Reducing barriers of communication across Indian languages: An AI and ES approach to mass media

SHREESH CHAUDHARY

Department of Humanities & Social Sciences, Indian Institute of Technology, Madras 600036, India

Abstract. The present paper identifies some nonlinguistic and linguistic barriers that will have to be overcome by any system for automatic and simultaneous communication of news, commercial advertisements, and other items of information and entertainment by mass media across some Indian languages.

The paper also presents a brief account of some theories for the representation of knowledge of language in a language-independent manner, because such theories can make simultaneous communication of an item across different languages rather easy. But more research is required in this field before the relevant knowledge can be represented in a language-independent manner. Presently automatic and simultaneous communication of an item from any language to many other languages does not seem easy. However, it seems that in a very limited way a beginning can be made in the direction of such communication by human editors aided by tools developed by computer scientists so far.

Keywords. Attention-worthiness; barrier; communication; database; knowledge; language; mass media.

1. Introduction

Newspapers, radio and television, the three most widely used media of mass communication, presently take considerable time and human and other resources for preparing language specific editions of the same or similar stories, features and commercials. Efficiency, economy and the reach of these media can increase substantially if tools are made available to communicate simultaneously across many languages.

Rather than prepare a copy, for instance, in English first and then translate the same into different languages through human translators, it would be economical and fast if copies for different languages could be prepared at the same time, without requiring human translators at all. Such a system of mass communication would carry much more information and entertainment to many more people much faster and at cheaper rates, and would also expose many more customers to different mass media, than possible at present because of the constraints of language.

It seems that a system of communication of this kind is likely to encounter difficulties, some linguistic and some nonlinguistic. In other words, a news item, for instance, published or broadcast in one language may not be understood by readers or listeners in other languages either because of differences between the languages known and those available, or because of differences in the reader/listener's assumed knowledge of the background of any news item in different languages.

In the following sections, the present paper is going to analyse the nature of these barriers in such communication. The paper also identifies some issues that can be addressed by artificial intelligence and expert systems (AI and ES) technologies to overcome these barriers and make such communication across languages possible.

Section 2 describes the difficulties of communication in the mass media arising out of nonlinguistic differences among groups of customers. Section 3 describes some difficulties of communication attributable to linguistic differences, and to different conventions of use of each language for mass communication. The final section of the paper indicates issues to be addressed for overcoming these difficulties for mass media in communication across some Indian languages.

2. Nonlinguistic barriers

A news item published or broadcast in Tamil may be difficult to understand for non-Tamil knowing readers/listeners, not only because of their ignorance of Tamil, but also because these readers may not know as much about a DMK politician, for instance, as the Tamil news editor may assume.

The demands on the producers of mass media seem to vary from one political segment to another, or from one economic stratum to another, just as they vary from one language community of readers/listeners to another. This appears to be a much bigger problem than may be readily realized. It is normal for a tabloid paper like *The Daily Mirror* of Britain (and India has its own tabloids like *The Blitz*, *The Telegraph* etc. as well), for instance, to headline Iraq's defeat as 'GOTTCHA' when *The Times* of London may headline this item with something like 'Iraq surrenders'. Any AI and ES technology intended to remove the barriers in simultaneous and automatic mass communication across many languages has a greater challenge here.

Because of the varying perceptions of any (news) item, it seems that adapting *The Times* of London, for instance, for readers of *Le Monde* of Paris may be easier than adapting the former for the readers of *The Sun*, or *The Daily Mirror*, or *The News of the World*, even when *The Times* of London and the three tabloids mentioned here are all in English, and *Le Monde* is in French.

The point is that news items or other features and commercials in mass media in different languages do not have only linguistic differences. Differences among them are induced also by political, economic, social and cultural factors, all of which are more varied and unstable than linguistic differences.

Analysis of multiedition newspapers in the same language even in India shows that

news items can have varying degrees of reader-attention-worthiness, and therefore, there can be varying editorial priorities and practices. As examples below indicate, these editorial practices do have linguistic manifestations. But they seem to be more influenced by the political, cultural and other sociological features of the target customer group or groups of each of these mass media. Given in the examples below are some headlines from two English language newspapers published from the same city, i.e. Madras.

Example 1. Same city, same language, different newspapers.

(a) February 6, 1992

Indian Express

Headline (HL) 1: Cong-I, Kabul faction get one Assembly seat each
 2: Pak not to allow militants to cross Line of Control
 Anchor HL : PM stalls drug price decontrol

The Hindu

HL 1: Amrinder, Cong(I) man elected unopposed
 2: Pakistan Govt. to prevent marchers
 Anchor HL : Temple in Tirunelveli ransacked

(b) February 7, 1992

Indian Express

HL 1: As Pak steps up anti-India drive...
 Deal firmly with JKLF-march, govt tells para-military forces
 2: Militants determined to cross border
 Anchor HL TN annual plan outlay enhanced to 1,751 cr

The Hindu

HL 1: India asks Big Five to counsel Pakistan
 2: Police action in Punjab
 Poll panel rejects State explanation
 Anchor HL Plan outlay for T. Nadu increased

(c) February 9, 1992

Indian Express

HL 1: India Prepared to Meet Pak threat: Solanki
 2: PM Convenes Meeting
 Anchor HL 1: Judges, lawyers told not to attack each other
 2: Railways in a hurry to favour MNC

The Hindu

HL 1: India Ready to Meet Pak. Threat: Solanki
 2: Busloads of "marchers" on Pak. side of border
 Anchor HL Strains in Cong (I)-AIADMK ties remain

These differences can be found also between editions of the same newspaper published from different cities. For instance, some headlines from two English language newspapers published from two cities in India are reproduced below.

