Andovar

LANGUAGES OF INDIA

WHITE PAPER

About Andovar

Andovar is a global provider of multilingual content solutions. Our services range from text translation and content creation, through audio and video recording, to turnkey localization of websites, software, eLearning and games. Our headquarters is in Singapore, and offices in Thailand, Colombia, USA and India.

About This White Paper

Andovar opened an office in Kolkata in 2014 to better serve our international clients and to add Indian language translation and audio recording to our suite of services. This white paper is an attempt to understand the localization situation when it comes to the major languages spoken in India. This has proven to be a challenging task. Not only are there dozens of languages and over ten scripts in everyday use in India, but it is also a rapidly developing country, with a huge and growing economy and new intiatives related to languages, encoding and translation technology support appear almost monthly. Not much has been written to date to give a holistic overview of the localization landscape and hopefully this white paper will be a useful reference to language professionals, even if it is far from perfect. As such, there may be mistakes or missing information which will be added in future updates. Please contact marketing@andovar.com with any questions or suggestions.

The data on number or speakers and native speakers of Indian languages is unreliable, with different sources following different conventions when deciding what is a language and what is merely a dialect*. Our first source is the Indian Census from 2001 (www.censusindia.gov.in/Census_Data_2001/), which provided comprehensive information about the whole country, but is now outdated. Unfortunately, the new data from the 2011 Census has not been released yet. To supplement this, our second source is Ethnologue (www.ethnologue.com) which provides more up to date data in most cases, as well as information about number of speakers outside of India. The source used is provided every time.

To find out more about Andovar, visit:

www.andovar.com

* Linguists generally distinguish the terms "language" and "dialect" on the basis of mutual intelligibility, however the Indian census uses two specific classifications: (1) "language" and (2) "mother tongue". The "mother tongues" are grouped within each "language". Many "mother tongues" so defined would be considered a language rather than a dialect by linguistic standards. This is especially so for many "mother tongues" with tens of millions of speakers that are officially grouped under the "language" of Hindi.

Andovar

Table of Contents

Introduction

India is a vast country and plurality is its hallmark. It is a federal union comprising 29 states, 7 union territories, and hundreds of languages and dialects. India is the seventh-largest country by area and second-most populous country in the world. The economy of India is the 10th largest in the world by nominal GDP and the third-largest by purchasing power parity. Each state of India, almost the size of a European country, has its own distinct culture and language. Most have their own distinct scripts. As a matter of fact, out of the 25 scripts invented by humankind, 10 are from India. For an Indian person travelling to another state in India, it is like going to a foreign country, i.e. the local language is not theirs.

As purchasing power of consumers in India grows, many international businesses seek to sell their products and services to the nation's linguistically diverse population. At the same time, many large organizations in India seek to expand into other markets. Considering India's linguistic, ethnic, social, cultural, and geographical diversities along with a growing economy, localization should be one of the salient features for international businesses. At the same time, localization into Indian languages is not a trivial matter. How does one even begin to approach a land with as much linguistic variety?

Every two miles the water doth change, and every four the dialect.

— Indian proverb —

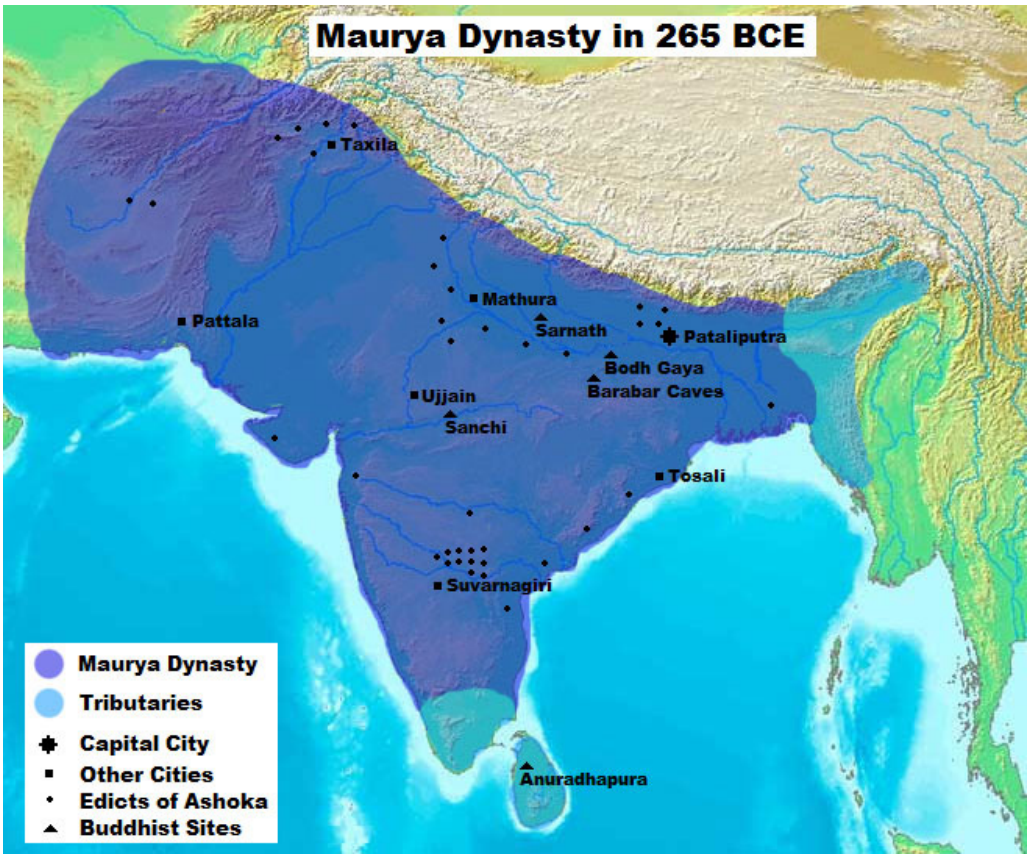
The Past

A Legacy of Multilingualism

The Indus Valley Civilization, which dominated the northwestern part of the Indian subcontinent from c. 3300 to 1300 BCE in present-day Pakistan and Northwest India, was the first major civilization in South Asia. After it collapsed at the start of the second millennium BCE, it was followed by the Iron Age Vedic civilization. The Vedic period is characterized by Indo-Aryan culture associated with the Vedic texts, which are sacred to Hindus, and were orally composed in the Vedic Sanskrit language. The Vedas are some of the oldest extant texts in India. The Vedic period, lasting from about 1750 to 500 BCE, contribute the foundations of Hinduism and other cultural aspects of Indian subcontinent to the present day.

