S

India and its Languages

Version:	1.0
Date:	2004-01-23
Restriction:	Public
Author:	Herbert S. Tropf Siemens AG, Munich

Executive summary

Commercially India is one of the most interesting markets in Asia, and linguistically it is most challenging because of the number of languages spoken. Speech enhanced products applying local languages will be sold very soon. So far, practically no speech databases for India are available.

After sketching the economy (section 1), the languages (section 2) and language resources centers (section 3) of India some features are proposed for the creation of Indian speech databases (section 4).

S

Contents

S

1 The country

1.1 Historical background

The Indus Valley civilization,* one of the oldest in the world, goes back at least 5,000 years. Aryan tribes from the northwest invaded about 1500 B.C.; their merger with the earlier inhabitants created the classical Indian culture. Arab incursions starting in the 8th century and Turkish in 12th were followed by European traders, beginning in the late 15th century. By the 19th century, Britain had assumed political control of virtually all Indian lands. Nonviolent resistance to British colonialism under Mohandas GANDHI and Jawaharlal NEHRU led to independence in 1947. The subcontinent was divided into the secular state of India and the smaller Muslim state of Pakistan. A third war between the two countries in 1971 resulted in East Pakistan becoming the separate nation of Bangladesh. Fundamental concerns in India include the ongoing dispute with Pakistan over Kashmir, massive overpopulation, environmental degradation, extensive poverty, and ethnic and religious strife, all this despite impressive gains in economic investment and output.

1.2 Geography

Fig. 1: Map of India (with major cities):

Area:
total: 3,287,590 sq km
land: 2,973,190 sq km
water: 314,400 sq km

* Editorial remark: Where not indicated otherwise, the material of section 1 comes from [http://www.cia.gov/cia/publications/factbook/geos/in.html]

1.3 People

Population: 1,049,700,118 (July 2003 est.)

Age structure:
0-14 years: 32.2% (male 173,973,350; female 163,979,116)
15-64 years: 63% (male 342,620,712; female 319,259,867)
65 years and over: 4.8% (male 25,281,756; female 24,585,317) (2003 est.)

Life expectancy at birth:
total population: 63.62 years; male: 62.92 years; female: 64.37 years (2003 est.)

Ethnic groups: Indo-Aryan 72%, Dravidian 25%, Mongoloid and other 3% (2000)

Religions: Hindu 81.3%, Muslim 12%, Christian 2.3%, Sikh 1.9%, other groups including Buddhist, Jain, Parsi 2.5% (2000)

Languages: English enjoys associate status but is the most important language for national, political, and commercial communication; Hindi is the national language and primary tongue of 30% of the people; there are 14 other official languages: Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Malayalam, Kannada, Oriya, Punjabi, Assamese, Kashmiri, Sindhi, and Sanskrit; Hindustani is a popular variant of Hindi/Urdu spoken widely throughout northern India but is not an official language

Literacy:
definition: age 15 and over can read and write; total population: 59.5%; male: 70.2%; female: 48.3% (2003 est.)

S

Table 1: State wise population totals with percentage of Urban population (Census of India 2001)

	State	*Population*	*Urban*
1	Uttar Pradesh	166.052.859	20,8
2	Maharashtra	96.752.247	42,4
3	Bihar	82.878.796	10,5
4	West Bengal	80.221.171	28,0
5	Andhra Pradesh	75.727.541	27,1
6	Tamil Nadu	62.110.839	43,9
7	Madhya Pradesh	60.385.118	26,7
8	Rajasthan	56.473.122	23,4
9	Karnataka	52.733.958	34,0
10	Gujarat	50.596.992	37,4
11	Orissa	36.706.920	15,0
12	Kerala	31.838.619	26,0
13	Jharkhand	26.909.428	22,3
14	Assam	26.638.407	12,7
15	Punjab	24.289.296	34,0
16	Haryana	21.082.989	29,0
17	Chhatisgarh	20.795.956	20,1
18	Delhi	13.782.976	93,0
19	Jammu and Kashmir	10.069.917	24,9
20	Uttaranchal	8.479.562	25,6
21	Himachal Pradesh	6.077.248	9,8
22	Tripura	3.191.168	17,0
23	Manipur	2.388.634	23,9
24	Meghalaya	2.306.069	19,6
25	Nagaland	1.988.636	17,7
26	Goa	1.343.998	49,5
27	Arunachal Pradesh	1.091.117	20,4
28	Pondicherry	973.829	66,6
29	Chandigarh	900.914	89,8
30	Mizoram	891.058	49,5
31	Sikkim	540.493	11,1
32	Andaman & Nicobar Islands	356.265	32,7
33	Dadra & Nagar Haveli	220.451	22,9
34	Daman & Diu	158.059	36,3
35	Lakshadweep	60.595	44,5
	INDIA Total	1.027.015.247	27,78

[Source: Census of India 2001: http://cyberjournalist.org.in/census/cenindia.html]

1.4 Governmental features

Country name:
conventional long form: Republic of India
conventional short form: India

Government type: federal republic

Capital: New Delhi

Administrative divisions:

28 states and 7 union territories*: Andaman and Nicobar Islands*, Andhra Pradesh, Arunachal Pradesh, Assam, Bihar, Chandigarh*, Chhattisgarh, Dadra and Nagar Haveli*, Daman and Diu*, Delhi*, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu and Kashmir, Jharkhand, Karnataka, Kerala, Lakshadweep*, Madhya Pradesh, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Orissa, Pondicherry*, Punjab, Rajasthan, Sikkim, Tamil Nadu, Tripura, Uttaranchal, Uttar Pradesh, West Bengal.

Major cities are shown in Fig. 1 above.

Fig 2: States and union territories of India

Fig. 3: Flag of India

1.5 Economy

Economy - overview: India's economy encompasses traditional village farming, modern agriculture, handicrafts, a wide range of modern industries, and a multitude of support services. Overpopulation severely handicaps the economy and about a quarter of the population is too poor to be able to afford an adequate diet. Government controls have been reduced on imports and foreign investment, and privatization of domestic output has proceeded slowly. The economy has posted an excellent average growth rate of 6% since 1990, reducing poverty by about 10 percentage points. India has large numbers of well-educated people skilled in the English language; India is a major exporter of software services and software workers; the information technology sector leads the strong growth pattern. The World Bank and others worry about the continuing public-sector budget deficit, running at approximately 10% of GDP in 1997-2002. In 2003 the state-owned Indian Bank substantially reduced non-performing loans, attracted new customers, and turned a profit. Deep-rooted problems remain, notably conflicts among political and cultural groups.

GDP: purchasing power parity - $2.664 trillion (2002 est.)

GDP - real growth rate: 4.3% (2002 est.)

GDP - per capita: purchasing power parity - $2,600 (2002 est.)

GDP - composition by sector:
agriculture: 25%
industry: 25%
services: 50% (2002 est.)

Population below poverty line: 25% (2002 est.)

Industries: textiles, chemicals, food processing, steel, transportation equipment, cement, mining, petroleum, machinery, software

1.6 Communicatons

Telephones - main lines in use: 27.7 million (October 2000)

Telephones - mobile cellular: 2.93 million (November 2000)

Telephone system:
general assessment: mediocre service; local and long distance service provided throughout all regions of the country, with services primarily concentrated in the urban areas; major objective is to continue to expand and modernize long-distance network to keep pace with rapidly growing number of local subscriber lines; steady improvement is taking place with the recent admission of private and private-public investors, but, with telephone density at about two for each 100 persons and a waiting list of over 2 million, demand for main line telephone service will not be satisfied for a very long time
domestic: local service is provided by microwave radio relay and coaxial cable, with open wire and obsolete electromechanical and manual switchboard systems still in use in rural areas; starting in the 1980s, a substantial amount of digital switch gear has been introduced for local and long-distance service; long-distance traffic is carried mostly by coaxial cable and low-capacity microwave radio relay; since 1985 significant trunk capacity has been added in

the form of fiber-optic cable and a domestic satellite system with 254 earth stations; mobile cellular service is provided in four metropolitan cities
international: satellite earth stations - 8 Intelsat (Indian Ocean) and 1 Inmarsat (Indian Ocean region); nine gateway exchanges; 4 submarine cables; India-SEA-ME-WE-3, SEA-ME-WE-2 with landing sites at Cochin and Mumbai (Bombay); Fiber-Optic Link Around the Globe (FLAG) with landing site at Mumbai (Bombay) (2000)

Liberalization of the telecom sector began in 1994 and become more serious in 2000 under the regulation of the Ministry of Communications. Mobile penetration is still 1%, while fixed-line is now around 4%-5%. these still very low numbers show the stage of the telecommunications industry in general, but also India's vast potentials. Currently, no domestic handset manufacturing is done in India.One can expect that foreign OEMs will pursue this market aggressively.

India's fixed-line telephony penetration is estimated to be at approximately 40-45 million which currently represents a teledensity of 4%-5%. UBS Warburg's Indian analysts team estimates that by 2007, fixed-line penetration will reach 10%.

In January of 2003, India had only a total of about 10 million wireless users. For the size of its population, the wireless market is clearly in its infancy, but will grow steadily. There are 21 GSM operators in India, the top three account for 60% of India's GSM subscriber market.

At the end of January 2003, there were 11 million GSM subscribers. Four million alone were added in the previous six to seven months, which shows that India's mobile subscriber growth has clearly began. As can be expected, the Metro region dominates.

[Source: Siemens Business Information Report ST030401: India's Telecommunications Market: Implementation Priorities and Differences with China, March 2003]

Internet Service Providers (ISPs): 43 (2000)

Internet users: 7 million (2002)

1.7 Automotive market

The automotive industry in India is one of the fastest growing in the world. The industry dates back to the 1930's and was once the domain of only a handful of Indian manufacturers. In 1993, foreign investment was allowed in the automotive industry through joint ventures with the Indian Government. Since the liberalization of the Inian economy and dismantling of restrictions on investment and imports, the domestic automotive industry has witnessed fierce competition from foreign firms.

India is the largest tractor manufacturer, the second-largest two wheeler manufacturer, and the fifth-largest maufacturer of commercial vehicles in the world. The automotive industry is the largest in India.

Demand for passenger cars in India is slated to grow at a compounded annual rate of growth of 8% till 2011-12. According to a study undertaken by the National Council of Applied Economic Research, the demand for passenger cars will grow from 613,000 units in 2002-02 to 1,227,000 units in 2011-12.

Interestingly, multi vehicles will see a higher compounded annual rate of growth than passenger cars, and will be giving active competition to the passenger cars segment in India. The total demand for multi vehicles is projected to grow from 130,000 units in 2002-03 to 282,000 units in 2011-12.
[Source: Siemens Business Information Report ST030403: Automotive Market in India, April 2003]

2 The Languages

2.1 Survey of Indian languages

The Indian subcontinent consists of a number of separate linguistic communities each of which share a common language and culture.

Some Indian languages have a long literary history--Sanskrit literature is more than 5,000 years old and Tamil 3,000. India also has some languages that do not have written forms. There are 18 officially recognized languages in India (Konkani, Manipuri and Nepali were added in 1992) and each has produced a literature of great vitality and richness.

Although some of the languages are called "tribal" or "aboriginal", their populations may be larger than those that speak some European languages. For example, Bhili and Santali, both tribal languages, each have more than 4 million speakers. Gondi is spoken by nearly 2 million people. India's schools teach 58 different languages. The nation has newspapers in 87 languages, radio programmes in 71, and films in 15.