Example 2. Same paper, same language, different cities.

(a) *The Hindu* February 7, 1992

Hyderabad Edition

- HL 1: India asks big five to counsel Pakistan
 2: Police action in Punjab
 Poll Panel Rejects State Explanation
 Anchor HL Move to give Ayacutdars minor tanks

Madras Edition

- HL 1: (Same as in Hyderabad Edition)
 HL 2: (Same as in Hyderabad Edition)
 Anchor HL Plan outlay for T. Nadu increased

(b) *India Express* February 9, 1992

Hyderabad Edition

- HL 1: India fully prepared to meet any threat: Solanki
 PM convenes joint meet to discuss Kashmir issue
 2: Sikh ultras held in Delhi, Calcutta
 7 killed in Haryana train bomb blast
 Anchor HL Locomotive Purchase
 Rlys in a hurry to favour MNC

Madras Edition

- HL 1: India Prepared to Meet Pak threat: Solanki
 2: PM Convenes Meeting
 Anchor HL 1: Judges, lawyers told not to attack each other
 2: Railways in a hurry to favour MNC

Examples 1 and 2 above very clearly show that there are many nonlinguistic differences between the ways in which the same news item is presented in two newspapers of the same language in the same city, or in two editions of the same newspaper in different cities.

These nonlinguistic differences are not just stylistic differences. These differences can be classified broadly into the stylistic, lexical, and perceptual. Stylistic differences, for instance, include those like *Indian Express* (IE) writing 'Pak' without a dot, as in example 1(a). 2 and *The Hindu* (TH) writing 'Pak' with a dot, as in example 1(c). 1 and example 1(c). 2, for instance.

There are many other differences of this kind. For Tamilnadu, IE writes 'TN', as in 1b. anchor. But TH presents it as 'T.Nadu', as in 1b.anchor itself. As King (1987, pp. 5–6) observes, the human mind is such a powerful processing device that these differences will pass as insignificant minor differences in the mind. However, for a mechanical system, which necessarily requires explicit description of knowledge and processing techniques, these stylistic differences might be significant. For instance,

block capitals have been used for the first letters of the two words 'Big Five' in TH, Madras edition, (1b). Human mind can perceive its significance and the logic behind it readily. That, however, may not be so for any mechanical system. But this stylistic feature will continue to remain important for newspapers, though most Indian languages do not present problems of capital and small letters. That, however, only indicates a range of differences of this kind.

Examples 1 and 2 above indicate that media within the same language too, and more so across languages, tend to use different lexical items to indicate the same thing. Whereas IE has used 'not to allow', see (1a), TH has used 'prevent'. In the same headlines, one has used 'militants', whereas for the same people the other has used 'marchers'. In (1c), IE has used 'prepared', TH has used 'ready'. Thus even among the examples 1 and 2 above, there are numerous differences.

Differences in the perception of the same event or news item can be most baffling for any mechanical system. These differences seem to be governed more by the socio-politico-cultural features of the target group of consumers than by any other. For instance, the event of the unanimous election of two members to the Punjab assembly is reported differently in the two newspapers of the same language published from the same city. IE reports it as 'Kabul faction', the faction of the political party to which the elected member belongs. But TH reports it as 'Amrinder', by the name of the elected member. IE reports it as 'Cong-I, Kabul faction'. But TH changes it partly, and reports the item as 'Amrinder, Cong-I'. Differences of anchor headlines between IE and TH of February 9, 1992, as in example 1(c) above, can be explained only by attributing them to their different perceptions by the editors and consumer groups. For IE readers, the alleged irregularities in the purchase of locomotives by the railways seem important. To the readers of TH, the strain in the ties between Cong-I and AIADMK, the erstwhile political allies, may be of greater importance. There are many other differences of this kind.

Such differences can be found in news, features and commercials pertaining to the same item even in radio and television. Different regional stations may present different items in different ways, even when they may be using the same language. Radio news from All India Radio (AIR) whether in Hindi or Tamil or any other language is likely to differ from one regional station to another.

If there are such differences among media within the same language, they are sure to be more so among media across different languages. These nonlinguistic differences of perception arising out of varying **attention worthiness** of different news, features or other items may be the biggest challenge for any system of communication across different languages.

3. Linguistic barriers

A more clearly definable problem in the journalistic field pertains to linguistic differences among media. Problems here arise out of differences of script, phonetics and phonology, lexicon, syntax and semantics, and of journalistic conventions of language use.

In other words, a news item or feature published in the press, or broadcast on radio and television in one language may not be legible or comprehensible in another language either because the script or pronunciation used even for shared lexical items

and/or syntactic features may be different, or because the media may be using different lexical and/or syntactic features for communicating a similar piece of information.

