

#### 2.6.1 Arabic

Arabic is a Semitic language spoken by about 206 million people across the world, especially in Middle East and North Africa, where it is also the national language of many countries. There are many dialectical variations of Arabic across this region. It is widely used as a medium of communication in schools,
government institutions and media in most of these Arabic speaking countries. Figure below shows the linguistic lineage of standard Arabic.

![Figure 2.1: Language Family Tree for Arabic](image)

Arabic script has evolved from the ancient Aramaic script, and has been in use since the 4th century AD. Earliest known Arabic inscriptions date back to 512 AD.

### 2.6.1.1 Character Set and Encoding

Unicode Arabic script block ranging from 0600-06FF is the standard character set encoding used for Arabic language. ISO 8859-6 is also widely used. This standard contains Arabic in addition to basic Latin characters and is an 8-bit standard. These standards have been derived from earlier standards, e.g. ASMO 449, CODAR-U and ISO 9036. Microsoft also used Arabic code page 1256 based on these earlier standards.

### 2.6.1.2 Fonts and Rendering

Arabic fonts are widely available. In addition, these fonts are well supported on multiple platforms.

### 2.6.1.3 Microsoft Platform

Microsoft ships an exclusive version of Windows and Office products in Arabic language. Microsoft Arabic Windows includes a rich inventory of fonts for Arabic, many of which are not available in the English version of Microsoft Windows.
2.6.1.4 Linux Platform

Arabic script is fully supported on all Linux based applications. However, Open Type fonts do not exhibit satisfactory results. Arabic distributions provide support for rendering only basic four shaped fonts.

2.6.1.5 Collation

Arabic collation is supported in Arabic Windows. LC_COLLATE for Arabic language has not been defined yet. Default sequence for sorting is used, which sorts data similar to original collation sequence for Arabic language.

2.6.1.6 Locale

Arabic (ar) locales are defined in IBM ICU library and CLDR 1.3 for different countries. They support Arabic date, time and number formats, currency symbol and collation. The locale is available for many countries where Arabic is spoken as a national language.

2.6.2 Burmese

Burmese belongs to Tibeto-Burman language family and derives from Sino-Tibetan, as shown in Figure 7. It is the official language of Myanmar, where 32 million people speak it as their first language. Some people in China and India also speak Burmese.

Figure 2.2: Language Family Tree of Burmese
Myanmar or Burmese script is used to write Burmese language. The script has been developed from the Mon script, adapted from southern Indian Pali script. The earliest known inscriptions in Burmese script date back to 11th century.

2.6.2.1 Character Set and Encoding

Unicode code chart 1000-109F is the internationally standardized character set encoding for Myanmar script but is not frequently used. Two other ad hoc character set encoding schemes, MyaZedi developed by Solveware Solutions) and Win/CE/Geocomp, are more frequently used at the national level.

2.6.2.2 Fonts and Rendering

Microsoft’s support for TTF and OTF fonts is able to render Myanmar fonts but fonts shipped by Microsoft do not support Myanmar. Many Myanmar Unicode fonts have been developed by local vendors and are available. Work is under progress to provide support in Pango rendering engine for GNOME. Mozilla (Firefox and Thunderbird) builds are also partially available in Myanmar.

2.6.2.3 Collation

There are two main collation sequences used for Myanmar, Pali order used for older dictionaries, and Spelling Book order used in modern dictionaries. Non-Unicode fonts allow variable sequence of keystrokes to generate the same surface string, making it difficult to develop sorting sequences. However, Unicode enables a unique input sequence, on which collation can be built. Details of how to develop a collation sequence based on modern lexicographic order are available. A Myanmar collation sequence developed by Myanmar NLP has been standardized nationally but is not widely known and used yet. Microsoft platform does not provide collation support for Burmese. Myanmar NLP Research Center has developed a Myanmar sorter, which can sort Myanmar text in Unicode. GeoComp has also developed a sorting engine based on GeoComp Myanmar font encoding. Myanmar collation is defined in IBM ICU and Glibc for open source platforms.
Chapter 2: Multilingual –Monolingual Information Retrieval

2.6.2.4 Locale

Burmese locale language name is “my” and country abbreviation is “mm” (earlier “bu” in ISO 3166). Myanmar locale data is not defined in the latest version of CLDR or IBM ICU. Microsoft does not provide support for Myanmar locale. Locale is being defined on Linux platform by Myanmar LUG and Myanmar NLP Research Center.

2.6.3 Chinese

Chinese is a Sino-Tibetan language spoken by about 867 million people across the globe. One fifth of people in the world speak some dialect of Chinese. It is the national and the official language of China, Taiwan, Singapore and United Nations. Chinese is spoken in more than fifty different dialects within China and also in Brunei, Cambodia, Indonesia (Java and Bali), Laos, Malaysia (Peninsular), Mauritius, Mongolia, Philippines, Russia (Asia), Singapore, Taiwan, Thailand, United Kingdom, USA and Vietnam.