The Indian languages belong to four language families: Indo-European, Dravidian, Mon-Khmer, and Sino-Tibetan. Indo-European and Dravidian languages are used by a large majority of India's population. The language families divide roughly into geographic groups. Languages of the Indo-European group are spoken mainly in northern and central regions.

The languages of southern India are mainly of the Dravidian group. Some ethnic groups in Assam and other parts of eastern India speak languages of the Mon-Khmer group. People in the northern Himalayan region and near the Burmese border speak Sino-Tibetan languages.

Speakers of 54 different languages of the Indo-European family make up about three-quarters of India's population. Twenty Dravidian languages are spoken by nearly a quarter of the people. Speakers of 20 Mon-Khmer languages and 98 Sino-Tibetan languages together make up about 2 per cent of the population.

[http://indiansaga.com/languages/index.html > home]

The Indians also distinguish between the general north Indian accent and general south Indian accent.

[http://adaniel.tripod.com]

About 80 percent of all Indians--nearly 750 million people based on 1995 population estimates--speak one of the Indo-Aryan group of languages. Persian and the languages of Afghanistan are close relatives, belonging, like the Indo-Aryan languages, to the Indo-Iranian branch of the Indo-European family. Brought into India from the northwest during the second millennium B.C., the Indo-Aryan tongues spread throughout the north, gradually displacing the earlier languages of the area.

Modern linguistic knowledge of this process of assimilation comes through the Sanskrit language employed in the sacred literature known as the Vedas . Over a period of centuries, Indo-Aryan languages came to predominate in the northern and central portions of South Asia .

As Indo-Aryan speakers spread across northern and central India, their languages experienced constant change and development. By about 500 B.C., Prakrits, or "common" forms of speech, were widespread throughout the north. By about the same time, the "sacred," "polished," or "pure" tongue--Sanskrit--used in religious rites had also developed along independent lines, changing significantly from the form used in the Vedas. However, its use in ritual settings encouraged the retention of archaic forms lost in the Prakrits. Concerns for the purity and correctness of Sanskrit gave rise to an elaborate science of grammar and phonetics and an alphabetical system seen by some scholars as superior to the Roman

system. By the fourth century B.C., these trends had culminated in the work of Panini, whose Sanskrit grammar, the *Ashtadhyayi* , set the basic form of Sanskrit for subsequent generations.

Around 18 percent of the Indian populace (about 169 million people in 1995) speak Dravidian languages. Most Dravidian speakers reside in South India, where Indo-Aryan influence was less extensive than in the north. Only a few isolated groups of Dravidian speakers, such as the Gonds in Madhya Pradesh and Orissa, and the Kurukhs in Madhya Pradesh and Bihar, remain in the north as representatives of the Dravidian speakers who presumably once dominated much more of South Asia. (The only other significant population of Dravidian speakers are the Brahuis in Pakistan.)

The oldest documented Dravidian Indian language is Tamil, with a substantial body of literature, particularly the Cankam poetry, going back to the first century A.D. Kannada and Telugu developed extensive bodies of literature after the sixth century, while Malayalam split from Tamil as a literary language by the twelfth century. In spite of the profound influence of the Sanskrit language and Sanskritic culture on the Dravidian languages, a strong consciousness of the distinctness of Dravidian languages from Sanskrit remained. All four major Dravidian languages had consciously differentiated styles varying in the amount of Sanskrit they contained. In the twentieth century, as part of an anti-Brahman movement in Tamil Nadu, a strong movement arose to "purify" Tamil of its Sanskrit elements, with mixed success. The other three Dravidian languages were not much affected by this trend.

There are smaller groups, mostly tribal peoples, who speak Sino-Tibetan and Austroasiatic languages. Sino-Tibetan speakers live along the Himalayan fringe from Jammu and Kashmir to eastern Assam . They comprise about 1.3 percent, or 12 million, of India's 1995 population. The Austroasiatic languages, composed of the Munda tongues and others thought to be related to them, are spoken by groups of tribal peoples from West Bengal through Bihar and Orissa and into Madhya Pradesh. These groups make up approximately 0.7 percent (about 6.5 million people) of the population.

Despite the extensive linguistic diversity in India, many scholars treat South Asia as a single linguistic area because the various language families share a number of features not found together outside South Asia. Languages entering South Asia were "Indianized." Scholars cite the presence of retroflex consonants, characteristic structures in verb formations, and a significant amount of vocabulary in Sanskrit with Dravidian or Austroasiatic origin as indications of mutual borrowing, influences, and counterinfluences. Retroflex consonants, for example, which are formed with the tongue curled back to the hard palate, appear to have been incorporated into Sanskrit and other Indo-Aryan languages through the medium of borrowed Dravidian words.

[http://www.india4world.com/indian-language/index.shtml]

2.2 Indian languages diversity, use, and policy

The languages of India belong to four major families: Indo-Aryan (a branch of the Indo-European family), Dravidian, Austroasiatic (Austric), and Sino-Tibetan, with the overwhelming majority of the population speaking languages belonging to the first two families. (A fifth family, Andamanese, is spoken by at most a few hundred among the indigenous tribal peoples in the Andaman Islands, and has no agreed upon connections with families outside them.) The four major families are as different in their form and construction as are, for example, the Indo-European and Semitic families. A variety of scripts are employed in writing the different languages. Furthermore, most of the more widely used Indian languages exist in a number of different forms or dialects influenced by complex geographic and social patterns.

The Indian constitution recognizes official languages . Articles 343 through 351 address the use of Hindi, English, and regional languages for official purposes, with the aim of a

nationwide use of Hindi while guaranteeing the use of minority languages at the state and local levels. Hindi has been designated India's official language, although many impediments to its official use exist.

The constitution's Eighth Schedule, as amended by Parliament in 1992, lists eighteen official or Scheduled Languages . They are Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, and Urdu. (Precise numbers of speakers of these languages are not known. They were not reported in the 1991 census, and estimates are subject to considerable variation because of the use of multiple languages by individual speakers.) Of the official languages, approximately 403 million people, or about 43 percent of the estimated total 1995 population, speak Hindi as their mother tongue. Telugu, Bengali, Marathi, and Tamil rank next, each the mother tongue of about 4 to 5 percent (about 37 million to 47 million people); Urdu, Gujarati, Malayalam, Kannada, and Oriya are claimed by between 2 and 3 percent (roughly 19 million to 28 million people); Bhojpuri, Punjabi, and Assamese by 1 to 2 percent (9 million to 19 million people); and all other languages by less than 1 percent (less than 9 million speakers) each.

Since independence in 1947, linguistic affinity has served as a basis for organizing interest groups; the "language question" itself has become an increasingly sensitive political issue. Efforts to reach a consensus on a single national language that transcends the myriad linguistic regions and is acceptable to diverse language communities have been largely unsuccessful.

Many Indian nationalists originally intended that Hindi would replace English--the language of British rule (1757-1947)--as a medium of common communication. Both Hindi and English are extensively used, and each has its own supporters. Native speakers of Hindi, who are concentrated in North India, contend that English, as a relic from the colonial past and spoken by only a small fraction of the population, is hopelessly elitist and unsuitable as the nation's official language. Proponents of English argue, in contrast, that the use of Hindi is unfair because it is a liability for those Indians who do not speak it as their native tongue. English, they say, at least represents an equal handicap for Indians of every region.

English continues to serve as the language of prestige in India. Efforts to switch to Hindi or other regional tongues encounter stiff opposition both from those who know English well and whose privileged position requires proficiency in that tongue and from those who see it as a means of upward mobility. Partisans of English also maintain it is useful and indeed necessary as a link to the rest of the world, that India is lucky that the colonial period left a language that is now the world's predominant international language in the fields of culture, science, technology, and commerce. They hold, too, that widespread knowledge of English is necessary for technological and economic progress and that reducing its role would leave India a backwater in world affairs.

Linguistic diversity is apparent on a variety of levels. Major regional languages have stylized literary forms, often with an extensive body of literature, which may date back from a few centuries to two millennia ago. These literary languages differ markedly from the spoken forms and village dialects that coexist with a plethora of caste idioms and regional lingua francas . Part of the reason for such linguistic diversity lies in the complex social realities of South Asia. Indian languages reflect the intricate levels of social hierarchy and caste. Individuals have in their speech repertoire a variety of styles and dialects appropriate to various social situations. In general, the higher the speaker's status, the more speech forms there are at his or her disposal. Speech is adapted in countless ways to reflect the specific social context and the relative standing of the speakers.

Determining what should be called a language or a dialect is more a political than a linguistic question. Sometimes the word *language* is applied to a standardized and prestigious form, recognized as such over a large geographic area, whereas the word *dialect* is used for the various forms of speech that lack prestige or that are restricted to certain regions or castes but are still regarded as forms of the same language. Sometimes mutual intelligibility is the

criterion: if the speakers can understand each other, even though with some difficulty, they are speaking the same language, although they may speak different dialects. However, speakers of Hindi, Urdu, and Punjabi can often understand each other, yet they are regarded as speakers of different languages. Whether or not one thinks Konkani--spoken in Goa, Karnataka, and the Konkan region of Maharashtra--is a distinct language or a dialect of Marathi has tended to be linked with whether or not one thinks Goa ought to be merged with Maharashtra. The question has been settled from the central government's point of view by making Goa a state and Konkani a Scheduled Language. Moreover, the fact that the Latin script is predominantly used for Konkani separates it further from Marathi, which uses the Devanagari script. However, Konkani is also sometimes written in Devanagari and Kannada scripts.

Regional India's languages are an issue in the politically charged atmosphere surrounding language policy. Throughout the 1950s and 1960s, attempts were made to redraw state boundaries to coincide with linguistic usage. Such efforts have had mixed results. Linguistic affinity has often failed to overcome other social and economic differences. In addition, most states have linguistic minorities, and questions surrounding the definition and use of the official language in those regions are fraught with controversy.

[http://www.india4world.com/indian-language/index.shtml]

States whose boundaries are based on languages are Kerala for Malyalam speakers. Tamil Nadu for Tamil speakers. Karnataka for Kanadda speakers. Andra Pradesh for Telugu speakers. Maharashtra for Marathi speakers. Orissa for Oriya speakers. West Bengal for Bengali speakers. Gujarat for Gujarati speakers. Punjab for Punjabi speakers. Assam for Assami speakers. Some of these states like Bengal and Orissa were provinces during British rule. Though many states were created based on language boundaries, there are other states which weren't created based on language boundaries and there are many language speaker who don't have their own state.

To name a few other languages spoken in India, one can name Dogri, Ladacki and Kashmiri which are spoken in different parts of Jammu and Kashmir state. In Sikkim, different languages are spoken. The main language there is Nepali. In Manipur the main language is Manipuri. In Madya Pradesh, Uttar Pradesh, Rajasthan, Haryana, Himachal Pradesh the main language is Hindi, which is also become the national language of India. Some languages of India aren't specific to a region of India, like Sindhi whose speakers came to India from Sindh (in present day Pakistan), but are scattered all over India. Urdu is spoken by many Muslims all over India. The different tribes of India (some of them only a few hundreds) also have their own languages.

As stated earlier most of the main Indian languages have different dialects and variations, sometimes very different from each other. Hindi has more than ten variations. Hindi spoken in Rajasthan is different from Hindi spoken in Bihar or Hindi of Himachal Pradesh. Sometimes the different variations of a language are considered as separate language with their own literature. One of Hindi dialects spoken in east India is Maithali. Many Maithali speakers regard their language as a different language from Hindi. Also Rajasthani from Rajasthan is considered sometimes as a different language and not as Hindi. But, actually Rajasthani also isn't one language but different tribal languages spoken by the people of Rajasthan and they all call their languages after the name of their region.