3.1 *Script and pronunciation*

Hindi and Urdu newspapers offer a good example of the barrier of communication due to different scripts. Ordinary spoken-Hindi and spoken-Urdu have a significant amount of mutual intelligibility. But in the printed form, this intelligibility is totally lost. Hindi is generally printed in the Devanagiri script, and Urdu in the Arabic. These scripts belong to different script families, and have almost no common features in either the symbols or the diacritics used by them. Hindi, and to a considerable extent this seems true of Urdu as well, is intelligible to a larger variety of people when written in a script known to them. The use of Hindi in the Roman script by the armed forces in India until recently can be cited as an example in this context. Given a common script, Urdu and Hindi may appear to be dialects of the same language. But somehow, perhaps due to some accident of history, scripts have got inseparably associated with the popular identity of languages.

Radio and television have a similar problem at the phonological level. Many languages show dialectal differences in the pronunciations of many words which otherwise appear to be similar. The most obvious example of such a variation is the different pronunciations of the name of one of our ancient languages, namely *Sanskrit*. In some parts of India the name of this language is pronounced as *Samskrit*, and in some others as *Samskrut*. There are arguments to justify all these pronunciations and spellings.

Researches during the last three decades in the phonology of different regional varieties of English strongly suggest that the same language might become unintelligible to different groups of speakers because of differences of pronunciation. For detailed discussions of this problem, see Bansal (1969), Chaudhary (1989), and relevant references therein.

3.2 *Lexical items*

Among many languages in India there is much similarity of several classes of lexical items, like words for numbers, units of distance, weights and measures, directions, days, dates, months and many words of common historical origin. Yet even these items become unintelligible across language groups because of differences in pronunciation. Publications in the Common Vocabulary Series of the Central Institute of Indian Languages, Mysore, are good illustrations of this phenomenon of the similarity of certain classes of lexical items and their different pronunciations. For instance, we can see Rajaram (1973) for similar lexical items in Hindi and Tamil.

There are formidable barriers in communication across many Indian languages due to differences of lexical items too. In recent years, possibly under populist programmes, these differences have been deliberately emphasized.

For instance, lexical items used in mass media in Hindi and Urdu are, whenever possible, very different from each other even when use of the same or at least very similar items does appear to be possible in these languages. Promoters of the sectarian identity of Urdu seem to have encouraged the greater use of lexical items of Arabic and Persian origin. The media in Hindi seems to have opted for greater use of Sanskrit words. The following examples from news broadcasts by All India Radio (AIR) only indicate this tendency.

Example 3.

- (a) satrah sipaahiyon kii mrityu ho gayii
“seventeen soldiers of death be went”
Seventeen soliders were killed.
- (b) satrah sipaahii halaak ho gaye
“seventeen soldiers killed be went”
Seventeen soldiers were killed.
- (c) satrah sipaahii maare gaye
“seventeen soldiers killed went”
Seventeen soldiers were killed.

Example 3(c) is widely attested in nonspecialist use in both Hindi and Urdu speaking communities. But, for certain reasons including those of the cultural perceptions of the consumer groups, mass media, even government-owned media like the AIR, prefer to use 3(a) for Hindi and 3(b) for Urdu.

Some examples of lexical differences for the same item from some Indian language newspapers are given in examples 4 and 5 below.

Example 4.

- (a) kendriyo sarkarer acomko siDDhaanTo
- (b) kendriiy sarkaar kaa achanak nirNay
- (c) keendriiya prabhutwapu aakasmika nirNayam
- (d) mattiya arasu tiDiir muDivu
- (e) “central government’s sudden decision”

Example 5.

- (a) panjaab, aasome widhaan sabhaa o lok sabhaa nirwaacana
- (b) panjaab, asam menbhii widhaan sabhaa, lok sabhaa cunaaw
- (c) panjaabu assaamulaloo widhaana sabhaa look sabhaa ennikalu
- (d) panjaab, asaamilum makkaLawai saTTapeerawai terdalhaL
- (e) “Punjab, Assam (in) (also) legislative assembly (and)
House of People election”
State Assembly and Parliament elections in Punjab and Assam also.

The headline in 4(a) is possible in any Bengali language newspaper, 4(b) in Hindi, 4(c) in Telugu, and 4(d) is from the Tamil newspaper *Dinamani*, Madras. Headlines have been given in the same order of languages in example 5 also.

We can see that among the four languages given here that there are varying degrees of lexical similarity, with Tamil perhaps being the most different from the other three. There are one or two items which are different in all these languages. For instance ‘decision’ in example 4 has different words in all of these languages. In example 5, “election” has different words in all of these languages.

Similar differences of lexicon have increased in recent years in many Indian languages, and media in all these languages have been under pressure to abandon common words for those perceived as specific to the cultural identities of their readers

and listeners. These differences can be seen clearly in the language specific translation of words and phrases borrowed from English and commonly used in many Indian languages. Some examples of differences in words of this kind have been given in example 6 below.

Example 6.

Hindi	Tamil	Gloss
Patraalaya	Tapaalnilayam	Post office
Duurdarshan	TolaikaTci	Television
Duurbhash	Tolaipeesi	Telephone
Betaar	Tandi	Telegram
Wimaanpattan	WimaanatLam	Airport
Kendriiya sarkaar	Mattiya arasu	Central Govt.