Chinese is written with characters known as Hanzi. Each Chinese character represents a syllable of spoken Chinese and also has a meaning. The characters were originally pictures of people, animals or other things, but over the centuries they have become increasingly stylized and no longer resemble the things they represent. Many characters are actually compounds of two or more characters. The simplified script (Simplified Chinese) was officially developed in the People's Republic of China in 1949 in an effort to improve literacy. The simplified script is also used in Singapore but the older traditional characters are still used in Taiwan, Hong Kong, Macau and Malaysia. Further simplifications were published in 1977 but proved very unpopular and abandoned in 1986.

2.6.3.1 Character Set and Encoding

There are three different sets of character encodings for Chinese, (i) Guobiao code for Simplified Chinese for Mainland China, (ii) Big5 for
Traditional Chinese for Hong Kong and Taiwan, and (iii) Unicode, which combines the two Chinese forms.

2.6.3.2 Fonts and Rendering

Free and vendor fonts for different encodings are available for Chinese. Rendering is also supported well on many different platforms. A number of Chinese fonts are available for both Microsoft and Linux platforms. On Debian Arphic TT Chinese fonts, xfonts-intlchinese, xfonts-cjk and Unifont, and on Red Hat taipeifonts, ttfonts-zh_CN, ttfonts-zh_TW, are among many Chinese fonts being used.

2.6.3.3 Collation

Like input methods, collation may also be done in a variety of ways, including phonetic, alphabetic (e.g. based on Latin input), stroke based or dictionary based, for example, Chinese Big5 order, PRC Chinese Phonetic order, Chinese Unicode order, PRC Chinese Stroke Count order and Traditional Chinese Bopomofo order. Many of these methods are available on various platforms, including Linux and Microsoft.

2.6.3.4 Locale

Locales for both Traditional and Simplified Chinese have been defined in IBM ICU [24]. Locales for Chinese are also available in the CLDR 1.3. The locales are zh_CN (for China), zh_TW (for Taiwan), zh_HK (for Hong Kong) and zh_SP (for Singapore).

2.6.4 Hindi

The word “Hindi” is derived from Sanskrit word ‘Hindva’ meaning 'language of Hind’. About 180 million people speak Hindi as their first language and many more across the globe use it as a second language. Hindi is the national language of India and is also widely spoken in Bangladesh, Fiji, Indonesia, Malaysia, Mauritius, Nepal, South Africa, Uganda and Yemen. It is the third most spoken language and comes after Chinese and English. Hindi belongs to the Indo-
Chapter 2: Multilingual – Monolingual Information Retrieval

European language family and has influences from Persian and Arabic. Its formal vocabulary is derived from Sanskrit and Prakrit.

![Language Family Tree for Hindi]

Figure 2.3: Language Family Tree for Hindi

### 2.6.4.1 Character Set and Encoding

ISCII (IS 13194:1991, earlier IS 13194:1988) is the national standard for Devanagari character set encoding, based on earlier standard IS 10402:1982. ISCII is a standard for Devanagari script and may be used for other languages. It is widely used in India. The standard contains ASCII in lower 128 slots and Devanagari alphabet superset in upper 128 slots and therefore it is a single byte standard. Though it is primarily an encoding standard (and sorting is usually not catered directly in such standards, e.g. see Collation section below), the standard was devised to do some implicit sorting directly on encoding. Official standard publication is available. Unicode provides an international standard for Devanagari character set encoding based on IS 13194:1988 from 0900 till 097F (and therefore is not exactly equivalent to IS 13194:1991; This may be used for Hindi and other Devanagari script based languages, including Marathi, Sanskrit, Prakrit, Sindhi, etc.

### 2.6.4.2 Fonts and Rendering

There are many fonts available to write Hindi on different platforms. Some are listed below.

### 2.6.4.3 Microsoft Platform

Windows provides Mangal font, which has been developed by CDAC, India, and other fonts for Hindi. Windows uses Uniscribe as the rendering engine,
which supports rendering of Open Type fonts for Devanagari script. Results of Hindi fonts rendered on Microsoft are shown in below.

(a) Mangal, (b) Kokila, and (c) Arial Fonts on Microsoft Office 2003

2.6.4.4 Keyboard

As for Hindi character set encoding formats, different software vendors have implemented many different keyboard layouts e.g. Godrej, Ramington, Phonetic, Shusha, and Traditional keyboard layout. Inscript is the standard Hindi keyboard layout and is the most commonly used. It is shown in Figure 9 below.

Figure 2.5: Inscript Keyboard for Hindi

2.6.4.5 Collation

Work had been in progress to finalize a single collation sequence standard for Government of India. However, ambiguities in the linguistic sorting order of the Hindi character set have hampered this standardization. The work is still in progress.
2.6.4.6 Linux Platform

Collation on Linux platform is done using LC_COLLATE in the locale definition. The current hi_IN locale file does not have collation data included, so default sort order is used. As such this suffices for basic Hindi sorting, because Devanagari range in Unicode is based on ISCII-8.

2.6.4.7 Locale

No significant effort at national or regional level has been undertaken to standardize Hindi locale document, which does implicit sorting for Hindi. (hi_IN), though it is included in CLDR 1.3. Some work has been done by CDAC in defining a locale for Microsoft to enable Hindi in Windows [27].