Most of the subcontinent was then conquered by the Maurya Empire led by Ashoka the Great during the 4th and 3rd centuries BCE, who rapidly expanded westwards across Central and Western India, taking advantage of the disruption of local powers in the wake of the western withdrawal by Alexander the Great's Greek armies. The Maurya Empire was one of the world's largest empires in its time and the largest ever on the Indian subcontinent. It was in those centuries that many holy texts of Buddhism, Jainism, and Hinduism were written in Pali, Sanskrit, Tamil, and Ardhamagadhi languages.

Map of the Maurya Dynasty



Source: Wikipedia

Parts of India were ruled by various so-called middle kingdoms for the next 1500 years. This period, witnessing a Hindu religious and intellectual resurgence, is known as the Classical Period or Golden Age of India. It was then that aspects of Indian civilization, administration, culture, and religion (Hinduism and Buddhism) spread across much of Asia, while kingdoms in Southern India had maritime business links with the Roman Empire from around 77 CE. The early medieval period of Indian mathematics influenced the development of mathematics and astronomy in the Arab world and Hindu numerals were introduced. Sanskrit, Sauraseni Prakrit, and, later, Sauraseni Apabhramsa served as languages of interregional communication.

Map of the British Indian Empire from Imperial Gazetteer of India - 1909

Source: Wikipedia

The middle period ended in the 13th century with the rise of the Muslim Delhi Sultanate. The Delhi Sultanate brought Persian and Arabic languages into prominence and ruled a major part of Northern India. It declined in the late 14th century, which saw the emergence of several powerful Hindu states. In the 16th century, Mughal rule came from Central Asia to cover most of the northern parts of India. Beginning with the Moghul Emperor Akbar's reign, Persian was used as the official language and over time gained such prestige that it enjoyed continued use as the official language in North India even after the end of Muslim rule. The Mughal Empire suffered a gradual decline in the early 18th century, which provided opportunities for the Maratha Empire, Sikh Empire, and Mysore Kingdom to exercise control over large areas on the subcontinent. During the 17th and 18th centuries, Hindi and Urdu developed into interregional communication languages.

Beginning in the late 1700s and over the next one hundred years, large areas of India were annexed by the British East India Company. English replaced Persian as the official language in 1837, though Persian and, to a lesser extent, Hindi were retained in some capacity at lower levels of administration. Similarly, as in many other parts of the post-colonial world, English became the language of the intellectual elite. During the first half of the 20th century, a nationwide struggle for independence was launched, and as a result, the subcontinent gained independence from the United Kingdom in 1947, after the British provinces were partitioned into India and Pakistan.

The Question of the Official Language of India

After gaining independence from the British in 1947, the leaders of the new Indian nation recognized an opportunity to unite the many regions of India with a common, universal language. Mahatma Gandhi felt that this was essential to the emergence of India as a bona fide nation. The task was crucial for the Indian government, but difficult, because choosing any one language would be controversial. Starting years before independence, Gandhi tirelessly supported Hindustani, which is a kind of compromise between Hindi and Urdu, as the best choice for a national language. However, after partition into India and Pakistan in 1947, and the subsequent emigration of millions of Muslims, Hindu political leaders saw little need for Gandhi's concessions to the Muslims. Instead, they focused on Hindi and left Urdu and Hindustani to their own fates.

Hindustani

Hindustani (हिन्दुस्तानी , ہندوستانی) is the *lingua franca* of North India and parts of Pakistan. It has two official forms: Modern Standard Hindi and Modern Standard Urdu and may be called Hindi-Urdu when taken together.

The colloquial languages are indistinguishable, and even though the official standards are nearly identical in grammar, they differ in literary conventions and in academic and technical vocabulary, with Urdu retaining stronger Persian, Central Asian and Arabic influences, and Hindi relying more heavily on Sanskrit. Before the Partition of India, the terms Hindustani, Urdu, and Hindi were synonymous; all covered what would be called Urdu and Hindi today.

Though it did not have an assured dominance over the other languages in India, Hindi seemed the clearest choice from the beginning. English, despite its prominence and haphazard yet even distribution throughout the nation was unacceptable for several reasons. As the language of the colonial power, which had just been ousted, English was seen by many as a symbol of oppression. More importantly, a foreign tongue such as English would not contribute to the national identity in the way that an indigenous one could.

Finally, English was spoken by only about 1% of India's population, while Hindi claimed the greatest number of speakers of all the Indian languages and was closely related to several of the other most widely spoken ones. Although linguistically it was not related to the South Indian languages of the Dravidian family, it was also thought that Hindi would not be entirely foreign to South Indians because of the strong Sanskrit influence they shared. Whether or not this thinking was correct, Hindi was chosen as the official language alongside Prime Minister Jawaharlal Nehru's assurance that it would never be imposed on people in non-Hindi areas. Therefore, the Indian constitution of 1950 declared Hindi, written in Devanagari script, to be the official language of the Union. Unless parliament decided otherwise, the use of English for official purposes was to cease after 15 years on 26 January 1965.

