

# Achieving the Goal of Language Technology for All

**Ramakrishnan A G**

MILE Laboratory, Department of Electrical Engineering

Indian Institute of Science, Bangalore, India

agr@iisc.ac.in

## Abstract

Given the advances in information technology, communication and internet, enabling individuals, organizations and governments to carry out their day-to-day transactions without being constrained in any way by the multitude of living languages is completely feasible, provided there is commitment to the cause and constant, undaunted efforts. In a country like India, this must be an ongoing process, since the various fields of knowledge are constantly advancing, giving rise to new terminologies and challenges. It requires systematic planning and execution by standing committees, both at the national and state levels, who work together, constantly communicating and collaborating with one another. Further, it may also be worthwhile looking at establishing a national level research organization for continuous upgrading of speech and translation technologies, especially in a code-mixed scenario.

**Keywords:** standardization, standing committees, translation, commitment, vision, policies, multilinguality, fonts, web content, multilingual dictionaries, open databases, transcription, computational linguistics.

## Résumé

பல வாழும் மொழிகள் இருந்தாலும், தகவல் தொழில்நுட்பம், தகவல் தொடர்பு மற்றும் இணையம் ஆகியவற்றின் சிறந்த முன்னேற்றங்களால், தனிநபர்கள், நிறுவனங்கள் மற்றும் அரசாங்கங்கள் தங்கள் அன்றாட பரிவர்த்தனைகளை எந்தவொரு வகையிலும் கட்டுப்படாமல், குறையின்றி செயல்படுத்துவது முற்றிலும் சாத்தியமானதே. ஆனால், அதற்கான அர்ப்பணிப்பும், நிலையான, இடைவிடாத முயற்சிகளும் தேவை. இந்தியா போன்ற ஒரு நாட்டில், இது ஒரு இடைவிடாது எப்போதும் தொடரும் செயல்முறையாக இருக்க வேண்டும், ஏனென்றால் பல்வேறு அறிவுத் துறைகள் தொடர்ந்து முன்னேறி வருகின்றன. இது புதிய சொற்களஞ்சியங்களுக்கும், சவால்களுக்கும் வழிவகுக்கிறது. இதற்கு, தேசிய மற்றும் மாநில மட்டங்களில், ஒருவருக்கொருவர் தொடர்ந்து தொடர்புகொண்டு, ஒன்றிணைந்து செயல்பட்டு, ஒத்துழைக்கும் நிலைக்குழுக்களும், அவற்றின் முறையான திட்டமிடல் மற்றும் செயல்படுத்தலும் தேவைப்படுகிறது.

## 1. Preamble

Even though more than 70 years have passed by since independence, none of the living languages, including the 22 scheduled languages, have kept pace with the scientific and technological advances and the consequent new terminologies. Thus, it is not possible for anyone, even if he/she is really interested, to pursue higher studies in science and technology in any of the Indian languages, including Hindi, which is being spoken by the majority.

The author feels that there has been no systematic planning to develop any of these languages to keep pace with the time. However, it is still not too late, since only in the recent past, the information and communication technologies have been advancing fast. Thus, with proper planning, commitment and systematic execution, India can still catch up and develop language technologies in all the major living languages and enable individuals, organizations and governments to improve their performance levels in spite of the presence of multiple, interacting and intersecting languages. It must also be realized that this must be planned as an eternal activity, and not as something that will get completed within a particular time.

### 1.1 Need for Multipronged Effort

There are multiple steps that need to be taken simultaneously and pursued vigorously. This involves standardization of terminologies in each language, technologies and the interfaces between themselves and other application softwares in multiple domains, wherein, again new developments will keep taking place.

### 1.2 Need for Multiple Standing Committees

Each state in India must have a standing committee, which consists of domain experts from different fields, linguists and technologists, who will periodically meet, plan and advise the respective arms of the industry and Government as to the steps taken for the next level of improvement of language technologies, newer applications, their deployment and outreach.

### 1.3 Ongoing Standardization of Terminology

To begin with, a systematic procedure must be evolved, which specifies the approach to be followed in coining a word for a new idea or object or an action. If possible, for languages such as Bangla and Tamil, which are official languages in countries other than India also, there can be understanding between the countries with respect to this activity, or even the committee may be constituted with members from all those countries. The committee for standardization of special terms to be used in business, art, science, technology, management and other special fields must periodically be creating new terms in the respective languages, publish them in dedicated websites for feedback from the experts and user community (people in general) for a specified length of time and then announce them as standard.

The members of the above committee may change with time; however, the committee itself needs to be planned as an ongoing, never-ending structure.

## 1.4 Promoting the Regular Use of Indian Languages by School Students

Each state must legislate that every QWERTY keyboard must have the local state script also printed or painted in the keys. Further, both the state educational departments and other interested organisations can conduct yearly competitions in the schools for fast typing in Indian languages using keyboards optimized for that language. By making the prizes attractive, we can easily ensure that the next generation is comfortable in typing in Indian languages using efficient key inputs, rather than using phonetic keyboards with input in Roman script.

## 1.5 Creating more Indian Language Web Content

Once again, attractive incentives may be given to school, college students as well as general public for creating quality web content in Indian languages. One simple way is to translate English Wikipedia content, as well as the content in Government tourism websites, etc. into the local language.

## 1.6 Role of Institution of Engineers, IETE, etc.

Professional bodies such as the Indian Academy of Sciences, Indian National Science Academy, Institution of Engineers (India), Institution of Electronics and Telecommunication Engineers can play a very significant role in promoting all the above. More importantly, they can announce awards for the best projects in undergraduate, postgraduate and even doctoral level work that create new applications in Indian languages, or promote the creation and widespread use of Indic language content. They must also play an important role in the policy making in these important areas.

