Language Engineering in a Multi-Lingual Environment - The Indian Context

Kavi Narayana Murthy
Department of Computer and Information Sciences,
University of Hyderabad,
Hyderabad, 500 046,
email: knmcs@uohyd.ernet.in

Abstract

In this paper[1] we look at the status of technology development in Indian languages, analyze the reasons for the slow progress and suggest some priorities for technology development.

1 Introduction

India is a land of One Billion people - about one sixth of the whole world. Our civilization dates back to many thousands of years. India is a land of many religions, many cultures and many languages. A life time is not sufficient to get even a glimpse of everything that is Indian.

There are about 150 different languages spoken in India, of which 18 have been given a kind of constitutional recognition and are considered to be the major languages. Indian languages encompass four language families - the Indo-Aryan, the Dravidian, the Tibeto-Burman and the Austro-Asiatic. Some of these languages have extensive literature going back to about 10th century AD. Many languages also exhibit a very rich oral 'literature'. The major languages are among the most widely spoken languages of the world. We have an extraordinarily systematic and scientific linguistic tradition from more than 2000 years now. Phonology, Morphology, Syntax, Semantics, Logic, Pragmatics have been studied extensively over the past several thousand years. It is a challenge even to make a survey of all the works and the various schools of thought that have originated, grown, changed and evolved

[1]The research reported in this paper was supported by the Department of Information Technology, Government of India under their TDIL (Technology Development for Indian Languages) programme.

in India over a continuum of thousands of years.

The three language formula has worked well in some cases but failed in others. Nevertheless, a large number of people know two, three or even more number of languages. This is a kind of natural multi-lingualism, very different from the kind of multi-lingualism you will find in countries that are essentially mono-lingual. You will find that more or less free mixing of several languages is very common. many documents are required to be in more than one language. With so many languages in use, identifying language is itself an critical step in many applications.

But where are we with respect to language technology? We should have been world the leaders in language technology. Instead, we are lagging far behind not only western languages but also the languages of the far east. We are still struggling to use computers as type-writers - type, compose and print!

This article attempts to describe the state of affairs as far as Indian language technologies are concerned. It also attempts to highlight the unique characteristics of our languages and of our traditional knowledge. In the end we venture to suggest some area of work that need to be given high priority.

2 The Status

We have stated above that there are about 150 different languages spoken in India. This is what linguists believe - we do not even have as yet an exact list of our languages! It takes a tremendous amount of effort to analyze each of these languages and the large number of dialects associated with each of them and study the vocabulary, the morphology, the syntax and the semantics. Such an effort has not been made of late. A majority of the recent work, especially in terms of technology, are limited to the 18 or so major languages.

Even in these major languages, resources available for technology development are scarce. Electronic dictionaries are only becoming available of late. There is no concept of a thesaurus in many languages. There is no computational grammar for any of these languages. Even the morphology (internal structure of words) has not been analyzed in enough depth and detail. It is not easy for an automatic system to say whether a given sequence of symbols is a valid word in a given language or not. We do not know exactly how many words are there in our language! There are no (good) spell checkers as yet in many of our languages.

The least that somebody would expect in today's times is a collection of texts in

electronic form. Such a large and representative collection of texts, called a corpus, has immense value for statistical and linguistic analysis and for developing technology at all levels right from dictionaries and spell checkers to intelligent information retrieval, automatic categorization, automatic summarization and automatic translation. Unfortunately, even large plain text corpora are not available. Only about 3 Million word corpora are available for most of the major languages. These corpora have not be thoroughly proof-read and hence are not very dependable.

There are very few web-sites and web pages in Indian languages. Most of them are not indexed by search engines because standard encoding schemes are not followed. Some sites use pictures instead of text! Many use font-encoded pages and either depend on local availability of fonts at client side, or more often, use dynamic font technology. A few have tried the plug-in technology. The WILIO technology developed at University of Hyderabad enables standard character encoded web pages to be used by clients without regard to the operating systems and browsers they may be using. WILIO also permits interactive web pages wherein the users can also type-in directly into the browsers in Indian languages.

Most of the newspapers, magazines and books will be in electronic form at some point of time during production but again, most often in completely non-standard, proprietary and secret encoding schemes and are thus useless for any further processing or analysis. In most cases, the electronic forms of the documents are never stored.