Another language named after its region is Konkani spoken in Goa and named as such because of the Konkan coast. To the north of Goa in the Konkan coast of Maharashtra there is another 'Konkani' language which is considered a dialect derived from the Marathi language and is different from Goa's Konkani language.

The Indian constitution uses the term 'mother tongue' instead of language or dialect. Officially the central government recognizes 18 languages, but each language includes in it many mother tongues. The Indian census records over 200 different mother tongues.

S

Despite the different languages and dialects, most of the official languages speakers have developed a standard of speaking language which has become the accepted style of speaking for that language. Sometimes, like in the case of Hindi this language is completely different from some of its dialects.

[http://adaniel.tripod.com/Languages1.htm]

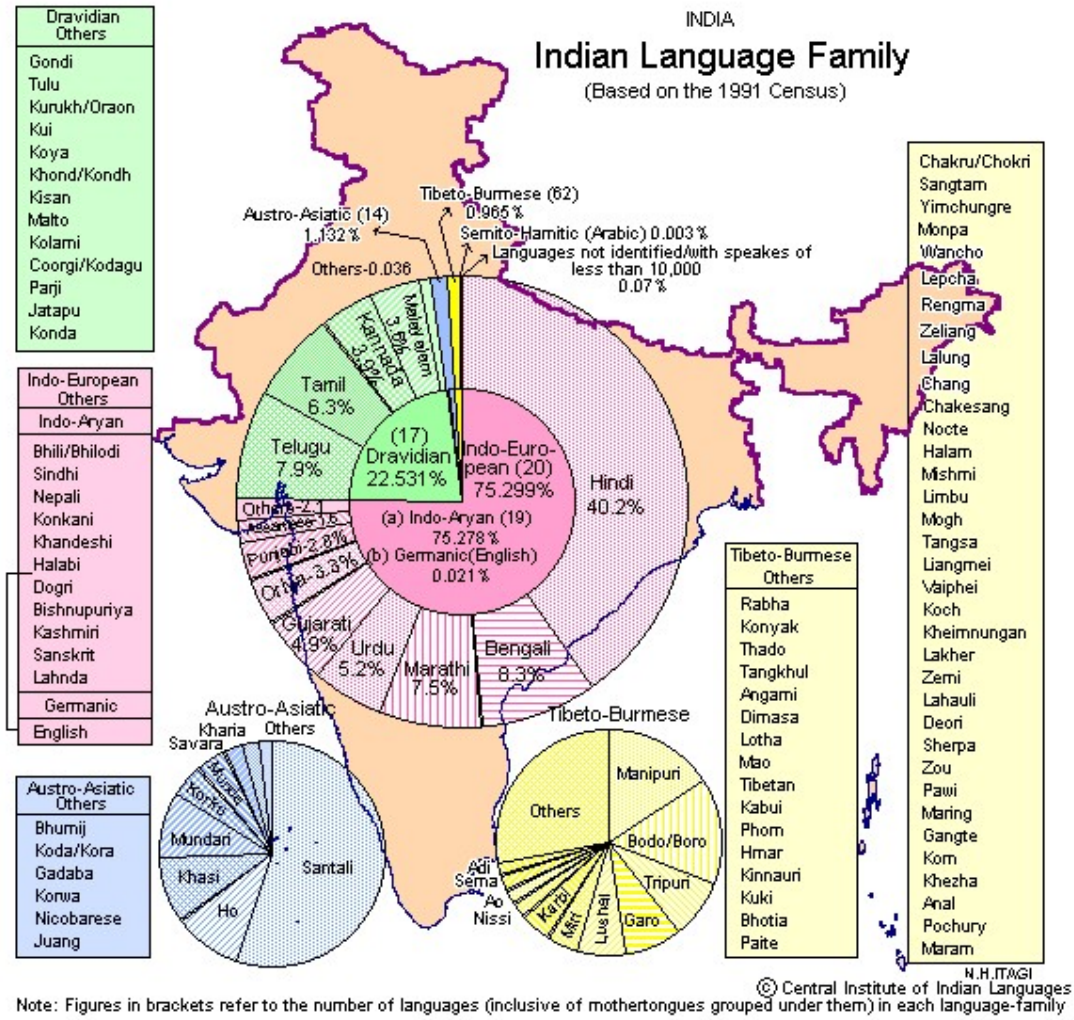
Table 2: Official languages spoken in India

	Language	*Number of speakers*	*Percentage*
1	Hindi	337,272,114	40.22%
2	Bengali	69,595,738	8.30%
3	Telugu	66,017,615	7.87%
4	Marathi	62,481,681	7.45%
5	Tamil	53,006,368	6.32%
6	Urdu	43,406,932	5.18%
7	Gujarati	40,673,814	4.85%
8	Kannada	32,753,676	3.91%
9	Malayalam	30,377,176	3.62%
10	Oriya	28,061,313	3.35%
11	Punjabi	23,378,744	2.79%
12	Assamese	13,079,696	1.56%
13	Sindhi	2,122,848	0.25%
14	Nepali	2,076,645	0.25%
15	Konkani	1,760,607	0.21%
16	Manipuri	1,270,216	0.15%
17	Kashmiri	56,693	0.01%
18	Sanskrit	49,736	0.01%
19	Other Languages	31,142,376	3.71%
	Total :	838,583,988	100.00%

Source : 1991 Census of India

S

Fig. 4: Indian language family



[http://www.ciil.org/languages/map4.html]

S

Table 3: Three Main Languages in every State (Census of India 1991)

State/UT	Languages	Number of Speakers	Per cent	State/UT	Languages	Number of Speakers	Per cent
I N D I A*				**Andhra Pradesh**			
	Hindi	**337,272,114**	**40.2**		Telugu	56,375,755	84.8
	Bengali	**69,595,738**	**8.3**		Urdu	5,560,154	8.4
	Telugu	**66,017,615**	**7.9**		Hindi	1,841,290	2.8
Arunachal Pradesh				**Assam**			
	Nissi/Daffla	172,149	19.9		Assamese	12,958,088	57.8
	Nepali	81,176	9.4		Bengali	2,523,040	11.3
	Bengali	70,771	8.2		Bodo/Boro	1,184,569	5.3
Bihar				**Goa**			
	Hindi	69,845,979	80.9		Konkani	602,626	51.5
	Urdu	8,542,463	9.9		Marathi	390,27	33.4
	Santhali	2,546,655	2.9		Kannada	54,323	4.6
Gujarat				**Haryana**			
	Gujarati	37,792,933	91.5		Hindi	14,982,409	91.0
	Hindi	1,215,825	2.9		Punjabi	1,170,225	7.1
	Sindhi	704,088	1.7		Urdu	261,82	1.6
Himachal Pradesh				**Karnataka**			
	Hindi	4,595,615	88.9		Kannada	29,785,004	66.2
	Punjabi	324,479	6.3		Urdu	4,480,038	10.0
	Kinnauri	61,794	1.2		Telugu	3,325,062	7.4
Kerala				**Madhya Pradesh**			
	Malayalam	28,096,376	96.6		Hindi	56,619,090	85.6
	Tamil	616,01	2.1		Bhili/Bhilodi	2,215,399	3.3
	Kannada	75,571	0.3		Gondi	1,481,265	2.2
Maharashtra				**Manipur**			
	Marathi	57,894,839	73.3		Manipuri	1,110,134	60.4
	Hindi	6,168,941	7.8		Thado	103,667	5.6
	Urdu	5,734,468	7.3		Tangkhul	100,088	5.4
Meghalaya				**Mizoram**			
	Khasi	879,192	49.5		Lushai/Mizo	518,099	75.1
	Garo	547,69	30.9		Bengali	59,092	8.6
	Bengali	144,261	8.1		Lakher	22,938	3.3
Nagaland				**Orissa**			
	Ao	169,837	14.0		Oriya	26,199,346	82.8
	Sema	152,123	12.6		Hindi	759,016	2.4
	Konyak	137,539	11.4		Telugu	502,102	1.6
Punjab				**Rajasthan**			
	Punjabi	18,704,461	92.2		Hindi	39,410,968	89.6
	Hindi	1,478,993	7.3		Bhili/Bhilodi	2,215,399	5.0
	Urdu	13,416	0.1		Urdu	953,497	2.2
Sikkim				**Tamil Nadu**			
	Nepali	256,418	63.1		Tamil	48,434,744	86.7
	Bhotia	32,593	8.0		Telugu	3,975,561	7.1
	Lepcha	29,854	7.3		Kannada	1,208,296	2.2
Tripura				**Uttar Pradesh**			
	Bengali	1,899,162	68.9		Hindi	125,348,492	90.1
	Tripuri	647,847	23.5		Urdu	12,492,927	9.0
	Hindi	45,803	1.7		Punjabi	661,215	0.5
West Bengal				**A & N Island**			
	Bengali	58,541,519	86.0		Bengali	64,706	23.1
	Hindi	4,479,170	6.6		Tamil	53,536	19.1
	Urdu	1,455,649	2.1		Hindi	49,469	17.6
Chandigarh				**D & N Haveli**			
	Hindi	392,054	61.1		Bhili/Bhilodi	76,207	55.0
	Punjabi	222,89	34.7		Gujarati	30,346	21.9
	Tamil	5,318	0.8		Konkani	17,062	12.3
Daman & Diu				**Delhi**			
	Gujarati	92,579	91.1		Hindi	7,690,631	81.6
	Hindi	3,645	3.6		Punjabi	748,145	7.9
	Marathi	1,256	1.2		Urdu	512,99	5.4
Lakshadweep				**Pondicherry**			
	Malayalam	43,678	84.5		Tamil	720,473	89.2
	Tamil	282	0.5		Malayalam	38,392	4.8
	Hindi	217	0.4		Telugu	34,799	4.3

* Excludes figures for Jammu and Kashmir where the 1991 census could not be conducted due to disturbed conditions.

Source of data: Statement 3,
ensus of India 1991,Paper 1 of 1997 - Language

[http://www.censusindia.net/cendat/datatable26.html]

2.3 Hindi and English: Political issues

For the speakers of the country's myriad tongues to function within a single administrative unit requires some medium of common communication. The choice of this tongue, known in India as the "link" language, has been a point of significant controversy since independence. Central government policy on the question has been necessarily equivocal. The vested interests proposing a number of language policies have made a decisive resolution of the "language question" all but impossible.

The central issue in the link-language controversy has been and remains whether Hindi should replace English. Proponents of Hindi as the link language assert that English is a foreign tongue left over from the British Raj . English is used fluently only by a small, privileged segment of the population; the role of English in public life and governmental affairs constitutes an effective bar to social mobility and further democratization. Hindi, in this view, is not only already spoken by a sizable minority of all Indians but also would be easier to spread because it would be more congenial to the cultural habits of the people. On the other hand, Dravidian-speaking southerners in particular feel that a switch to Hindi in the well-paid, nationwide bureaucracies, such as the Indian Administrative Service, the military, and other forms of national service would give northerners an unfair advantage in government examinations . If the learning of English is burdensome, they argue, at least the burden weighs equally on Indians from all parts of the country. In the meantime, an increasing percentage of Indians send their children to private English-medium schools, to help assure their offspring a chance at high-privilege positions in business, education, the professions, and government.

[http://www.india4world.com/indian-language/index.shtml]

One of the main political issues in Indian politics is connected to language problem. After India’s independence the government decided that the official language of India will be Hindi. Hindi belongs to the family of Aryan languages. Speakers of other languages, especially the Dravidian languages, saw in this decision an attempt to erase their language cultures. But the Indian constitution has declared that English can also be used for official purposes. Hindi has at least 13 different dialects and she is the most commonly spoken language in India. But the reason Hindi was chosen to be the official language of India wasn’t because it is the most commonly spoken language in India, but it has connection with India history before it’s independence.