## 2. Standardization of Fonts for Display Boards

With the existence of multiple languages, and with expanding travel of people for both business and tourism, there must be development of focused technology development, as well as standardization of different kinds, so that any traveler is facilitated to easily navigate through any place in spite of the display boards being in an unknown language or script. With camera captured document image analysis and recognition technologies, it is now possible to automatically detect and extract the text from images captured by a mobile phone. The suggestion is to standardize one or two fonts and even font sizes for any public display boards, so that the recognition engines can be optimized for these fonts and font sizes. By also standardizing the colour of the text and the background wherever possible, we can significantly increase the text recognition performance, making it a technology usable on a daily basis. The recognized text can then be translated and easily understood by the traveler. In many circumstances, most of the words (at least the key ones) in such boards will be proper nouns and hence, even a transliteration or transcription in the target script may suffice.

The same suggestion is also advanced for all the official printed documents of the government, which will guarantee exceptional accuracy of the respective OCRs, simplifying the process of digitization, editing or updating of existing

documents, wherein the source e-text is not easily available or accessible.

## 3. Create Specialized Institutes for Language Technology

New institutions must be created, whose mandate is to primarily research and develop new language technologies, and also train people for the industry.

### 3.1 Both Hardware and Software

These institutions will look at both hardware and software aspects. Hardware includes design of new devices, both primary and peripherals, such as input and output. For example, an extremely desirable device for a multilingual country such as India is a handheld, dedicated, handwriting input device, which will wirelessly transmit the recognized text to any computing device available nearby. The same universal device can be used for different languages by different people, by downloading the appropriate recognition engine, may be after a fee paid online. This has the potential to become as ubiquitous as the QWERTY keyboard, if not more.

### 3.2 Creating Computational Linguistic Studies

The computer science curriculum in all the engineering institutions must necessarily involve study of natural language processing and basics of computational linguistics. Also, the curricula for the different specialties in linguistics in universities must ensure basic training of the graduates in computer applications and computational linguistic tools. Research leading to up to Ph D degree in computational linguistics must be introduced in higher educational institutions.

## 4. Designing for Multilinguality

### 4.1 Website Design

All websites must necessarily be designed to be multilingual. The design specifications must take into account issues such as standardized Tables in each of the languages, wherever lists are involved, and provision for the users to quickly and periodically update such lists, without requiring technical experts to carry out such tasks. Once multilinguality becomes a basic, mandatory feature of websites, such an updation will be regularly required.

### 4.2 Design for Ease of Adding Another Language for User Interface

While the internal representations may be in English, the graphical user interface must be in the local language. In fact, it is preferable to make a provision for the user to customize the user interface to the language of his choice. This must be designed in a way that a technical expert is not required to add the new language interface.

### 4.3 Standardization of Terms for Lists

The terms for the most common lists that can occur in popular applications must be standardized for all the scheduled languages and a Table, giving the lists of corresponding terms across all the languages must be published and made openly available. This is because, in the past, there have been many efforts, both by individual

groups and many industries to come out with such terms in Indian languages, which led to the existence of multiple, unrelated words for the same object, adding to the confusion.

#### **4.4 Creating Multilingual Digital Dictionaries**

Structured digital lists containing equivalent verb roots, verb phrases, adjectives, adverbs, common nouns and noun phrases must be created, standardized and openly available for software developers. Laws must be created, which make it mandatory to use these standardized lists in all applications. These multilingual dictionaries will also form an extremely useful component of the machine translation. Further, they may ensure that information is not lost, when a sentence or even a document is translated between a number of successive pairs of languages.

### **5. Policy Changes Needed**

#### **5.1 Making Research Data Easily Available**

Every time an academic institute is funded to develop any aspect related to language technology, there must be a clear condition that the data collected and may be annotated as part of the funded project must be made available to researchers in some standard format, as soon as the project term ends, unless the researcher starts an industry or transfers the technology to an industry. In the latter case, there can be appropriate new policies that are applicable.

There could be a national level standing committee, which can look into these policies and the changes required from time to time.

#### **5.2 Publications Arising out of Government Funding being Freely Available for Researchers**

Just as special copyright laws exist in USA for Government funded research, exclusive copyright laws must be enacted, by which the publications arising out of any Government funded research must be openly available and/or the copyright must rest with the Government or the researcher. In any case, the idea is to ensure that the results of public funded research are readily available to the research community, without having to again pay to access them.

#### **5.3 Synergy between Multiple Institutions**

Currently, there are many planning, policy making and funding agencies, whose mandates or activities overlap; however, many of them operate in silos. This may lead to inefficient use of public money, and outcomes that fall short of desirable performance or quality standards or expectations. The national level standing committee must also look into such matters and advise the government as to how to bring in synergy between such agencies.

### **6. Conclusion**

Several suggestions have been made to the researchers, funding agencies and the Governments, which, in the opinion of the author, will go a long way in reaching the benefits of language technology to one and all.

There are a number of other things that need to be considered and this article is by no means exhaustive. However, if the spirit of the article is understood, one can

come up with meaningful suggestions to tackle each and every issue not addressed here.

### **7. Acknowledgements**

The author gratefully acknowledges Tata Trust Travel Grant for funding him to travel and participate in this conference. Immense thanks are also due to the Technology Development for Indian Languages (TDIL), Ministry of Information Technology, Government of India, for funding many of his projects in language technology, which has made it possible for him to be invited to this conference. Acknowledgment is also due to many students, research staff, colleagues, collaborators and interns, who enriched his knowledge and experience.