There is of course no question of more advanced technologies such as Automatic Document Categorization, Automatic Summarization, Intelligent Information Retrieval/Search Engines, Information Extraction, Speech Recognition, etc.

Speech technologies are especially important for a country like India. There are again certain characteristics of Indian languages which are quite distinct from English. For example, stress is relatively less important and other prosodic features such as duration are more significant. Aspiration is a contrastive feature. A deeper understanding of characteristics of our languages is essential and technology developed for other languages cannot be simply borrowed. There is a lot of work that needs to be done. However, very little has been done so far. We do not even have speech corpora!

Our scripts are more complex too. There are rounded features and it is not simply a linear sequence of shapes - shape units are arranged in 2 dimensions in complex ways. OCR systems have started appearing for Indian scripts only recently and it will take some more time before they become fully usable.

Basic issues such as language identification are still at research stage. Thus multi-lingual and cross-lingual applications are far from reality.

Most of higher education is imparted through the medium of English. Most people prefer to send their children to English medium schools. The quality of books and teachers in local languages are generally considered to be inferior, there is little material available in these languages and scope for gainful employment are relatively less. English is the language of choice in business, medicine, law and even the government, although there are efforts to encourage the use of our own languages. Most people speak their own mother tongue at home but use English for everything else. With each generation, there is a slow decay in the use of our languages. As the languages slowly go into oblivion, so do the vast treasures of knowledge that are encoded in those languages. Technology development for Indian languages has the potential to slow down and reverse this trend, at least to some extent.

The intention here is not to belittle the commendable work being carried out by several centres across the country over the past several decades. There are more than a dozen centres of excellence working dedicatedly on language technologies both within the TDIL mission of the Department of Information Technology and outside. Nevertheless, compared to what we could have achieved and what we should have achieved, what is achieved so far is meager. We only intend to understand some of the reasons for this state of affairs.

3 The Reasons

3.1 Characteristics of Indian Languages

Indian languages are characterized by several unique features that make them very different from other major languages of the world. Thus the technologies developed for English or Japanese cannot be borrowed in a more or less direct fashion and applied to our languages.

English and many other western languages use an alphabetic writing system. Any word is simply a (linear) sequence of the letters of the alphabet. On the other hand, ideographic languages such as Chinese and Japanese use pictures to depict meanings. Indian languages use a syllabic writing system.

Speaking and listening comes naturally whereas writing and reading come much later. Not all are capable of reading and writing - there are illiterates. Language *is* speech - writing is an artifact. Given this, our writing system has naturally developed into a system of shapes for representing sounds.

Syllables are units of sounds that can be uttered independently. However, the units of writing, called *aksharas* do not have an exact one-to-one correspondence with the syllables. The aksharas, sometimes called 'orthographic syllables', correspond to sequences of zero, one or more consonant sounds followed by a vowel sound. This definition implies that the number of possible aksharas is infinite. In practice, consonant clusters usually have only a small number of consonants. The largest number of consonants in a single cluster known is 5. If every possible 5 consonant combination is considered possible, the number of aksharas will be about ten Billion! Of course in practice many of these combinations never occur. Analysis of the available corpora show that only about 20,000 aksharas are in use. Out of these, about 5,000 aksharas account for more than 99% of all words. In any case the number of units of writing we have to consider is so large that typing in, editing, storing, processing, displaying and printing of Indian languages are all inherently very complex.

3.1.1 Script Grammar:

It is clearly not feasible to give a separate numerical code to each syllable. Fortunately we have an excellent solution. Indian languages are unique in having a script grammar, a 'grammar' of syllables, - a way of specifying all valid aksharas. The potentially infinite set of valid aksharas is specified by the following grammar, a very simple Finite State Machine.