Thus, words of even common historical origin are getting localised, and to that extent, communication across languages even of the same family is becoming difficult.

3.3 Syntax

Another difficulty in communication of this kind appears to be there in the differences of syntactic features of various languages and in differences of conventions of use of these features in the media. It is well known that India has several groups of languages belonging to several language families which are significantly different from one another in their typological features.

There are, for instance, Naga, Mizo and some other languages of the North-East. Many of these languages are mutually unintelligible even among themselves. Then there are languages like Kashmiri, Laddakhi and a few others in the North-West which are very different from other Indian languages in their typological features. There are languages like Mundari, Oraon, Kurukh, Santhal, and a few others spoken by the so-called tribal communities in the plateau of Central India which have little in common with even other tribal languages.

Among many languages listed in the eighth schedule to the Indian Constitution, such as Bengali, Hindi, Marathi, Tamil, Telugu etc. however, there is a considerable amount of typological similarity. They are, for instance, mostly of the subject-object-verb (SOV) type, unlike, for instance, an SVO type language like English. Many of these languages also have grammatical features of number, gender, person, case, honorifics and tense, though the exact forms of these features differ significantly.

In Telugu, for instance, case is realized through inflexions, whereas in Hindi it is realized through lexical items. The phenomenon of grammatical gender in Hindi/Urdu is much more pervasive than in other Indian languages. Many other Indian languages do not have this phenomenon of grammatical gender. We have already seen some instances of these differences in examples 4 and 5 above. These differences are significant barriers for any medium of mass communication across these languages.

3.4 Journalese

In addition to lexical and syntactic differences among these languages, mass media in each of these languages also have specific conventions of language use, with their own syntactic and semantic features. From examples 4 and 5 above, it seems that some newspapers in Indian languages prefer nominal and nonfinite forms to verbal or finite ones for headlines and the first lead. But, for the second lead, their preferences for finite and nonfinite forms differ, as may be seen in the following examples from Bangla, Hindi and Tamil, in that order in each.

Example 7.

- (a) ameerikaar tattaawdhaane aalocnaa
- (b) ameerikaa kii dekhrekh men honii caahiye
- (c) amerikkaa meerpaarwayil naDakka weeNDum
- (d) “America (of) auspices (in) (should be) (conference)”
(Conference) should be under American auspices

Example 8.

- (a) arab raashtro shonge shorashori aalocnaar jonne izraael ekmoT
- (b) arab raashtron se siiDhii baaTciiT ke liye izraael sahmaT
- (c) arabu naaDuhaLuDan neeraDi peeccu isreel sammadam
- (d) “Arab nations with direct talk Israel agreement”

Israel Agrees for Direct Talks with Arab Countries

Whereas English language newspapers in India show almost an equal preference for nonfinite and finite forms, simple present tense being their preferred finite form, in Indian languages the situation in this regard does not seem to be so clear. Newspapers in these languages demand at least one of the two headlines, and there usually are two headlines for many items, in the finite form, usually the simple past.

Hence, a system intending to facilitate communication across different languages faces a two-fold problem: first, deciding about which headline has to have a finite form, and then converting the finite or nonfinite form of headline in one language to a finite or nonfinite form in another.

Another problem for mass media in communication across languages is in the different linguistic conventions for journalistic and other references to sources of information. In many Indian languages there are already standard phrases for such journalistic expressions as – reportedly, allegedly, according to informed sources, according to reports reaching here, according to a spokesman, a spokesman/person said etc.

But not all languages have the same conventions of using these phrases in the beginning, the end or the middle of a sentence, as the following examples indicate:

Example 9.

- (a) It (Union Government) has left it to the Election
Commission to decide the dates, a spokesman of the
Home Ministry said tonight.

- (b) eko sorkaarii probokTaa moTe nirbaacon komiishan
 “a govt spokesman said election commission”
 eijonne nirbaaconer Diin ghoshnaa korben
 “for this election for dates announce will”
- (c) ek sarkaari prawakTaa ne baTaayaa ki cunnaw aayog
 “a govt spokesman said that election commission”
 ko iske liye TiThiyaan nirDhaariT karne ko kahaa gayaa hai
 “to this for dates fixed do to said has been”
- (d) idarkaana tedihaLai teerdal kamishan muDiwu seyyum enru
 “for this dates election commission decide will make that”
 arasin peeccaaLar oruwar teriwittaar
 “of govt spokesman one said”

The example in 9(a) is from the English newspaper *The Hindu*, Madras, and the one in 9(d) is from the Tamil daily newspaper *Dinamani*, Madras. Those in 9(b) and 9(c) are possible in Bangla and Hindi newspapers. These examples indicate that newspapers, and possibly other media too, in many Indian languages do not follow the same conventions for acknowledging their sources of information.

To summarise then, the following are some of the major issues that any system for simultaneous communication across some Indian languages may have to encounter.

Nonlinguistic

- (a) Varying attention-worthiness of items according to groups of readers/listeners/viewers.
- (b) Varying amounts of complementary information required for different items according to groups of readers/listeners.