Before its independence, most of India was a British colony. Before the British the most dominant Empire of north India was the Moghul Empire. The Moghuls were Muslim invaders who arrived in India from the present day Afghanistan. The official language of the Moghul courts was Persian. The Moghuls, like other residents who lived to the west of the Indian sub-continent named India as ‘Hind’ or ‘Hindustan’, after the river Indus which flows in the present day Pakistan. The language spoken in ‘Hind’ was called by them Hindi or Hindustani. This language and its script were based on an ancient Indian language called Sanskrit. Most of the sacred books of Hinduism are written in Sanskrit and the script is called Devanagiri.

Some of the Moghul family members were great patrons of poetry and music. Slowly there developed a ‘Hindustani’ poetry, based on Hindustani language which used words from Arabic and Persian and was written in Perso-Arabic script. This language was called Urdu. Urdu also replaced Persian as the language of the Moghul courtyards. And so there developed two languages with different writings but were actually one language when spoken except for their higher vocabularies. For example, rulers were titled in Urdu language as Shah, Nawab or Nizam. While in Hindi they were called Raja or Maharaja. Among the Hindustani speakers of north India, Urdu became the language of the Muslims while Hindi became the language of others.

After the collapse of the Moghuls the British became the rulers of north India. The British introduced English to India and continued using Urdu for official purposes. But nationalist Hindus demanded from the British to change the official language from Urdu to Hindi which is written in Indian script. Even Hindus whose mother tongue were not Hindi supported this argument. This debate between the Hindus and the Muslims continued right up to the independence of India. Against this stand of two different languages two of India's notable leaders, Jawaharlal Nehru and Mahatma Gandhi, supported the idea of one Hindustani language which could be written in both forms. But when British India was divided in two countries, India and Pakistan. Muslims who got Pakistan made Urdu their official language and Indians made Hindi with Devanagiri script as their official language. But the debate over the official language didn't end up with choosing Hindi with Devanagiri script as the official language. New debates occurred because of this decision.

One problem was connected to the different dialects of Hindi and the second problem was connected to other languages which exist in India. The first problem was which dialect of Hindi is the right Hindi. Hindi has at least 13 dialects, some of them completely different from each other. Two reasons caused to it that Hindi language includes in it so many different dialects. One reason was related to the fact that India is called Hind in many languages spoken west from it up to the Middle East. Muslim invaders of India like the Moghuls came from these regions and called the language spoken in 'Hind' as Hindi. The Indians also began calling their different languages as 'Hindi'. The other reason which concerns to the fact that Hindi has so many different dialects is related to the independence period of India and the debate of the official language of India.

Most of the Indians belong religiously to Hinduism and they perceive Urdu, written in Perso-Arabic script as Muslim language. Before the independence of India the Muslims supported the continuation of Urdu as the official language of India, while the Hindus supported Hindi. In order to secure Hindi's position as the sole official language of India the political leaders convinced the north Indians to claim that they speak a Hindi dialect and so different dialect speakers were put together in the Hindi speaking category by the British bureaucrats. After India's independence when Hindi was chosen as the official language of India, different 'Hindi' language speakers began demanding official recognition of their languages. Maithali and Punjabi speakers also demanded to recognize their languages as separate languages from Hindi. Of the different 'Hindi' languages, only Punjabi got this recognition. Other 'Hindi' languages are considered dialects of Hindi and their status in the different states of India isn't clear and is interpreted differently by different parties. The official Hindi is based on the dialect which was spoken in the Delhi-Agra region with a Sanskrit vocabulary. While the popular Hindi spoken by majority of Indians is based on this dialect, it is also affected by the different cultures of India mainly the Hindi cinema based in Mumbai(formerly Bombay) in west India and it includes many English words.

Among the other language speakers of India, the decision to choose Hindi as the official language was seen as an attempt to erase their cultures. After different struggles – political, violent and passive – the central government decided to allow the state governments to pick their official languages and recognized constitutionally other languages of India. For now the Indian constitution recognizes 18 Indian languages. One of meanings of the constitutional recognition is the right to use any of these languages for government service examinations. But, in reality this possibility isn't always given to the examinee.

The different states of India have different official languages, some of them not recognized by the central government. Some states have more then one official language. Bihar in east India has three official languages - Hindi, Urdu and Bengali – which are all recognized by the central government. But Sikkim, also in east India, has four official languages of which only Nepali is recognized by the central government. Besides the languages officially recognized by central or state governments, there are other languages which don't have this recognition and their speakers are running political struggles to get this recognition. Anyway as stated earlier the central government decided that Hindi is the official language of India and

therefore it has also the status of official language in the states. Another language that has a official status in all states is English.
[http://adaniel.tripod.com/Languages2.htm]

According to "India Today" (August 18, 1997) following opinion poll about the language issue was carried out. The first question was "Do you think there should be one language across the nation?".
The result was: YES 61, NO 33, Rest: Don' t Know / Can't Say.
The second question was: "Which language?"
The result was: HINDI 77, ENGLISH 8, Others: 15.
Hindi: North 97, East 83, West 75, South 31.
(All figures in precentage)

2.4 Hindi

SIL code: HND ; ISO 639-1:hi ; ISO 639-2: hin
Population: [HND] India. 180,000,000 in India (1991 UBS), 363,839,000 or nearly 50% of the population including second language users in India (1997 IMA). Population total all countries 366,000,000 first language speakers (1999 WA), 487,000,000 including second language users (1999 WA). Alternate names: KHARI BOLI, KHADI BOLI. Classification: Indo-European, Indo-Iranian, Indo-Aryan, Central zone, Western Hindi, Hindustani.
The following 4 languages have a name, alternate name, or dialect name (in at least one country) that is similar to the name for this ISO code, but they are not encompassed by the code.

- BUNDELI: [BNS] India. 644,000 (1997 IMA) to 8,000,000 or more (1997). Alternate names: BUNDEL KHANDI. Dialects: STANDARD BUNDELI, PAWARI (POWARI), LODHANTI (RATHORA), KHATOLA, BANAPHARI, KUNDRI, NIBHATTA, TIRHARI, BHADAURI (TOWARGARHI), LODHI, KOSTI, KUMBHARI, GAOLI, KIRARI, RAGHOBANSI, NAGPURI HINDI, CHHINDWARA BUNDELI. Classification: Indo-European, Indo-Iranian, Indo-Aryan, Central zone, Western Hindi, Bundeli.
- DOGRI-KANGRI: [DOJ] India. 2,200,000 including 2,105,000 Dogri (1997 IMA), 95,000 Kangri (1997 IMA). Alternate names: DOGRI, DHOGARYALI, DOGARI, DOGRI JAMMU, DOGRI PAHARI, DONGARI, HINDI DOGRI, TOKKARU, DOGRI-KANGRA. Dialects: BHATBALI, EAST DOGRI, KANDIALI, KANGRI (KANGRA), NORTH DOGRI, DOGRI. Classification: Indo-European, Indo-Iranian, Indo-Aryan, Northern zone, Western Pahari. More information.
- HINDUSTANI, CARIBBEAN: [HNS] Suriname. 150,000 in Suriname, about 38% of population (1986). Population total all countries 165,600 or more. Dialects: TRINIDAD BHOJPURI, SARNAMI HINDUSTANI (SARNAMI HINDI, AILI GAILI). Classification: Indo-European, Indo-Iranian, Indo-Aryan, Eastern zone, Bihari.
- HINDUSTANI, FIJIAN: [HIF] Fiji. 380,000 (1991 UBS) or 48.6% of the population (1987 Honolulu Star-Bulletin). Population total all countries 380,000. Alternate names: FIJIAN HINDI. Classification: Indo-European, Indo-Iranian, Indo-Aryan, East Central zone.

Region: Throughout northern India: Delhi; Uttar Pradesh; Rajasthan; Punjab; Madhya Pradesh; northern Bihar; Himachal Pradesh. Also spoken in Bangladesh, Belize, Botswana, Germany, Kenya, Nepal, New Zealand, Philippines, Singapore, South Africa, Uganda, UAE, United Kingdom, USA, Yemen, Zambia.
Alternate names: KHARI BOLI, KHADI BOLI

Classification: Indo-European, Indo-Iranian, Indo-Aryan, Central zone, Western Hindi, Hindustani.
Comments: Formal vocabulary is borrowed from Sanskrit, de-Persianized, de-Arabicized. Literary Hindi, or Hindi-Urdu, has four varieties: Hindi (High Hindi, Nagari Hindi, Literary Hindi, Standard Hindi); Urdu; Dakhini; Rekhta. State language of Delhi, Uttar Pradesh, Rajasthan, Madhya Pradesh, Bihar, Himachal Pradesh. Languages and dialects in the Western Hindi group are Hindustani, Haryanvi, Braj Bhasha, Kanauji, Bundeli; see separate entries. Spoken as mother tongue by the Saharia in Madhya Pradesh. Hindi, Hindustani, Urdu could be considered co-dialects, but have important sociolinguistic differences. National language. Grammar. SOV. Devanagari script. Hindu. Bible 1818-1987.
Also spoken in: Nepal, South Africa, Uganda
[http://www.ethnologue.com/show_iso639.asp?code=hi]
[http://www.ethnologue.com/show_language.asp?code=HND]

Dialects of Hindi: Marwari, Braj, Bundeli, Kanauji, Urdu, Chattisgarhi, Bagheli, Avadhi, Bhojpuri and many others. It is not easy to delimit the borders of the Hindi speaking region. There has been considerable controversy on the status of Punjabi and Maithili. Sometimes they are regarded to be independent languages and sometimes dialects of Hindi. A 1997 survey found that 66% of all Indians can speak Hindi, and 77% of the Indians regard Hindi as "one language across the nation".
[http://www.cs.colostate.edu/~malaiya/hindiint.html]
Western Hindi (the Khariboli speech of Delhi) It has become the national language while Maithili, Magahi, Bhojpuri, Awadhi, Bagheli, Brajabhasa, Chattisgarhi with other Central and Western Himalayan dialects being described as dialects of Hindi.
[http://www.esamskriti.com/html/new_inside.asp?cat_name=cultphil&cid=513&sid=9001]
The Hindi language comprise of a number of dialects of which those used for literary composition are Khariboli, Rajasthani, Maithli, Brijbhasha and Awadhi.
[http://indiansaga.com/languages/index.html > hindi]
Hindi is numerically the biggest of the Indo-Aryan family and is the official language of India. Among the various dialects of Hindi, the dialect chosen as official Hindi is the standard khariboli, written in the Devanagiri script. From the literary point of view, the term Hindi covers not only the khariboli form but also a number of other dialects like the Brajbhasha, Bundeli, and Awadhi, early Marhwadi of Rajasthan and the Maithili and Bhojpuri of Bihar. In six states and union Territories, Hindi is the official language.
[http://www.birminghamuk.com/asianlanguage.htm]
The title ‘Hindi’ (originally a Persian word meaning ‘Indian’) actually embraces a wide range of dialects, the geographical extremes of which may be so diverse as to be mutually unintelligible. However the modern standard form of Hindi (based on the speech of Delhi and Uttar Pradesh) is widely accepted.
[http://www.linguaphone.co.uk/language.cfm?language_id=14]
Amongst its interesting features is a three-tier level of honorifics, allowing great subtlety in adjusting the level of communication to suit ‘formal’, ‘familiar’ and ‘intimate’ conversational contexts. Thus, the polite communicating of gratitude etc. is an intrinsic part of the language itself and does not rely solely on separate words for ‘please’ and ‘thank you’.
[http://www.linguaphone.co.uk/language.cfm?language_id=14]

Hindustani languages: "Hindustani" is the term formerly used to collectively name the Hindi and Urdu languages. The two are largely similar in grammar and vocabulary. However, Urdu, spoken in largely-Muslim Pakistan, has more loanwords from Persian and Arabic and is written in the Arabic alphabet, while Hindi, spoken in largely-Hindu northern India, has more influence from the southern Dravidian languages and classical Sanskrit and is written in

Devanagari script.
[http://en2.wikipedia.org/wiki/Hindustani]

Hindi as second/third language:

According to the 1991 Census of India 49,767,917 speakers have Hindi as second language and 20,976,588 speakers have it as third language. Depending on the first language the command of Hindi as second (or third) language varies considerably, e.g. from 0.7 % for Tamil speakers to 44.4 % for Sanskrit speakers
[see Annex 1: HINDI AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991].