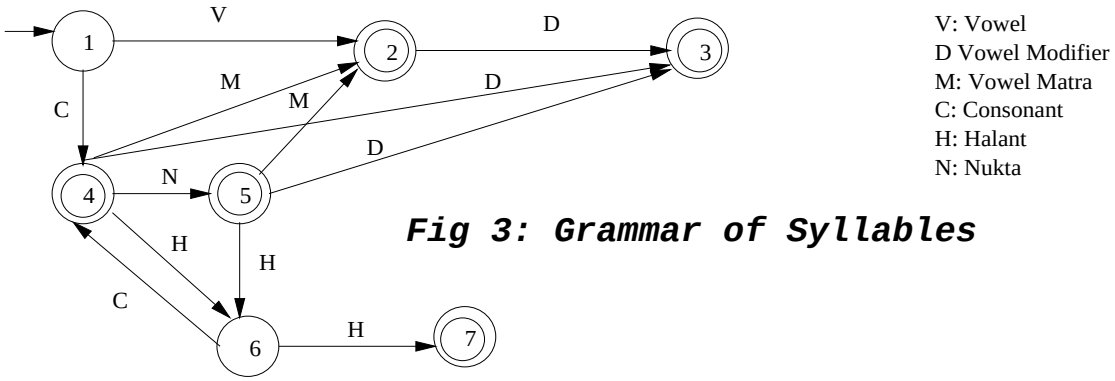

Fig 3: Grammar of Syllables

Note that the syllables are composed of more basic units such as Vowels and Consonants. All sequences of vowels, consonants, maatras and vowel modifiers are not valid akshara's. Thus it is not possible to have an isolated maatra, an akshara beginning with a maatra or an akshara with more than one maatra. Sequences can be grammatically valid or invalid, a concept non existent in English and other

languages of the world. Any sequence of alphabets is either a valid word or an invalid word in English but there is nothing ungrammatical or impossible about any sequence. Ungrammatical sequences, as compared to spelling errors, can never ever occur, even in proper names. Thus one level of checking for correctness is built into our scripts. Any character encoding scheme for Indian language scripts must ideally define a script grammar and implement it.

3.1.2 Character Encoding Standards: ISCII and UNICODE:

ISCII is a National Standard for character encoding of major scripts of Indian languages. Refer to the 1991 BIS standard for more information.

Barring a few languages that are written in Perso-Arabic scripts, Indian languages are written in 10 different scripts, all of which have the origins in the ancient Brahmi script. All these Brahmi based scripts have the same phonetic structure. It is therefore most appropriate to have a common script code for all these scripts, as indeed ISCII does. The ISCII scheme makes it possible to transliterate from one script to another trivially. Note, however, that there can be no multi-lingual (multi-script) plain ISCII text! UNICODE, the evolving de-facto international standard, on the other hand, provides separate code space for each language so that identification of language in a multi-lingual plain text becomes trivial.

ISCII is not a registered standard and is not supported by Operating Systems and Web Browsers. This makes it necessary for application software to handle all the issues, especially, the ISCII-to-Font mapping.

UNICODE for Indian languages is yet to pick up. Fonts are not yet widely available. Applications will need to be migrated to support UNICODE.

3.1.3 Fonts, Glyphs and Glyph Encoding Standards:

Since the number of syllables is very large, it is also not feasible to encode or store all the font shapes corresponding to each of these syllables separately. Also, the written syllables are graphically much more complex than the letters of English. It is therefore not feasible to store one shape for each basic ISCII character and then compose these shapes to get the required shapes for all the syllables.

The shapes we use in defining a font should be selected based on the simplicity of their being composed to obtain combined shapes for displaying full syllables. This dictates that we use a set of basic shapes which may not have a one to one correspondence with the basic characters encoded in a character encoding scheme such

as ISCII or UNICODE. The basic shapes we use in a font are called 'glyphs'. In English there is a near one to one correspondence between letters of the alphabets and glyphs used for rendering them. Not so in Indian languages.

There is no glyph encoding standard - each font uses a possibly different set of glyphs and positions them in possible different ways in the code table. Thus simply selecting a piece of text and changing the font can render the text as junk. Indian language Fonts have remained proprietary, non-standard and incompatible with one another.

Converting from a character encoding such as ISCII or UNICODE into and from a given font encoding is an additional step that is essential for Indian languages. There is as yet no 'standard' way of mapping from ISCII or UNICODE into various fonts - there is no standard 'grammar' to specify this mapping.

A large number of texts in Indian languages have been encoded in proprietary fonts and thus cannot be used directly for any language engineering application. If only we had respected the standards, we should be having large corpora in all the major languages by now.

3.1.4 Rich Morphology

Morphology plays a much greater role in Indian languages because our languages are highly inflectional. While the English verb *eat* gives rise to only a few variants such as *eats, ate, eaten and eating*, the corresponding verb in Telugu can give rise to a very large number of variants. Words in Dravidian languages like Telugu and Kannada are long and complex, built up from many affixes that combine with one another according to complex rules of saMdhi. For example,

nilapeTTukooleekapootunnaaDaa?
which means something like "Is it true that he is finding it difficult to hold on to (his words/something)?"