Linguistic

- (a) Different scripts for the same language.
- (b) Different pronunciations of the same item.
- (c) Different lexical items for the same thing.
- (d) Different rules of sentence construction.
- (e) Different conventions of language use in mass media.

Overcoming these difficulties in simultaneous communication across several languages may not be easy in the near future. However, progress in Artificial Intelligence and Expert Systems and in other fields of computer science indicate that some success can be achieved with the proper use of tools developed so far.

4. Knowledge of language

The solution to the problem of communication across several languages appears to depend upon the description of what constitutes the knowledge of language. It appears that application of this knowledge may presuppose a description and quantification

of this knowledge, possibly in machine representable form. But what kind of information constitutes knowledge of language?

Can we say that the knowledge of words, letters and sounds in them, their meanings and rules of usage, constitutes the knowledge of language? Or, can we say that to know the rules of sentence construction, as given in the grammar book of any language, constitutes the knowledge of language? Or, can we say that this knowledge is of a very abstract kind not specific to any language but is a common property of the human mind which enables nearly all human beings to learn at least one language?

Since the late 1950s, Noam Chomsky has pioneered a new approach to the study of the structure of language and of how the knowledge of language seems to be acquired and organised. Chomsky (1981, p. 4) argues that knowledge of language essentially comprises a knowledge of its rules and representations. It is this body of rules and representations, according to Chomsky (1986, pp. 12–13), which “relates sound and meaning and assigns structural properties to physical events in certain ways, not others”.

Chomsky (1986, pp. 12–13), therefore, believes that “We should think of knowledge of language as a certain state of the mind/brain, a relatively stable element in transitory mental states once it is attained; furthermore, as a state of some distinguishable faculty of the mind – the language faculty – with its specific properties, structure and organisation, one ‘module’ of the mind”. For a detailed description of this view of the knowledge of language, popularly known as ‘Generative Linguistics’ or ‘Generative Grammar’, we may see Chomsky (1981, p. 4, 1986, pp. 12–13) and other references.

This view of language may be very convenient for any system of application as it presupposes only one underlying representation for all languages of the world. It also assumes that representation of knowledge of language is not language-specific. To take a crude example, it is like representing the knowledge of an English word like cat with a picture of a four-legged creature. Such a representation enables easy and instant access to knowledge in a very language-free manner. Given such a representation of knowledge, any system for communication across languages will reduce all texts from any language to this form of knowledge, i.e. a visual form, and then generate the same text through any number of desired languages.

During the last 30 years, many languages have been described within the framework of generative linguistics, and a considerable body of knowledge about natural languages has been built up. But these works are so lacking in some crucial details that they are considered to be of limited use for engineering in natural languages, or for application of this knowledge for any practical purpose.

For instance, the entire phonological component of the generative linguistics is based on a presupposition that there is a discrete unit of sound, called *phoneme*, even in continuous speech such that it is distinct from other such units and can always be isolated. Researches in segmentation of continuous speech, however, have shown that it is extremely difficult, or nearly impossible, to indicate the particular point where one sound ends and the other begins in a normal continuous utterance,

The human mind nevertheless recognises some discrete sound units, irrespective of variations and other physiological factors. But no reliable description seems to be available so far of how exactly this recognition takes place. For want of such a description, studies in the field of speech synthesis and similar fields of engineering and application of knowledge of speech have used a term like ‘diphone’ which includes adjacent parts of two discrete sounds (see Klatt 1987). Consequent upon our inability

to identify the phoneme, studies in speech engineering have also used clusters of sounds upto the maximal unit of a syllable for identifying discrete sound sequences. Works like the ones reported in Yegnanarayana *et al* (1991, pp. 467–475) make use of a large number of sound sequence units, nearly 350, for speech synthesis in Hindi, whereas any language requires typically about 45 sound units (phonemes).

The approach which codifies knowledge of speech through diphones and/or discrete representation of different syllables may be useful from the engineering and application points of view, but appears somewhat *ad hoc* theoretically and also cumbersome in practice. It simply ignores the question of quantifying the knowledge that enables the human mind to recognise the beginning and the end of one sound from those of another, or recognise the presence of other sounds in any other utterance. From an engineering point of view also such an approach does not seem to be efficient as it increases the processing time and cost. For a discussion of engineering issues in this field see Klatt (1987).

In recent years, however, there have been attempts to write rules and representations in machine usable form for speech understanding and natural language processing. Ostensibly in the form of conventional parsing rules, these rules and representations are logically much more constrained, have mathematical precision, and describe many abstract concepts with an accuracy not available earlier. For illustration, we shall refer to two of these attempts here.

One of these, first reported in Schank (1973, pp. 45–6 & 192–4), has to do with the ‘Identification of conceptualizations underlying natural languages’. The work reported here addresses itself to the problem of the ‘representation of meaning in an unambiguous language-free manner’. Schank believes that ‘there exists a conceptual base into which utterances in natural languages are mapped during understanding’. So, for instance, Schank (1973, pp. 45–6 & 192–4) suggests that we can refer to a nominal concept as a “Picture Producer (PP) such that it is the concept of a general thing, a man, a duck, or the Grand Canyon... An action (ACT) is what a nominal can be said to be doing. For a concept to qualify as an action it must be something that an animate nominal can do to some object... the heart of any conceptualization is the relationship between the actor and the action in the event”. Following this approach, Schank can paraphrase the difference between the following pairs of sentences:

Example 10.