Despite planning, this decision led to violent protests by groups who felt that Hindi was being imposed on them, especially in the Dravidian-speaking states in the South. As a result, parliament enacted the Official Languages Act in 1963, which provided for the continued use of Hindi for official purposes along with English even after 1965. The Act was amended in 1967 to stipulate that the use of English would not be ended until a resolution to that effect was passed by the legislature of every state that had not adopted Hindi as its official language as well as by each house of the Indian parliament.

The Present

The Eighth Schedule Languages

Nowadays, the basis for official status of languages in India remains in the constitution along with various amendments, especially the so-called Eighth Schedule. The 1950 Constitution of India listed 14 languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya/ Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu) on the Eighth Schedule. Since then, the list has been expanded three times as follows: 1. in 1967 to include Sindhi 2. in 1992 to include Konkani, Manipuri, and Nepali; and 3. in 2003 to include Bodo, Santali, Maithili, and Dogri. Over 30 additional languages are under consideration for inclusion on the schedule and the list is expected to grow in the future.

22 Languages Included on the Eighth Schedule to the Indian Constitution

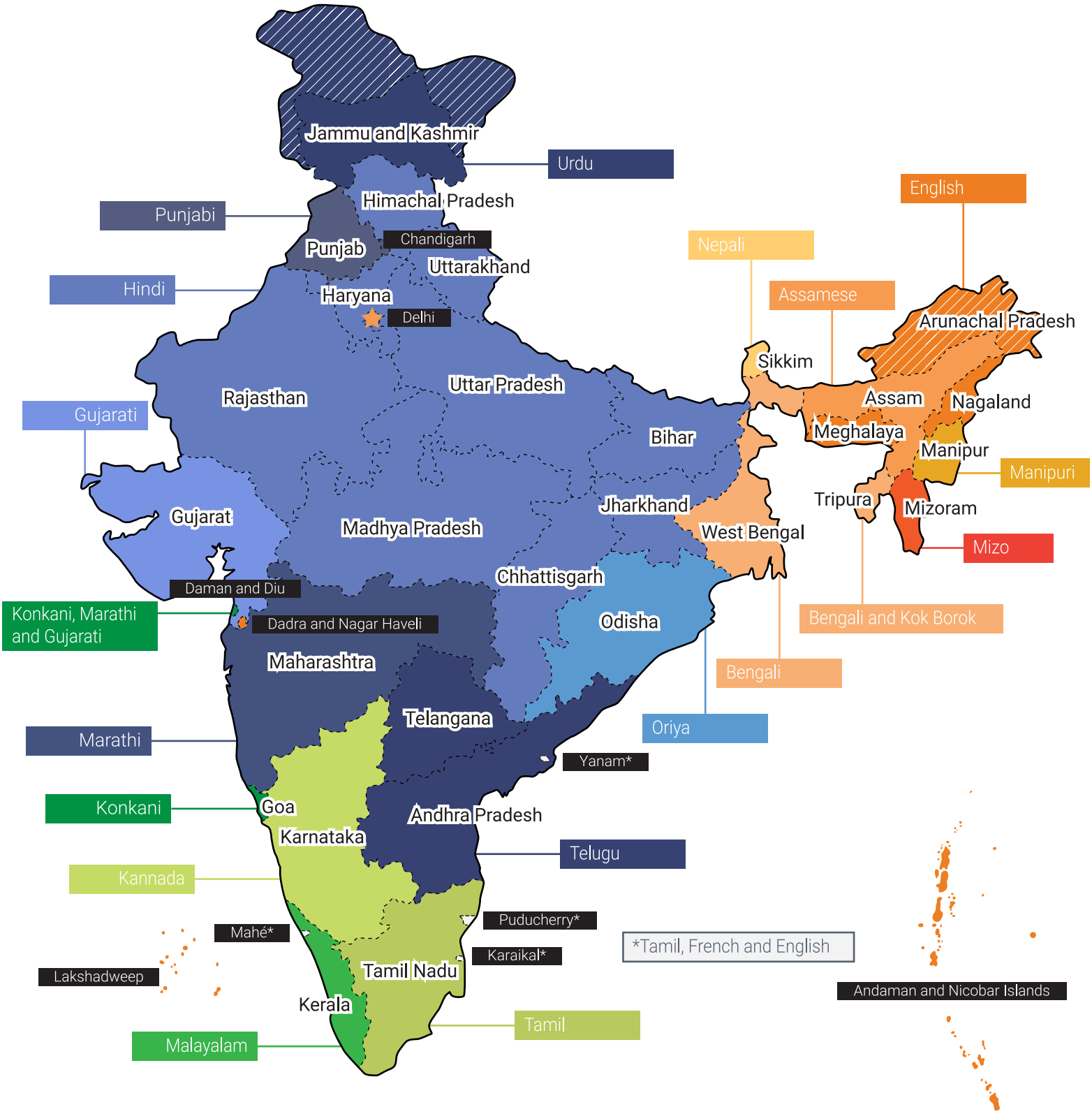
Assamese	Kannada	Marathi	Sindhi
Bengali	Kashmiri	Nepali	Tamil
Bodo	Konkani	Oriya/ Odia	Telugu
Dogri	Maithili	Punjabi	Urdu
Gujarati	Malayalam	Sanskrit	
Hindi	Manipuri	Santhali	

At the time the constitution was enacted, inclusion on the Eighth Schedule list meant that the language was entitled to representation on the Official Languages Commission and that the language would influence and enrich Hindi, the official language of the Union. Since then, the list has acquired greater significance. The Government of India is now obligated to take measures for the development of these languages, such that "they grow rapidly in richness and become effective means of communicating modern knowledge." In addition, a candidate appearing in an examination conducted for public service at a higher level is entitled to use any of these languages as the medium.

Although Hindi written in the Devanagari script remains as the only official language at the national level, the government continues to use English in addition to Hindi for official purposes as a "subsidiary official language," but also prepares and executes programs to progressively increase its use of Hindi.

