Asian Language Processing: History and Perspectives

S.T. Nandasara
*University of Colombo, Sri Lanka*

Yoshiki Mikami
*Nagaoka University of Technology, Japan*

This special issue focuses on Asian language processing in India, Sri Lanka, and Thailand. Specifically, it discusses historical background of specific writing systems; script structure and major issues relevant to text processing; important development stages in type-printing and the typewriter era; milestones marked by the design of character code standards for computing; and issues involved in designing a mechanism for input and display interfaces in Asian language computing.

Asia has a great variety of genetical and typological diversity of languages—close to 7,000. Of those, more than 3,070 languages are spoken in Asian and Pacific countries. The world’s largest language family, Indo-European, is found in Central, South, and some parts of the East Asia. Indo-Aryan, an Indo-European subfamily, is used throughout South Asia. The second-largest language family, Sino-Tibetan, is used in many parts of Asia—most commonly in East and Southeast Asia.

In addition to its rich language diversity, Asia is especially rich in scripts and writing style. Three types of scripts originated in three separate geographic areas. Brahmi script originated in South Asia in and around the Indian subcontinent and the continental part of Southeast Asia; Chinese ideographs originated in East Asia; and the Aramaic script originated in Southwest Asia and spread to West and Central Asia.

Writing style played a significantly different cultural role in traditional South Asia (i.e., the Indian subcontinent) than elsewhere in the ancient world. The earliest writing system of South Asia was based on the Indus script, which originated in the third millennium BC but seems to have died out in the following millennium. Indic scripts—Brahmi and Kharoshthi—appeared much later, and evolved from scripts other than Indus. The earliest datable written documents in these scripts are the rock and pillar inscriptions of the Mauryan emperor Asoka (3rd century BC).

The later history of South Asian scripts consists essentially of the Brahmi script, which was the ancestor of literally dozens of Asian scripts. Early regional varieties of Brahmi eventually developed into distinct scripts often associated with particular languages of the Indo-Aryan, Dravidian, and other families. But in systematic terms, these scripts typically share the same basic principles of the script system—that is, “diacritically modified consonant syllabic scripts.”

The spread of Brahmi script into Southeast Asia is generally established by the earliest known examples of some brief inscriptions on precious objects discovered in southern Vietnam. However, there is no concrete evidence that the objects were actually inscribed in Southeast Asia, since the area was an important trading center during the period from the second to the fifth century AD. Note that the types of script that have spread in Southeast Asia all seem to have originated from the Pallava and Grantha scripts of the south of India. These include those of the Burmese, the Lao, the Khmer, the Thai; the obsolescent national scripts of the Mon, the Javanese, the Sundanese, the Balinese, and the Cham, among others, in addition to tribal scripts in Sumatra and the Philippines.

In recent centuries, under Islamic influence, the Arabic script has become the written tool for some South Asian languages (e.g., Urdu, Sindi, Kashmiri, and Divehi). The disappearance of Asian language writing systems declined in the 16th and 17th centuries.
This situation resulted in part from the exploration and widening economic interest of a number of Europeans—Portuguese, Dutch, English, French, Spaniard, Russians, Siberians, and others. As a result, the roman script was introduced for Southeast Asian countries such as Vietnam, Malaysia, Philippines, and for Indonesian language scripts such as the Javanese, the Sundanese, and the Balinese.

Before computing came into the region, printing in Asian languages was the major challenge. Printing technology was an adaptation of what was available in western countries for the roman alphabet. For Asian type-based printing, a large set of precast conjuncts, individual characters, and symbols running into the thousands, of varying size and shapes, were used for manual composing. The Printing Press in India addressed a letter on this subject to Rome in 1608 and the following extract below shows the difficulties of casting hundreds of molds.

Before I end this letter I wish to bring before Your Paternity’s mind the fact that for many years I very strongly desired to see in this Province some books printed in the language and alphabet of the land, as there are in Malabar with great benefit for that Christian community. And this could not be achieved for two reasons; the first because it looked impossible to cast so many moulds amounting to six hundred, whilst as our twenty-four in Europe… A Jesuit Friar’s letter to Rome, 1608

Nearly three centuries later, in about 1890, major improvement in the quality of European typeface design and layout were achieved with the arrival of a new mechanical typesetting machine, the Monotype composer. However, it took almost seven decades to introduce some of the Asian scripts such as Sinhala for the Monotype composer. This was a common problem for many of the Asian languages, since they are not based on the Latin alphabet.

Complex writing systems used in many Asian countries, as we’ve mentioned, are made up of vowels, consonants, matra symbols, diacritical marks, and special symbols—and, in several cases, tone marks. Writing style is from left to right and usually written in three levels. For the Thai case, tone marks go into the fourth level. The matra symbols are attached to the left, right, top, or bottom of the consonants. Several components are combined, and some are reshaped to form complex ligatures. Sometimes the total number of these different ligatures runs into several thousands. Early typewriting, printing, and computing equipment for Asian language processing needed to provide for this degree of complexity in both display and printing.

Early computer technologies—that of the early 1970s and 1980s—were well designed predominantly for English and other Latin-based languages in particular; on the other hand, there was an eminent lack of support for the accented letters used in Western European languages and for the combined conjuncts in non-Latin scripts. The early technology rendered compatibilities to a certain extent for the application of Asian languages; however, the existing character codes conflicted with the 8-bit-coded character sets used in Asian languages. Nevertheless, complex Asian text processing was represented on early computers using additional hardware with 8-bit extended ASCII encoding, including the ISO 8859 character set from nearly 30 years ago.