- (a) John ate the steak with the odor.
- (b) John ate the steak with the fork.

Example 11.

- (a) John shot the girl with a rifle.
- (b) John shot the girl with long hair.

For details, we can see Schank (1973, pp. 45–6 & 192–4). However, this approach to machine understanding of natural languages has its own limitations. It seems to have limited efficiency with sentences which refer to abstract concepts, or to information preceding or following in the discourse. For instance, it may be difficult for this approach to paraphrase the difference in the semantic roles of John in the following sentences:

Example 12.

- (a) John is easy to please.
- (b) John is eager to please.

In 12(a) John is easy for someone else, not mentioned in the sentence, to please. In 12(b) John is eager to please someone else, not mentioned in the sentence. All these may require extra-sentential knowledge along with knowledge to paraphrase abstract concepts like "please" in a language-free manner.

There are other kinds of difficulties with the identification of conceptualizations underlying natural languages. These difficulties pertain to, as mentioned by King (1987), the knowledge specific to the culture of any language community. Some of these difficulties also stem from idioms, unfinished sentences, or assumptions and such other deformities of discourse in any natural language. It seems that these deformities account for a greater part of natural language use. Nevertheless, the approach indicated in Schank (1973, pp. 45–6 & 192–4) holds considerable promise.

Goodman & Reddy (1980, pp. 234–46) in a later work of this kind are more specific about a possible model for "Alternative control structures for speech understanding systems". They assume that "control structures are an essential part of any speech recognition system. They are the devices by which passive knowledge about the task and language is transferred into active and effective processes".

Goodman & Reddy (1980, pp. 234–46) assume that control structure may be divided into three areas: knowledge source interaction, knowledge source activation, and knowledge source focussing. The sources of knowledge (KS), according to them, include the characteristics of speech sound (*phonetics*), variability in phonetics (*phonology*), the stress and intonation patterns of speech (*prosody*), the sound patterns of words (*lexicon*), the grammatical structure of language (*syntax*), the meaning of words and sentences (*semantics*), and the context of the conversation (*pragmatics*).

Goodman & Reddy (1980, pp. 234–46) then go on to propose various models of KS interaction, activation and focussing. For a model of natural language understanding in a language-free manner their proposal seems to be very close to the model of universal grammar assumed in generative linguistics in Chomsky (1981, p. 4, 1986, pp. 12–13), and elsewhere. If and when this model of representation of knowledge succeeds, it should be possible to reduce the text into language-free representations and then generate the same text in any number of desired languages.

There, however, are certain areas of natural language understanding where this model may not be very helpful. Most difficulties of this model have to do with inter-componential knowledge. It may not be easy, for instance, to decide if word stress pertains to lexicon or prosody, or whether case inflexions pertain to syntax, lexicon or phonology etc. It may be difficult within this model to say if the phenomenon of stress shift in words like photograph, photography and photographic can be decided by prosody, or by phonology, or by lexicon. This model does not appear to be adequate for understanding of overlapping phenomena of the kinds mentioned here.

Ignoring certain gaps of detail in the generative linguistics point of view of the knowledge of language, it seems logical to say, that a multilingual person creates a module, or a system of knowledge, which includes knowledge of different languages and the knowledge of different parts of these languages, such as phonetics, phonology, morphology, lexicon, syntax, semantics and pragmatics, and their mutual relationships, as subsystems or as sub-subsystems.

In this module all subsystems, such as those of sound, word, phrase, sentence and

discourse structure, are related to one another. Hence, given an utterance of any size, even with distortion, it seems that the module of rules and representations built by the multilingual person is immediately activated, and then the module initiates multilateral processing of the linguistic data most of the time to generate an appropriate verbal or nonverbal response. This module is so efficient that it handles even complex input correctly and quickly.

To be able to recognise and use different languages automatically, like a multilingual human being, a computer or any other potentially intelligent device will require multiple, such as visual, auditory, olfactory, and other sensory and nonsensory channels of data input, a large, though adequately constrained, lexicon, and inferential and evaluative rules of global reach for data processing. In the present state of knowledge and engineering in this field it does not seem possible.

A fair idea of difficulties in such a task can be had from the account by Steiner *et al* (1988) of the progress of the European Community supported project of simultaneous machine translation in eight European languages including English, French, German, Italian, Portuguese and Spanish.

Engineering in the field of neural networks (see Yegnanarayana 1991, pp. 103–29, for instance), however, seems to be a promising development in this area. But it may still be many years before we can justifiably place any hopes on it. To be able to match the computing powers of the human mind, it seems that a machine will not only require multiple channels for receiving data, such as visual, auditory etc., but also multitier simultaneous processing. Such a machine will also require the power of nonlinear, nonsequential and fuzzy-logic computation for processing the data and then for responding to it in an appropriate manner, occasionally on the basis of instincts and intuition like human beings. Clearly, computers of the power required for this purpose are not in sight yet.