Each of India's 29 states and 7 union territories has chosen an official language to be used in formal proceedings. The "Hindi-speaking belt" (see map on the next page) is made up of the states and union territories that have Hindi as the principal official language. The rest use their own state or regional language along with English translations for those who don't understand it. While the Eighth Schedule languages and English make up the core of linguistic variety in India, many other languages enjoy a semi-official status.

Map of Official Languages in Indian States and Union Territories



The status of Arunachal Pradesh and parts of Jammu and Kashmir states is disputed by Pakistan, China and India, as marked by diagonal lines on the map. The borders shown on this map are not meant as a statement in this dispute by Andovar.

Andovar

Classical Languages

In 2004, the Government of India declared that languages that met certain requirements could be accorded the status of a "Classical Language in India". Languages thus far declared to be Classical are Tamil (2004), Sanskrit (2005), Telugu (2008), Kannada (2008), Malayalam (2013), and Oriya/ Odia (2014). The criteria laid out by the Ministry of Culture to determine the eligibility of languages for classification as a "Classical Language" are as follows:

> High antiquity of its early texts/recorded history over a period of 1500–2000 years; a body of ancient literature/texts, which is considered a valuable heritage by generations of speakers; the literary tradition being original and not borrowed from another speech community; the classical language and literature being distinct from modern, there may also be a discontinuity between the classical language and its later forms or its offshoots.

The Indian Government has been criticized for not including Pali as a classical language as experts have argued it fits all of the above criteria.

Current Localization Landscape

The localization industry in India is not as mature as in the West. Until recently, the government did not pay much attention to it, with most government websites predominantly in English, even though it was against the official language policy. There has been a lack of excitement about localization in the country and there have not been any major deployments. A plausible reason for this might have been insufficient support for Indian languages at the operating system (OS) level. Interestingly, even though most foreign organizations use Unicode fonts for Indian languages, many Indian ones still use non-standard fonts. The popularity of non-standard fonts has hampered interoperability, collaboration, and development of software tools in Indian languages.

Localization in India: SWOT analysis

Strengths

- Presence of IT giants, IT service providers, and the BPO boom has created a demand for language professionals;
- Owing to the growing Indian economy, many foreign companies are expanding to India and many Indian companies are expanding across different Indian cities and abroad. Thus, the demand for translation in Indian languages is on the rise;
- Agencies, institutions, universities, diplomatic missions, corporate houses, government bodies, BPOs, publishing houses, and software companies all already use the services of language professionals.

Weaknesses

- Lack of universally accepted standards for scripts, fonts, input methods, and Romanization;
- Limited availability of text editing and translation software;
- Low availability of local language content;
- Lack of awareness about the language industry;
- Training and certification institutions are in short supply.

Threats

- Popularity of non-standard fonts and character sets has hampered interoperability, collaboration, and development of software tools in Indian languages;
- Popularity and status of English language in India makes it less desirable for companies to translate into local languages;
- Low exposure of Indian translators to translation technology.

Opportunities

- Tenth-largest economy in the world by nominal GDP and third-largest by purchasing power parity;
- Seventh-largest country by area;
- Second-most populous country in the world;
- Fast spread of internet and mobile technologies make content available to the masses;
- Governmental and private sector initiatives.

Two Indian languages (Hindi and Bengali) are amongst top 10 most popular languages spoken in the world. However, there are no Indian languages among the top 10 languages used online.

TOP 10 Languages Used on the Internet

TOP 10 Languages on the Internet	Internet Users by Language	Internet Penetration by Language	Growth in Internet Penetration (2000 – 2011)	Internet Users, % of Total	World Population for This Language (2011 Estimate)
English	565,004,126	43.40%	301.40%	26.80%	1,302,275,670
Chinese	509,965,013	37.20%	1478.70%	24.20%	1,372,226,042
Spanish	164,968,742	39.00%	807.40%	7.80%	423,085,806
Japanese	99,182,000	78.40%	110.70%	4.70%	126,475,664
Portuguese	82,586,600	32.50%	990.10%	3.90%	253,947,594
German	75,422,674	79.50%	174.10%	3.60%	94,842,656
Arabic	65,365,400	18.80%	2501.20%	3.30%	347,002,991
French	59,779,525	17.20%	398.20%	3.00%	347,932,305
Russian	59,700,000	42.80%	1825.80%	3.00%	139,390,205
Korean	39,440,000	55.20%	107.10%	2.00%	71,393,343

Source: Internet World Stats 2010

Top 10 Languages by Number of Native Speakers

	Language	Approximate # of Native Speakers
1	Mandarin Chinese	1,197 million
2	Spanish	414 million
3	English	335 million
4	Hindi	260 million
5	Arabic	237 million
6	Portuguese	203 million
7	Bengali	193 million
8	Russian	167 million
9	Japanese	122 million
10	Javanese	84 million

Source: Ethnologue

On one hand, IT is improving the quality of life in India, but on the other, the use of technology is still out of reach for many. A person literate in Indian languages, but not well versed in English is deprived of access to a vast store of information, thereby creating a "Digital Divide". A study by W3Techs shows that 55.9% of all content online is in English, while Hindi makes up less than 0.1%. In a multilingual country like India, it is essential that tools for information processing in local languages are developed and made available at a low cost for wider proliferation of information and communication technology. It is interesting to note that though the IT boom has brought a revolution to India and Indian computer experts are making waves in Silicon Valley, the Digital Divide continues to plague the nation at home.

Given that not even 10% of Indian population can communicate in English, the impact of localization into various Indian languages cannot be overstated. The situation is no different in most developing and underdeveloped nations. This also means that if the benefits of information and communication technologies are to reach below the top-layers of a society, they need to be adapted to fit into the "world-view" of the people. The communities differ not only in terms of using different languages, but also in a variety of cultural factors, such as the use of colors, meaning of gestures, symbols, analogies, etc. The major task, however, continues to be speaking the user's language. This involves presenting menus, instructions, error messages, and the entire product documentation including online help in all the languages relevant to the communities being addressed.