2.5 English

According to the 1991 Census of India 64,602,299 speakers (8.0 %) have English as second language and 25,440,188 speakers (3.1 %) have it as third language. Depending on the first language the command of English as second (or third) language varies considerably, e.g. from 1.5 % for Gujarati speakers to 22.0 % for Malayalam speakers
[see Annex 2: ENGLISH AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991].
However, "India Today" (August 18, 1997) reports: Contrary to the census myth that English is the language of a microscopic minority, the poll indicates that almost one in three Indians claims to understand English, although less than 20 per cent are confident of speaking it.

2.6 Present-day scripts in India

There are 10 Indic scripts in vogue. They follow similar script and language grammars. Alphabetic order is similar. Some languages use common script, especially Devanagari.

Eighteen constitutional Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*); Assamese (*Assamese*), Manipuri (*Manipuri*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada *(Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) and Urdu (*Urdu*).

[Language Technology Development in India, Dr. Om Vikas, Ministry of Information Technology, New Delhi, India; Email: omvikas@mit.gov.in]

[http://www.emille.lancs.ac.uk/lesal/omvikas.pdf]

The general script of the Aryan languages is different from the general script of Dravidian languages.

Hindi has inherited its writing system from Sanskrit. The Devanagari script is derived from the ancient Brahmi and is closely related to other Indian scripts such as Gujarati and Bengali.

Devanagari is a logical writing system which has a phonetic basis so there are relatively few spelling problems.

The general appearance of the Devanagari script is that of letters ‘hanging from a line’. This ‘line’, evident in many South Asian scripts, is actually a component part of most of the letters and is drawn as the writing proceeds. The script has no capital letters.

[http://www.linguaphone.co.uk/language.cfm?language_id=14]

3 Indian Language Resources and Resources Centers

3.1 LDC, ELRA

At LDC 2 (very) small Hindi speech databases are available:

- LDC96S52 CALLFRIEND Hindi
- LDC94S17 OGI Multilanguage Corpus

[http://www.ldc.upenn.edu]

At ELRA no Hindi language resources are currently available.
[http://www.elra.info/]

Other speech databases of Indian languages are not known.

3.2 Central Institute of Indian Languages, (CIIL) Mysore

Machine Readable Corpora:

The source of corpora is Printed Books, Journals, Magazines, Newspapers and Government Documents published during 1981-1990. It has been categorised into six main categories viz. Aesthetics, Social Sciences, Natural, Physical & Professional Sciences, Commerce, Official and Media Languages and Translated Material. The Tag Set consists of Finite Verb (FV), Non-Finite Verb (NV), Noun (NN), Pronoun (PN), Adjective (AJ), Adverb (AV), Indeclinables (ID). Corpus Manager and KWIC Concordance s/ws have also been developed.Corpora of about 30 Lakh words in each of the Indian Languages viz. Hindi, Punjabi, English, Telugu, Malyalam, Tamil, Kannada, Sanskrit, Urdu, Kashmere, Marathi, Gujrati, Oriya, Assamese and Bengali has been developed at various centres and is now being centrally maintained at CIIL, Mysore. It is being distributed for educational and research purpose. Three more languages viz. Konkani, Manipuri & Nepali were later on added to the eighth schedule of the constitution, hence corpora development for these languages was also taken up.

Corpora of Konkani Language has been completed at Asmitai Pratishtan, Goa. Thirty Lakh words of Konkani Corpora in machine readable form and s/w for tagging the corpora, word count and frequency count has been developed. Spell Checker for use in conjunction with corpora has also been developed. This will also be maintained at CIIL, Mysore and made available for distribution.

Corpora of Nepali Language is under development at Centre for Computers and Communications Technology, Gangtok. Nepali Corpora of 1.2 Lakh words in machine readable form and s/w for tagging the corpora, word count and frequency count has been developed.

Corpora of Manipuri Language has been undertaken in University of Manipur, Manipur. Data collection has already been completed for 25 lakh words and data entry is in progress.

[http://tdil.mit.gov.in]

Lexical Resources in machine readable form:

The lexical resources of a language contain the information like Head word, Stem alterants, Stem type, detailed grammatical information, syntactic information, all types of meanings, citation for each meaning, paradigms, derived words, cross reference for the derived words,

compound words, synonyms, antonyms, idioms, encyclopeadic information, etymological information, statistical information. The lexical resource database will be useful to liguists and Computer Scientists who are working in linguistic research, machine translation, expert systems and Artificial Intelligence. It can be used for generation of learners dictionary, historical dictionary, machine readable grammatical dictionary, electronic dictionary, computational lexicon etc. Lexical Resources in five Indian Languages viz. Bengali, Hindi, Marathi, Tamil and Telugu is under advanced stage of development. Lexical Resources provide Lexical Information on the basis of concepts, more grammatical information, Syntactic and Semantic conditioning for the usage of lexical items, Synonym-set and their usages, Compound forms & Idioms. The categories for which Lexical Resources are being developed are Verb, noun, Adjective, Adverb and Function Word.

These can be used for research in the areas of Machine Translation systems during the Lexical Transfer Phase, Lexical Resource of Source Language in the analysis phase and Lexical Resource of Target Language in the synthesis phase etc.

[http://tdil.mit.gov.in]

3.3 EMILLE project

The EMILLE Corpus (Full release):

The EMILLE Corpus has been constructed as part of a collaborative venture between the EMILLE project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India.

The corpus consists of three components: monolingual, parallel and annotated corpora. There are fourteen monolingual corpora, including both written and (for some languages) spoken data for fourteen South Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu. The EMILLE monolingual corpora contains approximately 96,157,000 words (including 2,627,000 words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu). The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. The annotated component includes the Urdu monolingual and parallel corpora annotated for parts-of-speech, together with twenty written Hindi corpus files annotated to show the nature of demonstrative use. The corpus is marked up using CES-compliant SGML, and encoded using Unicode.

The EMILLE Corpus is distributed free of charge for use in non-profit-making research.

[http://www.ling.lancs.ac.uk/corplang/emille/default.htm]

[http://www.emille.lancs.ac.uk/home.htm]

EMILLE Corpus Beta Version

At the end of March 2003, the beta version of the EMILLE Corpus was released. Beta users include researchers in a variety of fields at institutions around the world. If you are interested in becoming a beta user, please email Andrew Hardie.

The beta version of the corpus consists of:

- 30 million words of monolingual written data (Gujarati, Tamil, Hindi, Punjabi)
- 600,000 words of monolingual spoken data (Hindi, Urdu, Punjabi, Bengali, Gujarati)
- 120,000 words of parallel data in each of (English, Hindi, Urdu, Punjabi, Bengali, Gujarati)

S

Further updates on available data and software, news of identified bugs in the corpus, and (where appropriate) patches to correct the same, will be available on this page in the months leading up to the final release of the EMILLE Corpus in summer of this year.
[http://www.emille.lancs.ac.uk/beta.htm]

3.4 Indian Resource Centers for Language Technology Solutions

The Ministry of IT has established thirteen Resource Centres for Indian Language Technology Solutions covering all the eighteen constitutionally approved official languages.

Organizations and associated Languages:

Indian Institute of Technology, Kanpur. (Hindi & Nepali)
Tel: 0512-597174 E-mail: rmk@iitk.ac.in

Indian Institute of Technology, Mumbai. (Marathi & Konkani)
Tel: 022-5767718 E-mail: pb@cse.iitb.ac.in

Indian Institute of Technology, Guwahati. (Assamese & Manipuri)
Tel: 0361-690321-28 E-mail: sbnair@iitg.ernet.in

Indian Institute of Science, Bangalore. (Kannada & Sanskrit Cognitive Models)
Tel: 080-3092377 E-mail: njrao@mgmt.iisc.ernet.in

Indian Statistical Institute, Kolkata. (Bengali)
Tel: 033-5778085 E-mail: bbc@isical.ac.in

Jawarharlal Nehru University, New Delhi. (Foreign Languages: Japanese, Chinese andSanskrit Language Learning Systems)
Tel: 011-6107676 E-mail:gvs@jnuniv.ernet.in

University of Hyderabad, Hyderabad. (Telugu)
Tel: 040-3010500 E-mail: knmcs@uohyd.ernet.in

Anna University, Chennai. (Tamil)
Tel: 044-2351723 E-mail: rp@annauniv.edu

MS University, Baroda. (Gujarati)
Tel: 0265-792959 E-mail:sityash@satyam.net.in

Utkal University, Bhuvaneshwar (Oriya)
Tel: 0674-580216 E-mail: sangham@sanchar.net.in

Orissa Computer Application Centre, Bhuvaneshwar (Oriya)
Tel: 0674-543113 E-mail: akp@ocac.ernet.in

Thapar Institute of Engg. & Tech., Patiala. (Punjabi)
Tel: 0175-393137 E-mail: gslehal@mailcity.com

Electronics Research & Development Ccenter (ER&DC), Trivendrum. (Malayalam)
Tel: 0471-325897 E-mail: ravi@erdcitvm.org

Center for Development of Advanced Computing (C-DAC), Pune. (Urdu, Sindhi & Kashmiri)
Tel: 020-5652461 E-mail: rkarora@cdac.ernet.in

The core objectives of these Resource Centres are:

- To act as a repository of all knowledge tools and products concerned with computer processing of Indian Languages and bring out yearly resource documents.
- To develop the methodologies and tools for seamless integration of language processing tools with existing and evolving software development environment.
- To network with Centres concerned with computer processing of Indian Languages and potential user agencies.
- To create content and databases on the resource information available in Indian languages and to put at least 10 most respected books (related to Indian Heritage) in

Indian language on the web. Also to work with local News Papers and to make it available on-line.

- To create awareness and organize training programmes for agencies and personnel concerned with the deployment of Indian language processing systems.
- To facilitate language technology research in Machine Aided Translation, Optical Character Recognition, Text-to-Speech and Speech Recognition for Hindi and other Indian languages.
- To organize IT localization clinics for small business to provide consultancy on use of Indian language tools in developing IT solutions and to take up development of requisite niche technologies

[Language Technology Development in India, Dr. Om Vikas, Ministry of Information Technology, New Delhi, India; Email: omvikas@mit.gov.in]

[http://www.emille.lancs.ac.uk/lesal/omvikas.pdf]

Dr. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology
Professor
Mumbai 400 076, India
Phone: +91 22 25767718
Fax: +91 22 25720290 / 25723480
Email: pb@cse.iitb.ac.in
Website: www.cse.iitb.ac.in/~pb

Dr. Pushpak Bhattacharyya is a professor of Computer Science and Engineering at the Indian Institute of Technology, Bombay. He did his bachelor of technology at IIT Kharagpur, M.Tech at IIT Kanpur and PhD at IIT Bombay. He was a visiting research fellow at the AI lab of MIT, USA.
His research interests include Natural Language Processing and Machine Learning. In the former he is contributing to the creation of lexical resources like the Indian language wordnets for Hindi and Marathi, Sense disambiguated lexicons in the context of the Universal Networking Language project of the UN and also a hierarchy of English and Hindi verbs. He has been publishing in the Journal of Machine Translation, AAAI, ACL, NLDB and such other fora.