So what can be done to facilitate communication across different languages through mass media at the moment? It appears that presently human editors using the available AI and ES technology can to some extent accomplish simultaneous communication of an item across several languages. An outline of such a solution is given in the following paragraphs.

5. Towards a solution

5.1 *Nonlinguistic barriers*

In the present state of the art, human editors seem indispensable for deciding what item of information or entertainment may be of interest to which group of customers. But once they decide on such items, providing complementary information about these items can be made significantly less difficult by the use of databases for storing, ordering and retrieving data on any given item in any given language.

Editors in the mass media for cross-linguistic communication may create and maintain some kind of a multilingual database on people, places, events, clientele etc. Monolingual databases are already in use by large business and public organisations for storing and retrieving information about their personnel, operations, clients etc. Multilingual databases are also possible now. Databases can be built and maintained in as many languages as the particular media organisation may wish to cater to.

These databases may be created on the basis of the frequency of the occurrence

of an item in a given language. If, for instance, a particular name occurs quite frequently in a language, then the amount of complementary information required for that item in that language is likely to be smaller than the amount of information required for this item in a language where it does not occur so frequently. To take a concrete instance, let us say that the name of a politician called Yadav is mentioned frequently in Hindi news media for clients in Bihar, but does not occur so frequently in the Tamil news media for clients in Tamilnadu. It is then logical that the amount of complementary information about Yadav in Tamil should be larger than that for him in Hindi. This can be a good working principle for editors to create their databases on a large number of items, i.e. people, places, events etc., and to use them accordingly.

Therefore, once the human editor decides to take up a particular news item in a particular language, all that he may need to do is to retrieve the information stored in the database about that item and put it in the script which can be published or broadcast as appropriate. With the phenomenally large storage capacity of modern computers, it may not be a problem to build databases of different kinds and sizes on a large number of items in several languages. This, however, may only be a tool which can enhance the efficiency and speed of the human editor in the goal of simultaneous communication of information or entertainment through several languages.

5.2 Linguistic barriers – difficulty of differing scripts

There are languages like Hindi and Urdu where communication is impeded largely due to different scripts being used for writing. In the spoken mode, these languages are mutually intelligible to their speakers. Therefore, with the machine transcription of text from the Devanagiri to the Arabic script, or vice versa, mutual intelligibility can be extended between these languages to their printed mode as well. Unfortunately, the arrangement of letters and diacritics in these scripts do not match. Yet, with the state-of-the-art in computer science, it should not be impossible to design an expert system for machine transcription of texts with human participation wherever necessary.

In the other major Indian languages, fortunately, there is a large extent of similarity between the arrangement of letters and diacritics so that it should be relatively easy to obtain such transcriptions among these languages. For instance, the arrangement of letters and diacritics in the scripts of Telugu and Kannada has a great degree of similarity. If, somehow, the text in one language can be transcribed into the other, a great deal of intelligibility can be obtained for the reader of the target language in this way as well. A similar facility can be created through transcription of texts between Hindi and Punjabi, between Bengali and Oriya, between Tamil and Malayalam etc.

Bilingual keyboards are already in use. All we need is a technology to identify and match letter and diacritics.

5.3 Difficulty of different lexical items & pronunciations

Sometimes the same lexical item with the same meaning also becomes unintelligible to users of related languages in India because of the different pronunciations of the particular lexical item in these languages. As an example we saw earlier how the name of one of the ancient languages of India is pronounced as 'Sanskrit', 'Samskrit' or 'Samskrut'. There are large number of lexical items, such as numbers, names of days, dates, months, distance, directions etc., which differ in similar ways. Overcoming

these difficulties of differing pronunciations can also make simultaneous communication across several languages considerably easy.

Advances in multimedia technology can be of use here. They are already being used for educational purposes in European countries. In India too, language-specific pronunciations of frequently occurring lexical items can be stored in laser disc-based multimedia systems, and can be used when needed both by the print and the voice media. This approach to lexical differences among related languages can greatly reduce the barriers of simultaneous communication across several languages.

For instance, language-specific lexical terms for the word 'five' can be stored in the multimedia memory, both in print and in voice. This can be retrieved by the human editor for print or voice media according to need. Building such databases no longer appears impossible. Similar things are already being provided in terms of value-added services by personal computer manufacturers. As we have seen earlier, lack of mutual intelligibility at least among major Indian languages is not so much due to differences of syntax as due to differences of lexical items. A database of lexical items, or an electronic dictionary, so to say, of languages, with cross language matching and referencing of this kind, especially in multimedia as suggested here, can greatly enhance the mutual intelligibility among these languages, and also facilitate simultaneous communication across them.

5.4 *Differences of syntax*

We have seen earlier that the major Indian languages do not have significant syntactic differences. They are all of the subject-object-verb type, the phrases in the sentences of these languages are accordingly arranged, and almost a perfect match can be obtained between sentences in these languages.

The major difference, however, lies in the way case marks are shown in the phrases of these languages. In Hindi, for instance, case marks are generally shown through discrete lexical items, whereas in many other major India languages case marks are shown through rule-governed inflections. This is a major problem and a hurdle between simultaneous and automatic communication of an item across these languages.