The first article in this special issue, “A Journey from Indian Scripts Processing to Indian Language Processing,” addresses common features, structural similarities, characteristics, and complexities of modern languages used in India. The article is based on the Devanagari script, which is used for writing classical Sanskrit, modern Hindi, and Marathi, for example, as well as for Nepali. In this article, the author examines the adaptation of Indic scripts to computing and how this facilitated development of Indian language processing. Emphasis is on the features common in Devanagari and on how the unifying characteristics of the script and language have been exploited to provide solutions applicable to almost all Indic scripts and languages. The article first provides an overview of early attempts made to manually compose, print, and type-write (mechanical and electronic) Indian scripts, then moves on to analyze the development of interfaces with computers concerning keyboard design, input method, and internal coding. The article also examines early standardization efforts for Devanagari script. Indian national standards have had a major impact on the encoding methodologies of ISO/IEC 10646 and Unicode. Other issues such as rendering, collation techniques, and sorting are also discussed. Finally, the article takes a look at the development of interfaces, applications, and language tools such as OCR and machine translation systems.

The second article, “From the Past to the Present: Evolution of Computing in Sinhala Language,” is based on the Sinhalese language, which underwent a largely separated process of development from early Brahmi. Sinhala was imported from North India along with
Guest Editors’ Introduction

Buddhism in the third century BC, but scripts were influenced at various stages of its development by South Indian scripts. Besides these major literary scripts, countless other local scripts and variations have been encountered through all periods of history.

The article provides an overview of the historical development of modern Sinhala scripts—which are used in Sri Lanka to write Sinhala, Pali, and Sanskrit. The author describes the philosophy behind the standardization and implementation of the 8-bit and 16-bit code design, and examines the structure, features, characteristics, complexities, and rendering issues of the language scripts. This is followed by an overview of early attempts to mechanize Sinhala scripts—manual composing, printing, and mechanized and electronic typewriting. The author explores how the language was standardized for printing books, government publications, and newspapers; and then explores how computer codes were developed for representing the language, including a discussion on the problems with sorting the language characters and codes. The article highlights the culturally encoded aspects of computing—especially at the very elemental level of representing non-Western languages.

In the third article, “Computers and the Thai Language,” the authors explore the language, script, writing system typewriter, Thai standards for computers, and I/O systems. The computerization of the script presented here differs from the previous two articles. Although the script is derived from Khmer, it is generally developed from Pallava in South India. This article examines the Thai language, which is characterized by its non-Indic vowels and consonants, writing from left to right without regular word spacing, and its tone marks (ingenious adaptations of the scripts).

Thai computerization history goes back to early 1960s—much earlier than any other South and Southeast Asian countries. The article highlights the authors’ experiences with standardization of Thai script in the mid-1980s. Interestingly, Thai script encoding contrasts sharply with the Indic one described in the first article. The authors address input methodology requirements, typewriter keyboards, and conventions for Thai support implementation, followed by a discussion of the development of display and printing with PCs, covering aspects of internal coding and collation principles. Finally, the research and development on language processing, such as OCR, hand-written character recognition, text-to-speech, and automatic speech recognition are reported.

Although we hear of plants or animals that have become extinct, it turns out that languages are threatened with extinction too—many languages and their scripts have disappeared from the planet. Some languages are rarely used in the domains of everyday use, whether public and private.

Information and communications technologies (ICTs) are continually changing our world. Data exchange is becoming ubiquitous worldwide. On the other hand, the digital divide remains in the developing world, including a number of Asian countries. The languages used on the Web are not those most widely spoken in these countries. The inability to do computing in local languages is a major obstacle to providing universal access to information, which may be considered a basic human right.

Not long ago, the year 2008 was proclaimed the International Year of Languages by the United Nations. In response, UNESCO’s director-general asked all stakeholders, including educational institutions and professional associations, to increase activities for “the promotion and protection of all languages, particularly endangered languages, in all contexts.” We believe that the pioneering experiences presented in this special issue of the Annals will benefit concerned engineers and practitioners who seek further development of their respective language-processing technologies.

Acknowledgments

This special issue would not have come into being without the insight and assistance of Annals editor in chief Jeffrey Yost and editorial board member Chigusa Ishikawa Kita—our sincere gratitude to them. We further thank Louise B. O’Donal, managing editor; Alkenia Winston, publications coordinator, and other Annals staff for their support. Last but not least, we thank our colleague A.B. Disanayaka, professor of archaeology, University of Colombo, for providing his translation of the Lumbini Asoka pillar inscription for the cover.

References and notes


S.T. Nandasara is on the Faculty of Science, University of Colombo, Sri Lanka, which he joined in 1981. He is a continuing member of the Sectoral Committee on Information Technology of the Sri Lanka Standard Institute and has closely worked with ISO and the Unicode Consortium for standardization of Sinhala scripts. He was a co-leader of the Asian Language Resource Network Project (2005–2008), and is a member of IFIP’s Working Group of the History of Computer Education. Nandasara holds a degree in development studies and statistics from the University of Colombo.

Contact Nandasara at stn@ucsc.cmb.ac.lk.

Yoshiki Mikami is a professor of management and information systems science at Nagaoka University of Technology, Japan. He has initiated several international collaborative projects, such as the Language Observatory Project, the Asian Language Resource Network Project, and the Country Domain Governance Project. He also serves as a chairman of Joint Advisory Committee for ISO registry of character codes. Mikami received a BE in mathematical engineering from the University of Tokyo and a PhD from the Graduate School of Media and Governance at Keio University.

Contact Mikami at mikami@kjs.nagaokaut.ac.jp.

For further information on this or any other computing topic, please visit our Digital Library at http://computer.org/csdl.