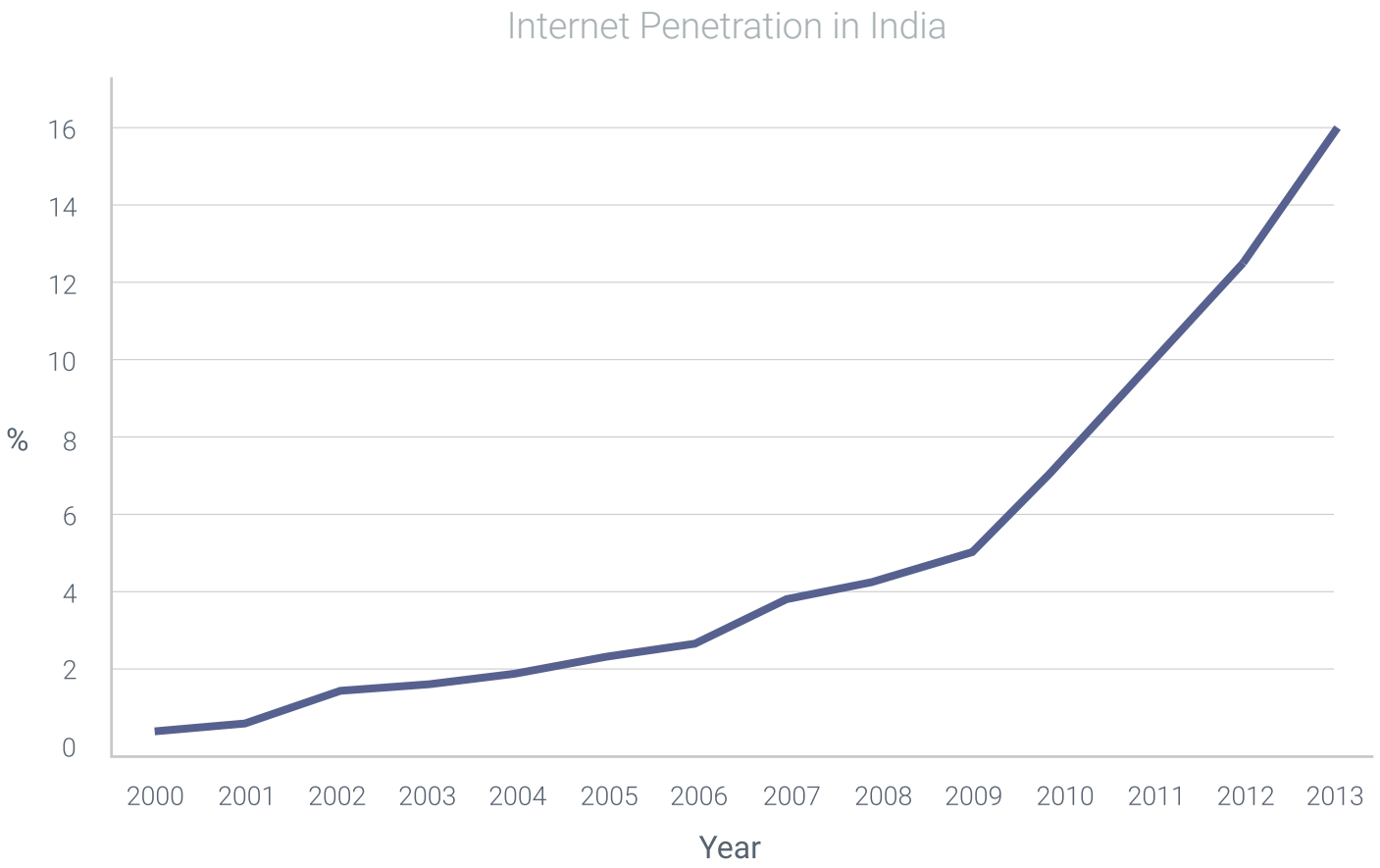
The earlier chaotic scene of encoding schemes has improved with the addition of more and more Indic scripts to Unicode, which is slowly growing in popularity in India. However, there are also issues when it comes to computer-aided translation (CAT) tools. Technically, all languages using scripts included in Unicode are supported by all major CAT tools; however, the level of support varies. Some languages can be used comfortably, while others will suffer from outdated dictionaries, unusable spell-checkers, and poor or non-existent segmentation engines. Additionally, few linguists can afford to pay for a full license, and as a result, CAT tools are rarely utilized. This leads to a vicious cycle where Indian linguists don't use them and CAT tool publishers don't see a market that would justify adding greater support for the languages of India. As a result, most linguists have little experience using the tools even in language pairs and types of projects where they would work well.

The Future

Rising Internet Penetration and Technology Support

Internet penetration in the country may not have crossed 20% of the population yet, but in absolute numbers, this percentage works out to nearly 10 times the population of Australia. According to a report released by the Internet and Mobile Association of India (IMAI) and market research company IMRB, by October of 2013, the nation had over 200 million internet users. The report estimates that in 2014 the number of internet users should overtake the US as the world's second largest internet base after China.

At the same time however, India has the slowest Internet connection speed in Asia, according to Akamai's 2014 State of the Internet report. Among the 14 countries in the Asia-Pacific region in the report, India came in last with an average connection speed of 1.7 Megabits per second. That's less than one tenth of the blazing speed web surfers can expect in South Korea. It's also one third the speed in Australia, Singapore or Thailand and one half the speed of China.



Source: ITU (International Telecommunication Union), Internet in India 2013 - IAMAI report

We are also starting to see a growth in consumption of local language content, driven by the growth of smartphones. As these devices get cheaper, making mobile Internet more accessible in smaller towns and rural parts of the country, the demand for content that is not in English is also rising. Even the cheapest and most basic mobile phones today offer access to the internet. A total of 42% of rural India's internet users prefer surfing in local languages and the high prevalence of content in English continues to be a hurdle for them.