If the machines can identify the boundaries of a phrase/word, then substituting that phrase/word in one language with the corresponding word/phrase from the target languages(s) cannot be very difficult, especially with the database of lexical items mentioned above, and human editors participating in it. Bharati *et al* (1991, pp. 277-93) and Rao & Yegnanarayana (1991) have also shown that computer technology can be used for identifying word boundaries in Hindi. Once such recognition takes place, the human editor will only have to find the corresponding item from the target languages(s) with help of the lexical database, and make appropriate substitutions. As King (1987, pp. 5-6) has shown, something similar is already being done by the Canadian Radio Organisation for making bilingual weather forecasts. In this manner, computerised databases and laser disc-based multimedia can become effective tools in the hands of the human editor in his objective of achieving simultaneous communication across several languages.

5.5 *Difficulty of different journalistic conventions*

Evolution of journalistic conventions in different languages has been influenced by many factors. Of these, two very important factors are the cultural perception of the

community using that language, and the constraints of space and time on the human editors of the media in that language at any given time.

It seems entirely a matter of cultural perception why for instance, *Union Government* is called *Central Government* (Kendriya Sarkar) in Hindi, whereas it is *Middle Government* (Matthiyaarasu) in Tamil. These terms can be stored and retrieved in a database in the manner stated above for each of the different languages used by any mass media organisation. A human editor will still have to participate and conduct the search, but the electronic dictionary mentioned above can greatly facilitate this search and make cross-linguistic communication less difficult and less expensive.

What, however, cannot be accomplished so easily is the conversion of journalistic items like sources of information from one language to another. For instance, it cannot be decided *a priori* which of the following should be used.

Example 13.

- (a) a spokesman of the government said
- (b) according to a spokesman of the government

The choice of 13(a) or (b) will depend on the context and the space/time available with the editor. In a business where every inch of space and every second of time counts, it may be of vital importance to the editor to decide which of the competing terms will be appropriate in the given context. But once the human editor decides to use a term of a certain length and kind, then he can look for matching terms in the database of the particular language with him, and use the most appropriate term. Many other decisions with regard to journalistic conventions will also have to be made in a similar manner.

6. Conclusions

Automatic and unrestricted simultaneous communication across languages through the mass media does not seem possible presently in view of some of the more formidable barriers mentioned in this paper. With the knowledge available in computers and in the AI and ES fields so far, certain kinds of communication across some Indian languages seem possible even now. Many (news) items, features and commercials involving only names, numbers, dates, days, months, and such other items can, even in the present state of knowledge, be communicated across many Indian languages through the mass media. In other words, the mass media can even now communicate weather forecasts, election results, market rates, company reports, railway, airline and radio and television timetables etc., across several Indian languages simultaneously. Copies of many commercial advertisements for different media can also be prepared in many languages simultaneously. As King (1987, pp. 5–6) reports, with progressively diminishing human participation, Canada has been having simultaneous weather forecasts in French and English from the late nineteen seventies.

I am grateful for the help I have received from Prof B Yegnanarayana and from an anonymous reviewer in the preparation of this article. However, I alone am responsible for mistakes, if any, here.

References

- Bansal R K 1969 *Intelligibility of Indian English*. Monograph of the Central Institute of English & Foreign Languages, Hyderabad
- Bharati A, Chaitanya V, R Sangal 1991 Local word grouping and its relevance to Indian languages. In *Frontiers in knowledge based computing* (eds) V P Bhatkar, K M Rege (New Delhi: Narosa)
- Chaudhary S C 1989 *Some aspects of the phonology of Indian English* (Ranchi: Jayswal)
- Chomsky N 1981 *Lectures on Government and binding* (Dordrecht: Foris)
- Chomsky N 1986 *Knowledge of language* (New York: Praeger)
- Goodman G, Reddy R 1980 Alternative control structures for speech understanding systems. In *Trends in speech understanding systems* (ed.) W A Lea (New York: Prentice Hall)
- King M 1987 *A tutorial on machine translation* (Geneve, Suisse: ISSCO)
- Klatt D H 1987 Review of text to speech conversion for English. *J. Acoustic Soc. Am.* 82: 737–793
- Rajaram S 1973 *Hindi–Tamil common vocabulary: CIIL Common Vocabulary Series – 3* (Mysore: Central Institute of Indian Languages)
- Rao G V R, Yegnanarayana B 1991 Word boundary hypothesization in Hindi speech. *Comput. Speech. Language* 5: 379–392
- Schank R C 1973 Identification of conceptualizations underlying natural languages. In *Computer models of thought and language* (eds) R C Schank, K M Colby (San Francisco: W H Freeman)
- Steiner E H, Schmidt P, Zelinsky-Wibbelt C 1988 *From syntax to semantics: Insights from machine translation* (London: Pinter)
- Yegnanarayana B 1991 Neural Networks. In *Artificial intelligence & expert system technologies in the Indian context* (ed.) V V S Sarma (New Delhi: Tata-McGraw Hill) vol. 2
- Yegnanarayana B, Murthy H A, Sundar R, Ramachandran V R, Madhukumar A S, Alwar N, Rajendran S 1991 Development of text to speech system for Indian languages. In *Frontiers in knowledge based computing* (eds) V P Bhatkar, K M Rege (New Delhi: Narosa)