[http://www.elda.fr/rubrique71.html]

3.5 TDIL Programme

The Technology Development for Indian Languages (TDIL) programme was launched by Ministry of Information Technology, Govt. of India in the year 1991-92. The programme aims at promotion of IT tools for Indian Languages
[http://tdil.mit.gov.in/]

Objective:

- To develop information processing tools to facilitate human machine interaction, in Indian Languages and multi-lingual knowledge resources.
- To promote the use of information processing tools for language studies and research.
- To support R&D efforts in the area of information processing in Indian Languages and to support research on Knowledge Tools: Representation, Integration, compression and learning methodologies.

- To consolidate technologies thus developed for Indian Languages and integrate these to develop innovative user products and services.

Potential IT Products & Services:
Multi-lingual Dictionaries, Thesauri,Educational software, Encyclopedia,Gyan-nidhi Creative Writing System, Translation Support Systems, OCR, Text-to-speech & Speech Recognition System, Pocket Translator, Personal Digital Assistants, Reading machine for blinds & deafs, Portals, e-governance / e-commerce / e-skills.

4 Features proposed for Indian speech databases

Some speech databases are suggested for following languages and ranked in the order given: Hindi, English as second/third language, Hindi as second/third language, Bengali, Telugu, Marathi, Tamil, Urdu. This list is ordered according to the number of speakers. An additional criterion could be "regional distribution", i.e. whether a language is used in many states or only in a few. Thus, if only speech databases for three languages can be collected, it should be Hindi, English as second/third languageand Hindi as second/third languageand.

4.1 Hindi

Language:
Only speakers with Hindi as primary language should be recorded, where "primary" means "mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].
Although closely related Urdu should not be part of this database because of practical reasons: Urdu is -- besides national language in Pakistan -- spoken throughout India, i.e. not concentrated in a few states and has only a relatively small number of speakers. Above all, it has another script than Hindi, which would be clumsy for handling prompt sheets and lexical entries in the same database.

Dialectal coverage:
Only speakers of the first 9 (northern) states with most Hindi speakers (see following table) should be recorded, where in addition administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
The number of speakers to be recorded per state (and administrative unit respectively) should be proportional to the total number of Hindi speakers in those 9 states. Eventually a bias for major cities and a threshhold for the minimal number of recorded speakers per state could be required in order to get the critical mass for an appropriate training of recognizers.

S

Table 4: States in which Hindi is among the three languages with most speakers

	State	# of Hindi speakers	% of Hindi speakers
1	Uttar Pradesh	125.348.492	90,1
2	Bihar	69.845.979	80,9
3	Madhya Pradesh	56.619.090	85,6
4	Rajasthan	39.410.968	89,6
5	Haryana	14.982.409	91,0
6	Delhi	7.690.631	81,6
7	Maharashtra	6.168.941	7,8
8	Himachal Pradesh	4.595.615	88,9
9	West Bengal	4.479.170	6,6
10	Andhra Pradesh	1.841.290	2,8
11	Punjab	1.478.993	7,3
12	Gujarat	1.215.825	2,9
13	Orissa	759.016	2,4
14	Chandigarh	392.054	61,1
15	A & N Islands	49.469	17,6
16	Tripura	45.803	1,7
17	Daman & Diu	3.645	3,6
18	Lakshadweep	217	0,4

Source : 1991 Census of India

4.2 English as second/third language

Language:
Only speakers with English as second/third language (and Indian background) should be recorded.

DIALECTAL COVERAGE:
Speakers with different first languages should be recorded. About the first 10 languages of the following table could be taken as first language.
The speakers should be resident in India, and the recordings should be carried out in India.

GENDER:
Genders should be equally distributed among the speakers.

AGE:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

NUMBER OF SPEAKERS:
The number of speakers of a given first language which have to be recorded should be in proportion to the number of speakers of that language in the country. Eventually an upper limit for Hindi and a lower limit for 'small' first languages could be required in order to get the critical mass for an appropriate training of recognizers.

Table 2 (repeated): Official languages spoken in India

	Language	*Number of speakers*	*Percentage*
1	Hindi	337,272,114	40.22%
2	Bengali	69,595,738	8.30%
3	Telugu	66,017,615	7.87%
4	Marathi	62,481,681	7.45%
5	Tamil	53,006,368	6.32%
6	Urdu	43,406,932	5.18%
7	Gujarati	40,673,814	4.85%
8	Kannada	32,753,676	3.91%
9	Malayalam	30,377,176	3.62%
10	Oriya	28,061,313	3.35%
11	Punjabi	23,378,744	2.79%
12	Assamese	13,079,696	1.56%
13	Sindhi	2,122,848	0.25%
14	Nepali	2,076,645	0.25%
15	Konkani	1,760,607	0.21%
16	Manipuri	1,270,216	0.15%
17	Kashmiri	56,693	0.01%
18	Sanskrit	49,736	0.01%
19	Other Languages	31,142,376	3.71%
	Total :	838,583,988	100.00%

Source : 1991 Census of India

4.3 Hindi as second/third language

Language:
Only speakers with Hindi as second or third language (Hindi-L2/3) and Indian background should be recorded.

DIALECTAL COVERAGE:
Speakers with different first languages should be recorded. About the first 10 languages of the above table (without Hindi and Urdu) could be taken as first language.
The speakers should be resident in India, and the recordings should be carried out in India.

GENDER:
Genders should be equally distributed among the speakers.

AGE:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

NUMBER OF SPEAKERS:
The percentage of Hindi-L2/3 speakers per state in the database should be equal to the percentage of Hindi-L2/3 of the given state in relation to all Hindi-L2/3 speakers in the states collected. Eventually a lower limit for 'small' first languages could be required in order to get the critical mass for an appropriate training of recognizers.

4.4 Bengali

Language:
Only speakers with Bengali as primary language should be recorded, where "primary" means

"mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].

Dialectal coverage:
Languages or dialects in the Bengali group according to Grierson: Central (Standard) Bengali, Western Bengali (Kharia Thar, Mal Paharia, Saraki), Southwestern Bengali, Northern Bengali (Koch, Siripuria), Rajbangsi, Bahe, Eastern Bengali (East Central, including Sylhetti), Haijong, Southeastern Bengali (Chakma), Ganda, Vanga, Chittagonian (possible dialect of Southeastern Bengali).
[http://www.ethnologue.com/show_language.asp?code=BNG]
Bengali is the main language in Bangladesh with a population of 100.000.000. Thus, it should be considered to to cover Bangladesh together with West Bengal and Tripura. In all other states the Bengali speaking population is marginal (see table 3 above). In addition administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
Eventually a bias for major cities and a threshhold for the minimal number of recorded speakers per state could be required in order to get the critical mass for an appropriate training of recognizers.

Comment:
Languages or dialects in the Bengali group according to Grierson: Central (Standard) Bengali, Western Bengali (Kharia Thar, Mal Paharia, Saraki), Southwestern Bengali, Northern Bengali (Koch, Siripuria), Rajbangsi, Bahe, Eastern Bengali (East Central, including Sylhetti), Haijong, Southeastern Bengali (Chakma), Ganda, Vanga, Chittagonian (possible dialect of Southeastern Bengali). National language. Bengali script used. Muslim.

4.5 Telugu

Language:
Only speakers with Telugu as primary language should be recorded, where "primary" means "mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].

Dialectal coverage:
It would be sufficient to record speakers of the states Andhra Pradesh (where it is the language with most speakers), Karnataka and Tamil Nadu only. In all other states the Telugu speaking population is marginal (see table 3 above). In addition, administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
Eventually a bias for major cities and a threshhold for the minimal number of recorded

speakers per state could be required in order to get the critical mass for an appropriate training of recognizers.

4.6 Marathi

Language:
Only speakers with Marathi as primary language should be recorded, where "primary" means "mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].

Dialectal coverage:
42 dialects. The dialect situation throughout the greater Marathi speaking area is complex. Dialects bordering other major language areas share many features with those languages [http://www.ethnologue.com/show_language.asp?code=MRT].
It would be sufficient to record speakers of the state Maharashtra only, where it is the language with most speakers. In all other states the Marathi speaking population is marginal (see table 3 above). Administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
Eventually a bias for major cities and a threshhold for the minimal number of recorded speakers per administrative units could be required in order to get the critical mass for an appropriate training of recognizers.

4.7 Tamil

Language:
Only speakers with Tamil as primary language should be recorded, where "primary" means "mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].

Dialectal coverage:
It would be sufficient to record speakers of the states Tamil Nadu and Pondicherry only, where it is the language with most speakers. In all other states the Tamil speaking population is marginal (see table 3 above). In addition, administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
Eventually a bias for major cities and a threshhold for the minimal number of recorded speakers per state could be required in order to get the critical mass for an appropriate training of recognizers.

4.8 Urdu

Since Urdu has a specific history, it is spoken in many states by a relatively low (but not neglectable) percentage of speakers (see table 3 above). Urdu is national language in Pakistan.

Language:
Only speakers with Urdu as primary language should be recorded, where "primary" means "mother tongue" or rather "Matri-Bhasa", the sanskritic word for "language ordinarily used" [http://www.censusindia.net/cendat/language/intro_language.html].

Dialectal coverage:
Speakers of the states Andhra Pradesh, Bihar, Karnataka, Maharashtra, Uttar Pradesh and West Bengal (see table 3 above) should be recorded. Administrative units (districts etc.) should be taken into account in order to cover intra-state dialect variation.
The speakers should be resident in India, and the recordings should be carried out in India.

Gender:
Genders should be equally distributed among the speakers.

Age:
Age groups should be according SpeechDat/SPEECON convention, eventually with a bias of the younger age groups.

Number of speakers:
The percentage of Urdu speakers per state in the database should be equal to the percentage of Urdu speakers of the given state in relation to all Urdu speakers in the states collected. Eventually a bias for major cities and a threshhold for the minimal number of recorded speakers per state could be required in order to get the critical mass for an appropriate training of recognizers.
Comment:
Urdu is also a language of Pakistan, where it is the mother tongue of 10,719,000 speakers. Thus, it should be considered to to cover India together with Pakistan.