> *Internet users in India could increase by 24% if local language content is provided on the Internet.*
>
> — Local Language Study 2013 by Internet and Mobile Association of India (IAMAI) and IMRB International.

Thanks to growing smartphone penetration, the typical web user is not necessarily sitting in New Delhi or Mumbai anymore and as access reaches a larger audience than ever before, the demand for local language content is rising. Google India believes that the next 300 million Internet users from India won't use English. Popular sites in Hindi include Facebook, YouTube, Wikipedia, Twitter, Google, WordPress, Bing, Blogger, Vube, and *indiatimes*, which shows that the audience wants the same content as everybody else and not just local language websites. There is a huge and untapped audience that wants to read predominantly English language websites in their native languages.

Ramkumar Pichai, from Microsoft India, said, "We have support for 12+ Indian languages now, and continue to invest in localization of the Office user experience. We have support for IME keyboard [through which you can enter non-English characters] built in, and we're going to keep trying to make it easier for everyone to use our software." Pichai adds that the growing number of smartphones has made this both easier and more important than ever, because "There is a definite drive to the interior today, and if you look at this Nokia Lumia phone that I'm using, you go to settings and you can activate Hindi and other languages. It's that simple. And these devices are becoming ubiquitous, so soon everyone will have the ability to use computing in the language they feel comfortable in." This is true not just for Nokia phones, but for other Windows Phone devices as well, you can navigate through the device in your mother tongue, though you'll still need English text to enter URLs to visit websites.

On a BlackBerry 10 device, there is out-of-the box support for Hindi as well. The keyboard even has the same excellent autocomplete features that the English keyboard has. This means that creating content is becoming extremely simple. In the first half of 2014, Micromax launched an Android phone, called the Unite 2, with support for 20 Indian languages including Hindi, Gujarati, Punjabi, and Malayalam. In March 2014, a third-party Indian language keyboard for Android was released. While iOS does not have system support for these languages, it comes with a Hindi keyboard. Google Translate currently supports the following: Bengali, Gujarati, Hindi, Kannada, Marathi, Nepali, Punjabi, Tamil, Telugu, and Urdu and recognizes handwriting in Gujarati, Kannada, Punjabi, Tamil, Telugu, Hindi, and Marathi. Additionally, in 2014, Google Translate added voice recognition for Hindi and seven other Indian languages. However, the different language keyboards of standardization of keypads (like QWERTY layout for English) and users can get confused if each manufacturer pushes a different layout.

This is a big change from the early days, where due to a lack of standards, you would find Hindi content saved as an image file, or as a PDF. Some sites were using proprietary fonts, which also meant that their page could not be easily indexed and computer and mobile phone users had to use third-party applications to enter text in Indian scripts. However, while there is growing activity from the industry giants, so far there are relatively few Indian language-only sites among the top 100 sites visited from India.

Government Initiatives

In a crucial step towards making online public services more inclusive, the Department of Electronics and Information Technology (DeitY) of India has launched a set of tools to develop online content in local languages and the website www.localization.gov.in. The site was introduced as part of the National e-Governance Plan launched in 2006 and currently offers tools like a text-to-speech plugin for Firefox, a JavaScript-based on-screen keyboard, and an on-screen Indian language keyboard for Android. Their localization portal offers basic tools and services for making e-government applications available in the official language of each Indian state. The tools currently support six Indian languages with the rest to be gradually added. In June 2013, DeitY created a repository of fonts for all 22 constitutionally recognized languages through TDIL. These fonts are available as a CD and digital downloads.

Another initiative from The Centre for Development of Advanced Computing (CDAC) is directed at helping ministries and departments with the localization process. The localization portal offers code converter APIs for converting legacy data to Unicode, JavaScript-based on-screen keyboards, Sakal Bharati OpenType font that supports all 22 Indian languages and transliteration services. Additionally, CDAC works on solutions for speech processing, machine translation, optical character recognition (OCR), fonts, tools for Indian languages GUI design, and transliteration standards.

W3C India was set up in 2006 to engage all stakeholders in the country to work towards the internationalization of W3C standards. The long-term goal is to enable all W3C standards in the 22 Indian languages to achieve a seamless web experience for every Indian language speaker. The second objective is to promote the adoption of W3C recommendations among developers, application builders, and standard setters.

Finally, a Government of India initiative known as the National Translation Mission intends to establish translation as an industry and facilitate higher education by providing translated study materials for students. The program also trains translators in different languages. It is not meant to replicate the efforts of other organizations like CDAC or TDIL that develop language support for computers. Their website also offers dictionaries and other resources for translators. One of their projects is a Machine Translation system that instantly translates text from all major Indian languages to English.