One linguist puts the number of variants for a single Telugu verb at nearly 200,000! The exact number of different forms that a verb can take in a language like Telugu is not yet clear.

While Indian languages in general are morphologically richer than languages like English, Dravidian languages are a lot more complex. The 12 Million word corpus of Telugu has nearly 20,000,000 different words and there should be many more as the growth rate studies indicate. In contrast, the Indo-Aryan languages have only

about 1,50,000 to 2,00,000 words forms in all. Dravidian languages including Telugu, Kannada, Malayalam and Tamil are among the most complex languages of the world and can only be placed along with languages such as Finnish and Turkish. Clearly, there is no way we can hope to list all forms of all words in a dictionary. We cannot build a spell checker, for example, by simply listing all forms of all words. Morphology is not just useful but absolutely essential. Our languages are inherently more complex than other languages such as English.

3.1.5 Syntax

Syntax of Indian languages is considered to be relatively simpler. It is however, quite different from that of English. Indian languages are relatively free word order languages. Word order is such an innate characteristic of English and other such languages that the grammar formalisms developed keeping such languages at the back of the mind are not suitable for Indian languages.

Computational grammars have not been developed for any of our languages so far. Many language engineering applications benefit significantly from syntactic analysis and the development of such applications is limited by the non-availability of computational grammars and parsing systems.

3.2 Available knowledge is not really 'available'

We have said above that we have an ocean of traditional knowledge, both in terms of breadth and depth, on almost all aspects of language, meaning, logic and understanding. Yet we are unable to leverage this wealth of knowledge and experience. This is due largely to the very nature of these traditional knowledge sources. Firstly, almost all the works are in Sanskrit and not many know Sanskrit. Secondly, knowing Sanskrit is not sufficient. According to Indian tradition, knowledge was not meant for all - in fact every effort was made to ensure that knowledge does not reach the hands of the "un-deserving". (If everybody is taught how to make bombs you know what happens.) Knowledge was meant only for those who are extremely serious - those who consider seeking knowledge as the main goal of life. Given these, it was expected that a seeker of knowledge should search for the right teacher (guru) and learn from him. Public knowledge was thus limited, sketchy and incomplete. An extremely cryptic style is followed - often in the form of sutras or formulae. It is part of the tradition that commentaries are written on the original texts to make them easier to understand and commentaries are written on such commentaries! To this day, getting an in depth understanding of these works requires spending years of your life with a guru. With the increasing shift to western life styles and cultures,

the number of serious students within the Indian tradition is steadily decreasing and so it is becoming more and more difficult even to find a good guru. Added to all this is the difficulty in communicating with traditional scholars in terms that make sense to a modern language engineer. Further, the original purpose of these works were very different from the purposes for which we wish to use them today. By and large, this wealth of traditional knowledge has remained dormant.

3.3 Need is not felt

Nearly 95% of our people are either completely ignorant or not at all comfortable in using English in their daily life. Thus the benefits of information revolution and all the technologies that we develop will never reach the majority of the people if we limit ourselves to English. However Indian languages are not used in daily life as much as may be expected and, naturally, the need for developing technology for these languages is not being felt in any great measure by the educated, rich and urban population.

3.4 Trained manpower

Language technology issues are not widely known or appreciated. These technologies are not part of the curriculum and trained manpower is in short supply. It is also not easy to motivate people to work in this area. Lack of adequate manpower is one of the main reasons for the slow progress.

4 Conclusion

We have noted that developments in technology for Indian languages has been painfully slow and the major reasons for this include non-availability of large scale data resources and slow development of basic tools technologies, partly due to inherent complexity of our languages.

The first and the most important step therefore is to develop large scale, representative, high quality data resources - plain and annotated corpora, electronic dictionaries, morphological analyzers, computational grammars etc. This would be greatly facilitated by following national and international standards. Data generation is always a tedious and time consuming affair and collaborative development is a model that needs to be explored seriously. Following the evolving standards with regard to internationalization is essential even as localization efforts go on.

Manpower development is an area that needs to be given highest priority. Availability of interesting and useful content motivates users and this in turn cranks the technology development cycle. Hence attention should be paid to the development of quality content and make them easily available to users. The web is an extremely powerful medium for this.