References

[http://adaniel.tripod.com/Languages1.htm]
[http://adaniel.tripod.com/Languages2.htm]
[http://adaniel.tripod.com]
[http://cyberjournalist.org.in/census/cenindia.html]
[http://en2.wikipedia.org/wiki/Hindustani]
[http://indiansaga.com/languages/index.html > hindi]
[http://indiansaga.com/languages/index.html > home]
[http://tdil.mit.gov.in]
[http://www.birminghamuk.com/asianlanguage.htm]
[http://www.censusindia.net/cendat/datatable26.html]
[http://www.censusindia.net/cendat/language/intro_language.html]
[http://www.cia.gov/cia/publications/factbook/geos/in.html]
[http://www.ciil.org/languages/map4.html]
[http://www.cs.colostate.edu/~malaiya/hindiint.html]
[http://www.elda.fr/rubrique71.html]
[http://www.elra.info/
[http://www.emille.lancs.ac.uk/beta.htm]
[http://www.emille.lancs.ac.uk/home.htm]
[http://www.emille.lancs.ac.uk/lesal/omvikas.pdf]
[http://www.esamskriti.com/html/new_inside.asp?cat_name=cultphil&cid=513&sid=9001]
[http://www.ethnologue.com/show_iso639.asp?code=hi]
[http://www.ethnologue.com/show_language.asp?code=HND]
[http://www.ethnologue.com/show_language.asp?code=BNG]
[http://www.ethnologue.com/show_language.asp?code=MRT]
[http://www.india4world.com/indian-language/index.shtml]
[http://www.ldc.upenn.edu]
[http://www.ling.lancs.ac.uk/corplang/emille/default.htm]
[http://www.linguaphone.co.uk/language.cfm?language_id=14]
[India Today, August 18, 1997]
[Language Technology Development in India, Dr. Om Vikas, Ministry of Information Technology, New Delhi, India; Email: omvikas@mit.gov.in]
[Source: Siemens Business Information Report ST030401: India's Telecommunications Market: Implementation Priorities and Differences with China, March 2003]
[Source: Siemens Business Information Report ST030403: Automotive Market in India, April 2003]

Additional sources

[http://adaniel.tripod.com/languagelist.htm]
[http://adaniel.tripod.com/Languages.htm]
[http://en2.wikipedia.org/wiki/Hindi]
[http://indiansaga.com/languages/index.html > home]
[http://www.esamskriti.com/html/new_inside.asp?cat_name=cultphil&cid=515&sid=9001&count1=2]

S

[http://www.linguaphone.co.uk/language.cfm?language_id=14]

S

Annex 1

ENGLISH AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991

ENGLISH AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991

Scheduled Languages		Number of persons who know English as the second language	Percentage of Col.3 to total speakers of the language	Number of persons know English as the third language	Percentage of Col.5 to total speakers of the language
Name	Total Speakers				
1	2	3	4	5	6
1.Assamese	13,079,696	1,322,488	10.11	538,088	4.11
2.Bengali	69,595,738	5,052,456	7.16	1,236,168	1.78
3.Gujarati	40,673,814	620,265	1.52	3,691,582	9.08
4.Hindi	337,272,114	27,569,676	8.17	2,288,498	0.68
5.Kannada	32,753,676	3,091,484	9.44	832,763	2.54
6.Kashmiri	56,693	7,638	13.47	10,841	19.12
7.Konkani	1,760,607	381,500	21.67	232,106	13.18
8.Malayalam	30,377,176	6,692,407	22.03	704,134	2.32
9.Manipuri	1,270,216	245,230	19.31	88,507	6.97
10.Marathi	62,481,681	1,082,168	1.73	6,479,857	10.37
11.Nepali	2,076,645	84,187	4.05	86,136	4.15
12.Oriya	28,061,313	2,933,330	10.45	619,819	2.21
13.Punjabi	23,378,744	1,467,992	6.28	4,076,792	17.44
14.Sanskrit	49,736	2,651	5.33	4,714	9.48
15.Sindhi	2,122,848	125,724	5.92	287,160	13.53
16.Tamil	53,006,368	7,092,118	13.38	355,490	0.67
17.Telugu	66,017,615	5,460,642	8.27	1,867,606	2.83
18.Urdu	43,406,932	1,370,343	3.16	2,039,927	4.70
Total	**807,441,612**	**64,602,299**	**800**	**25,440,188**	**3.15**

Note :

1.The Statement excludes the figures for Jammu & Kashmir where the 1991 Census was not held due to disturbed conditions.

2.The 1991 Census could not be conducted in 33 villages of Akrani and Akkalkuwa tahsils of Dhule district of Maharashtra. The population of these villages (i.e. 16,052 Persons) has been obtained from secondary sources and included in the population of Maharashtra and India. However, their language data are not available.

S

Annex 2

HINDI AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991

HINDI AS SECOND AND THIRD LANGUAGE AMONG THE SPEAKERS OF SCHEDULED LANGAUGES – 1991

Scheduled Languages		Number of persons who know Hindi as the second language	Percentage of Col.3 to total speakers of the language	Number of persons know Hindi as the third language	Percentage of Col.5 to total speakers of the language
Name	Total Speakers				
1	2	3	4	5	6
1.Assamese	13,079,696	1,153,300	8.82	1,060,637	8.11
2.Bengali	69,595,738	2,776,137	3.99	1,843,962	2.65
3.Gujarati	40,673,814	9,001,474	22.13	723,326	1.78
4.Hindi	337,272,114	@	@	@	@
5.Kannada	32,753,676	1,272,857	3.89	1,666,139	5.09
6.Kashmiri	56,693	23,903	42.16	5,872	10.36
7.Konkani	1,760,607	150,995	8.58	289,949	16.47
8.Malayalam	30,377,176	817,081	2.69	4,974,577	16.38
9.Manipuri	1,270,216	127,285	10.02	181,456	14.29
10.Marathi	62,481,681	14,861,627	23.79	1,252,031	2.00
11.Nepali	2,076,645	482,855	23.25	253,575	12.21
12.Oriya	28,061,313	1,280,564	4.56	1,908,953	6.80
13.Punjabi	23,378,744	7,189,178	30.75	1,286,207	5.50
14.Sanskrit	49,736	22,057	44.35	1,173	2.36
15.Sindhi	2,122,848	864,893	40.74	210,338	9.91
16.Tamil	53,006,368	372,954	0.70	453,342	0.86
17.Telugu	66,017,615	2,164,963	3.28	3,124,524	4.73
18.Urdu	43,406,932	7,205,794	16.60	1,740,527	4.01
Total	**807,441,612**	**49,767,917**	**6.16**	**20,976,588**	**2.60**

Note :

1.The Statement excludes the figures for Jammu & Kashmir where the 1991 Census was not held due to disturbed conditions.

2.The 1991 Census could not be conducted in 33 villages of Akrani and Akkalkuwa tahsils of Dhule district of Maharashtra. The population of these villages (i.e. 16,052 Persons) has been obtained from secondary sources and included in the population of Maharashtra and India. However, their language data are not available.

@ Not applicable.

Annex 3

Languages in Descending Order of Strength – India, States and Union Territories – 1991 Census

Languages in Descending Order of Strength – India, States and Union Territories – 1991 Census

Serial No.	Language	Number of persons who returned the language as their mother tongue		Serial No.	Language	Number of persons who returned the language as their mother tongue	
		Total	Percent to toal piopulation of India			Total	Percent to toal piopulation of India
1	2	3	4	1	2	3	4
	India						
1	Hindi	337,272,114	22.000	23	Khandeshi	973,709	0.116
2	Bengali	69,595,738	8.299	24	Ho	949,216	0.113
3	Telugu	66,017,615	7.873	25	Khasi	912,283	0.109
4	Marathi	62,481,681	7.451	26	Mundari	861,378	0.103
5	Tamil	53,006,368	6.321	27	Tripuri	694,940	0.083
6	Urdu	43,406,932	5.176	28	Garo	675,642	0.081
7	Gujarati	40,673,814	4.850	29	Kui	641,662	0.077
8	Kannada	32,753,676	3.906	30	Lushai/Mizo	538,842	0.064
9	Malayalam	30,377,176	3.623	31	Halabi	534,313	0.064
10	Oriya	28,061,313	3.346	32	Korku	466,073	0.056
11	Punjabi	23,378,744	2.788	33	Munda	413,894	0.049
12	Assamese	13,079,696	1.560	34	Miri/Mishing	390,583	0.047
13	Bhili/Bhilodi	5,572,308	0.665	35	Karbi/Mikir	366,229	0.044
14	Santali	5,216,325	0.622	36	Savara	273,168	0.033
15	Gondi	2,124,852	0.253	37	Koya	270,994	0.032
16	Sindhi	2,122,848	0.253	38	Kharia	225,556	0.027
17	Nepali	2,076,645	0.248	39	Khond/Kondh	220,783	0.026
18	Konkani	1,760,607	0.210	40	English	178,598	0.021
19	Tulu	1,552,259	0.185	41	Nissi/Dafla	173,791	0.021
20	Kurukh/Oraon	1,426,618	0.170	42	Ao	172,449	0.021
21	Manipuri	1,270,216	0.151	43	Sema	166,157	0.020
22	Bodo/Boro	1,221,881	0.146	44	Kisan	162,088	0.019

Serial No.	Language	Number of persons who returned the language as their mother tongue	
		Total	Percent to toal piopulation of India
1	2	3	4
45	Adi	158,409	0.019
46	Rabha	139,365	0.017
47	Konyak	137,722	0.016
48	Malto	108,148	0.013
49	Thado	107,992	0.013
50	Tangkhul	101,841	0.012
51	Kolami	98,281	0.012
52	Angami	97,631	0.012
53	Kurgi/Kodagu	97,011	0.012
54	Dogri	89,681	0.011
55	Dimasa	88,543	0.011
56	Lotha	85,802	0.010
57	Mao	77,810	0.009
58	Tibetan	69,146	0.008
59	Kabui	68,925	0.008
60	Phom	65,350	0.008
61	Hmar	65,204	0.008
62	Kinnauri	61,794	0.007
63	Bishnupuriya	59,233	0.007
64	Kuki	58,253	0.007
65	Kashmiri	56,693	0.007
66	Bhotia	55,483	0.007
67	Sanskrit	49,736	0.006
68	Paite	49,237	0.006
69	Chakru/Choki	48,207	0.006
70	Sangtam	47,461	0.006
71	Yimchungre	47,227	0.006
72	Bhumij	45,302	0.005
73	Parji	44,001	0.005
74	Monpa	43,226	0.005
75	Wancho	39,600	0.005
76	Lepcha	39,342	0.005
77	Rengma	37,521	0.004
78	Zeliang	35,079	0.004
79	Lalung	33,746	0.004
80	Chang	32,478	0.004
81	Chakhesang	30,985	0.004
82	Nocte	30,441	0.004
83	Halam	29,322	0.003
84	Mishmi	29,000	0.003
85	Koda/Kora	28,200	0.003
86	Limbu	28,174	0.003
87	Gadaba	28,158	0.003
88	Mogh	28,135	0.003
89	Tangsa	28,121	0.003
90	Korwa	27,485	0.003
91	Liangmei	27,478	0.003
92	Lahnda	27,386	0.003
93	Nicobarese	26,261	0.003
94	Vaiphei	26,185	0.003
95	Koch	26,179	0.003
96	Jatapu	25,730	0.003

Serial No.	Language	Number of persons who returned the language as their mother tongue	
		Total	Percent to toal piopulation of India
1	2	3	4
97	Khiemnungan	23,544	0.003
98	Lakher	22,947	0.003
99	Zemi	22,634	0.003
100	Lahauli	22,027	0.003
101	Arabic/Arbi	21,975	0.003
102	Deori	17,901	0.002
103	Konda	17,864	0.002
104	Juang	16,858	0.002
105	Sherpa	16,105	0.002
106	Zou	15,966	0.002
107	Pawi	15,346	0.002
108	Maring	15,268	0.002
109	Gangte	13,695	0.002
110	Kom	13,548	0.002
111	Khezha	13,004	0.002
112	Anal	12,156	0.001
113	Pochury	11,231	0.001
114	Maram	10,144	0.001
	Total of Other Languages	565,949	0.067
	All Languages Total	838,567,936	100.000

S

Annex 4

LANGUAGES AND MOTHER TONGUES AND THEIR STRENGTH IN 1991 CENSUS

LANGUAGES AND MOTHER TONGUES AND THEIR STRENGTH IN 1991 CENSUS

Presented below is an alphabetical abstract of languages and the mother tongues with strength of 10,000 and above at all India level, grouped under each language. There are a total of 114 languages and 216 mother tongues. The 18 languages specified in the Eighth Schedule to the Constitution of India are given in Part A and languages other than those specified in the Eighth Schedule (numbering 96) are given in Part B.

The population of Jammu and Kashmir is not included in these figures, as the 1991 census was not conducted there due to disturbed conditions.

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue	Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
1.ADI	**158,409**	**6.BHILI/BHILODI (Continued)**	
1.Adi	70,739	2.Barel	467,328
2.Adi Gallong/Gallong	45,616	3.Bhilali	468,519
3.Adi Miniyong/Miniyong	18,417	4.Bhili/Bhilodi	1,891,109
Others	23,637	5.Dhodia	15,770
2.ANAL	**12,156**	6.Gamti/Gavit	12,500
1.Anal	10,355	7.Kokna/Kokni/Kukna	130,526
Others	1,801	8.Mawchi	80,850
3.ANGAMI	**97,631**	9.Paradhi	33,683
1.Angami	58,567	10.Pawri	123,078
Others	39,064	11.Rathi	20,617
4.AO	**172,449**	12.Tadavi	19,067
1.Ao	170,911	13.Varli	91,763
Others	1,538	14.Wagdi	2,156,851
5.ARABIC/ARBI	**21,975**	Others	38,546
1.Arabic/Arbi	21,975	**7.BHOTIA**	**55,483**
6.BHILI/BHILODI	**5,572,308**	1.Bhotia	55,413
1.Baori	22,101	Others	70

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
8.BHUMIJ	**45,302**
1.Bhumij	39,389
Others	5,913
9.BISHNUPURIYA	**59,233**
1.Bishnupuriya Manipuri/ Manipuri Bishnupuriya	55,647
Others	3,586
10.BODO/BORO	**1,221,881**
1.Bodo/Boro	1,201,957
2.Kachari	11,588
Others	8,336
11.CHAKHESANG	**30,985**
1.Chakhesang	30,985
12.CHAKRU/CHOKRI	**48,207**
1.Chakru/Chokri	48,207
13.CHANG	**32,478**
1.Chang	32,478
14.COORGI/KODAGU	**97,011**
1.Coorgi/Kodagu	97,011
15.DEORI	**17,901**
1.Deori	17,901
16.DIMASA	**88,543**
1.Dimasa	87,284
Others	1,259
17.DOGRI	**89,681**
1.Dogri	89,648
Others	33
18.ENGLISH	**178,598**
1.English	178,598
19.GADABA	**28,158**
1.Gadaba	28,016
Others	142
20.GANGTE	**13,695**
1.Gangte	13,577
Others	118
21.GARO	**675,642**
1.Garo	671,908
Others	3,734
22.GONDI	**2,124,852**
1.Dorli	31,885
2.Ganda/Gando	12,845
3.Gondi	1,958,883
4.Kalari	12,323
5.Maria	80,436
6.Muria	16,556
Others	11,924
23.HALABI	**534,313**
1.Halabi	533,839

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
23.HALABI (Continued) Others	
Others	474
24.HALAM	**29,322**
1.Halam	12,147
Others	17,175
25.HMAR	**65,204**
1.Hmar	65,204
26.HO	**949,216**
1.Ho	942,845
Others	6,371
27.JATAPU	**25,730**
1.Jatapu	25,503
Others	227
28.JUANG	**16,858**
1.Juang	16,858
29.KABUI	**68,925**
1.Kabui	28,475
2.Rongmei	40,324
Others	126
30.KARBI/MIKIR	**366,229**
1.Karbi/Mikir	363,715
Others	2,514
31.KHANDESHI	**973,709**
1.Ahirani	836,925
31.KHANDESHI (Continued)	
2.Dangi	108,947
3.Khandeshi	19,192
Others	8,645
32.KHARIA	**225,556**
1.Kharia	224,786
Others	770
33.KHASI	**912,283**
1.Khasi	700,047
2.Pnar/Synteng	169,388
3.War	26,735
Others	16,113
34.KHEZHA[1]	**13,004**
Mother tongues grouped	13,004
35.KHIEMNUNGAN	**23,544**
1.Khiemnungan	23,544
36.KHOND/KONDH	**220,783**
1.Khond/Kondh	197,762
2.Kuvi	22,450
Others	571
37.KINNAURI	**61,794**
1.Kinnauri	61,225
Others	569

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue	Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
38.KISAN	**162,088**	**47.KOYA**	**270,994**
1.Kisan	162,088	1.Koya	270,994
39.KOCH	**26,179**	**48.KUI**	**641,662**
1.Koch	24,296	1.Kui	641,600
Others	1,883	Others	62
40.KODA/KORA	**28,200**	**49.KUKI**	**58,263**
1.Koda/Kora	27,703	1.Kuki	52,511
Others	497	Others	5,752
41.KOLAMI	**98,281**	**50.KURUKH/ORAON**	**1,426,618**
1.Kolami	98,281	1.Kurukh/Oraon	1,417,856
42.KOM	**13,548**	Others	8,762
1.Kom	13,548	**51.LAHAULI**	**22,027**
43.KONDA[2]	**17,864**	1.Lahauli	21,907
1.Kodu	10,009	Others	120
Others	7,855	**52.LAHNDA**[3]	**27,386**
44.KONYAK	**137,722**	1.Multani	20,589
1.Konyak	137,722	Others	6,797
45.KORKU	**466,073**	**53.LAKHER**	**22,947**
1.Korku	433,852	1.Lakher	22,947
2.Muwasi	28,704	**54.LALUNG**	**33,746**
Others	3,517	1.Lalung	33,746
46.KORWA	**27,485**	**55.LEPCHA**	**39,342**
1.Koraku	15,716	1.Lepcha	39,342
2.Korwa	11,169	**56.LIANGMEI**	**27,478**
Others	600	1.Liangmei	23,517

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue	Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
56.LIANGMEI (Continued)		**66.MOGH (Continued)**	
Others	3,961	Others	9
57.LIMBU	**28,174**	**67.MONPA**	**43,226**
1.Limbu	18,931	1.Monpa	40,620
Others	9,243	Others	2,606
58.LOTHA	**85,802**	**68.MUNDA**	**413,894**
1.Lotha	85,802	1.Kol	89,538
59.LUSHAI/MIZO	**538,842**	2.Munda	319,977
1.Lushai/Mizo	537,627	Others	4,379
Others	1,215	**69.MUNDARI**	**861,378**
60.MALTO[4]	**108,148**	1.Mundari	852,398
1.Pahariya	106,902	Others	8,980
Others	1,246	**70.NICOBARESE**	**26,261**
61.MAO	**77,810**	1.Nicobarese	26,261
1.Mao	63,178	**71.NISSI/DAFLA**	**173,791**
Others	14,632	1.Apatani	21,453
62.MARAM	**10,144**	2.Bangni	35,339
1.Maram	10,144	3.Nishang	16,976
63.MARING	**15,268**	4.Nissi/Dafla	68,176
1.Maring	15,268	5.Tagin	31,845
64.MIRI/MISHING	**390,583**	Others	2
1.Miri/Mishing	390,567	**72.NOCTE**	**30,441**
Others	16	1.Nocte	24,574
65.MISHMI[5]	**29,000**	Others	5,867
1.Mothers tongues grouped	29,000	**73.PAITE**	**49,237**
66.MOGH	**28,135**	1.Paite	49,052
1.Mogh	28,126	Others	185

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue	Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
74.PARJI[6]	**44,001**	**83.SEMA**	**166,157**
1.Dhurwa	41,278	1.Sema	166,157
Others	2,723	**84.SHERPA**	**16,105**
75.Pawi	**15,346**	1.Sherpa	16,105
1.Pawi	15,346	**85.TANGKHUL**	**101,841**
76.PHOM	**65,350**	1.Tangkhul	101,805
1.Phom	65,350	Others	36
77.POCHURY	**11,231**	**86.TANGSA**[7]	**28,121**
1.Pochury	10,750	Mother tongues grouped	28,121
Others	481	**87.THADO**	**107,992**
78.RABHA	**139,365**	1.Thado	102,001
1.Rabha	139,341	Others	5,991
Others	24	**88.TIBETAN**	**69,416**
79.RENGMA	**37,521**	1.Tibetan	62,891
1.Rengma	37,521	Others	6,525
80.SANGTAM	**47,461**	**89.TRIPURI**	**694,940**
1.Sangtam	47,454	1.Kokbarak	517,664
Others	7	2.Reang	94,421
81.SANTALI	**5,216,325**	3.Tripuri	81,119
1.Karmali	233,766	Others	1,736
2.Mahili	14,995	**90.TULU**	**1,552,259**
3.Santali	4,915,808	1.Tulu	1,550,334
Others	51,756	Others	1,925
82.SAVARA	**273,168**	**91.VAIPHEI**	**26,185**
1.Savara	273,165	1.Vaiphei	26,185
Others	3		

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue		Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue
92.WANCHO	**39,600**		**95.ZEMI**	**22,634**
1.Wancho	39,600		1.Zemi	22,627
93.YIMCHUNGRE	**47,227**		Others	7
1.Yimchungre	35,598		**96.ZOU**	**15,966**
Others	11,629		1.Zou	15,966
94.ZELIANG	**35,079**			
1.Zeliang	35,079			

Note:

[1]**KHEZHA :** Since none of the mother tongues grouped under this language fulfils the criterion of 10,000 or more speakers at the all-India level, the name of no mother tongue appears under this language.

[2]**KONDA :** A number of mother tongues including Konda have been grouped together under the language name Konda on the basis of their linguistic affiliation but out of these only Kodu fulfils the criterion of 10,000 or more speakers at the all-India level and hence only Kodu appears by name as a mother tongues and the rest are included under 'Others'.

[3]**LAHNDA :** A number of mother tongues including Lahnda have been grouped together under the language name Lahnda on the basis of the Linguistic Survey of India (LSI) classification of G.A.Grierson, but out of these only Multani fulfils the criterion of 10,000 or more speakers at the all-India level and hence only Multani appears by name as a mother tongues and the rest are included under 'Others'. Since the main area of Lahnda speech is now in Pakistan, India's population of Lahnda speakers is rather small.

[4]**MALTO :** Malto represents a number of mother tongues including Malto itself of which only Pahariya fulfils the criterion of 10,000 or more speakers at the all-India level and hence only Pahariya appears by name as a mother tongue and the rest are included under 'Others'. The name Malto is used by the people to denote their language and its status as an independent language has been established.

[5]**MISHMI :** Since none of the mother tongues grouped under this language fulfils the criterion of 10,000 speakers or more at the all-India level, the name of no mother tongue appears under this language.

[6]**PARJI :** A number of mother tongues including Parji have been grouped together under the language name Parji, but out of these only Dhurwa fulfils the criterion of 10,000 or more speakers at the all-India level and hence only Dhurwa appears by name as a mother tongues and the rest are included under 'Others'.

Part B – Languages not specified in the Eighth Schedule (non-Scheduled Languages)

Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue		Name of language and mother tongue(s) grouped under each language	Number of persons who returned the language (and the mother tongues grouped under each) as their mother tongue

[7]**TANGSA :** Since none of the mother tongues grouped under this language fulfils the criterion of 10,000 or more speakers at the all-India level, the name of no mother tongue appears under this language.