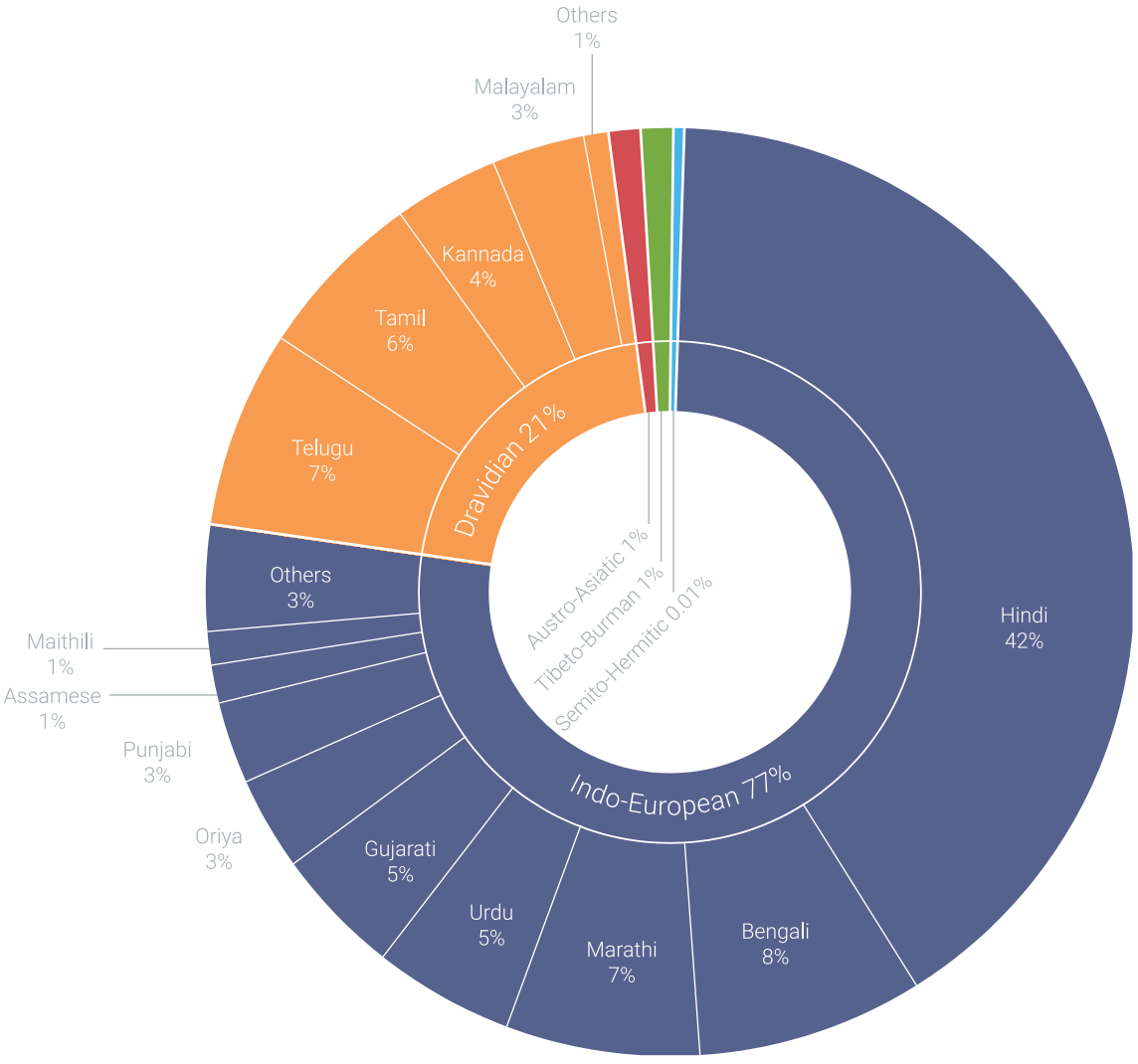
By now, the Indian government has invested 500-600 million Rupee in creating language resources – fonts, dictionaries, thesaurus, and OCR tools and technologies, but most have not been widely adapted. If they are to be accessible to people outside government and academic institutions, they need to be open sourced. Speaking at #NAMA: The Digital Future of Indic Languages event in July 2014, Summit Information Systems MD Rakesh Kapoor said he wants the government to look into five things:

1. Government should have a policy that they will not buy a device that is not localised in all Indian languages.
2. All government tenders should be published in Indic languages.
3. All the e-governance should be done using a 3 language formula: state language, Hindi and English.
4. Government funded technology should be made open source.
5. Any law that is not published in the language that I understand should not be applicable to me. If I don't understand the law, how can I abide by it?

Language Families

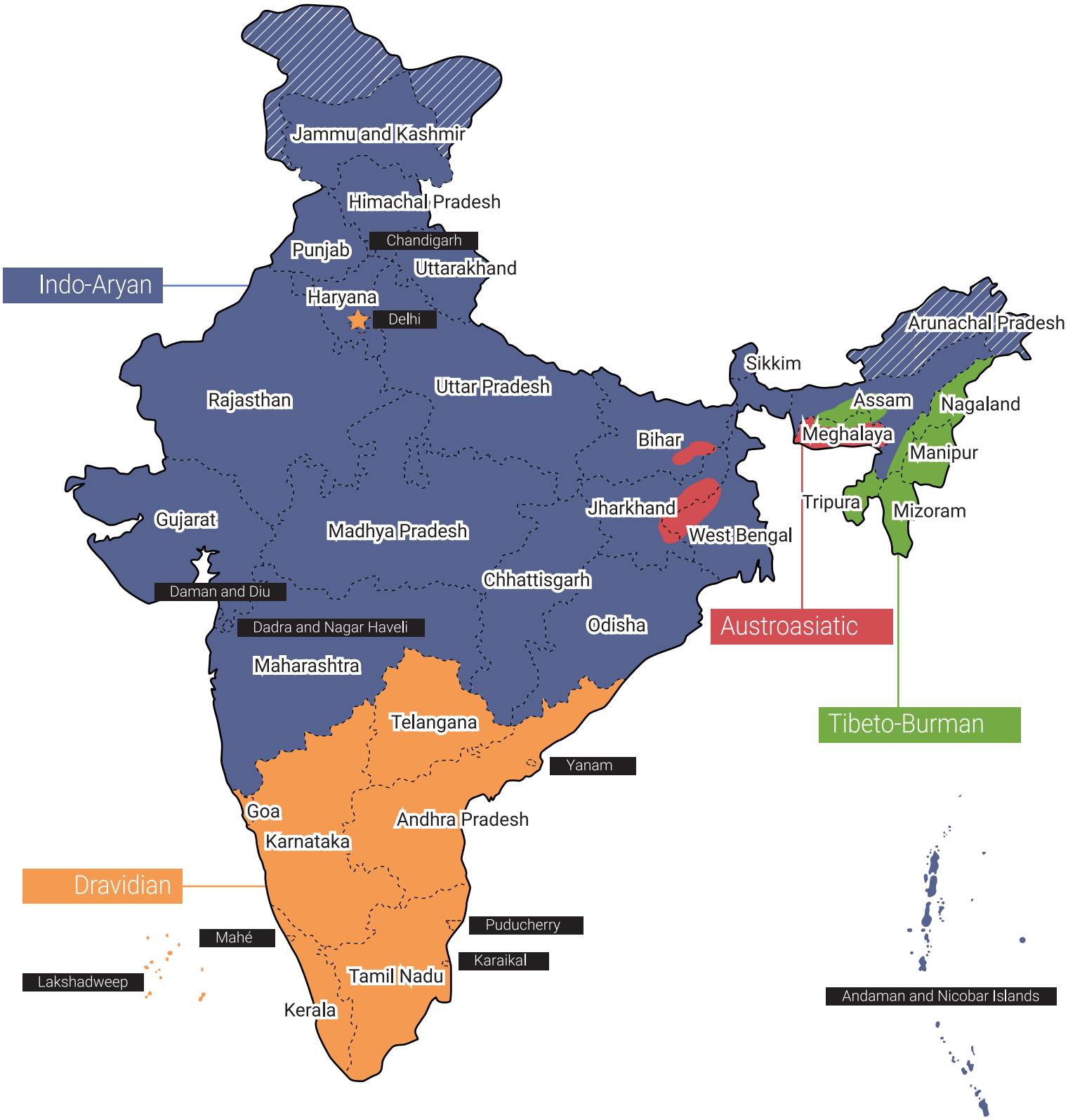
In the 2001 Indian census, 122 languages were grouped into the following families:

- Indo-European: 24 languages, with speakers that make up 76.89% of the total population, divided into Indo-Aryan (21), Iranian (2), and Germanic (1 - English);
- Dravidian: 17 languages, with speakers that make up 20.82% of the total population;
- Austro-Asiatic: 14 languages, with speakers that make up 1.11% of the total population;
- Tibeto-Burman: 66 languages with speakers that make up 1% of the total population;
- Semito-Hermitic: 1 language (Arabic), with speakers that make up 0.01% of the total population.



Source: 2001 Indian Census

Map of Language Families in India



The status of Arunachal Pradesh and parts of Jammu and Kashmir states is disputed by Pakistan, China and India, as marked by diagonal lines on the map. The borders shown on this map are not meant as a statement in this dispute by Andovar.

Andovar

Indo-Aryan

Indo-Aryan languages constitute a branch of the Indo-Iranian languages, itself a branch of the Indo-European family. This is the dominant language family of the Indian subcontinent. The largest in terms of native speakers are Hindustani (Hindi-Urdu, about 240 million), Bengali (about 230 million), Punjabi (about 110 million), Marathi (about 70 million), Gujarati (about 45 million), Bhojpuri (about 40 million), Oriya/ Odia (about 30 million), Sindhi (about 20 million), Sinhala (about 16 million), Nepali (about 14 million), and Assamese (about 13 million) with a total number of native speakers of more than 900 million.

Dravidian

The Dravidian languages are spoken mainly in Southern India and parts of Eastern and Central India, as well as in Northeastern Sri Lanka, Pakistan, Nepal, Bangladesh, and overseas in other countries such as Malaysia and Singapore. The most populous Dravidian languages are Telugu, Tamil, Kannada, and Malayalam. There are also small groups of Dravidian-speaking tribes, who live beyond the mainstream communities. Dravidian languages have been in use since the 2nd century BCE and it is often speculated that they are native to India. Only two Dravidian languages are exclusively spoken outside India: 1. Brahui in Pakistan; and 2. Dhangar (a dialect of Kurukh) in Nepal.

Dravidian place-names along the northwest coast as well as Dravidian grammatical influences on Marathi, Konkani, Gujarati, Marwari, and Sindhi languages suggest that Dravidian languages were once spoken widely across the Indian subcontinent.

Austroasiatic

The Munda languages are a language family spoken by about nine million people in Central and Eastern India and Bangladesh. They constitute a branch of the Austroasiatic language family, which means they are distantly related to Vietnamese and Khmer (Cambodian). The origins of the Munda languages are not known and are thought to predate the other languages of Eastern India. The family is divided into two branches as follows: 1. North Munda: spoken in the Chota Nagpur Plateau of Jharkhand, Chhattisgarh, West Bengal, and Odisha; and 2. South Munda: spoken in Central Odisha and along the border between Odisha and Andhra Pradesh.

Tibeto-Burman

The Tibeto-Burman languages are the non-Sinitic members of the Sino-Tibetan language family, over 400 of which are spoken throughout the highlands of Southeast Asia, as well as lowland areas in Burma (Myanmar). The name derives from the most widely spoken of these languages, Burmese (over 32 million speakers) and the Tibetic languages (over 8 million speakers). Most of the other languages are spoken by much smaller communities and many have not been described in detail.

Scripts of India

Introduction

Apart from Perso-Arabic, all major scripts used for Indian languages have evolved from the ancient Brahmi script (used during Emperor Ashoka's time, roughly 300 BCE) and have a common phonetic structure. Brahmic-derived scripts are used throughout South Asia (excluding Pakistan and Afghanistan), Southeast Asia, and parts of Central and East Asia. They are used by languages of several language families including Indo-European, Dravidian, Tibeto-Burman, Mongolic, Austroasiatic, Austronesian, and Tai.

In most places, it is safe to assume that a given language is written with just one alphabet, but this is not true in India. Indian languages are written in more than 14 scripts. One script can be used to write many languages, and most languages can be written in several scripts. All the languages of India use a similar set of consonants and vowels, but there are significant differences between the writing systems characterized by their respective scripts.

Matra

In most Indian scripts, many characters have a horizontal line at the upper part, which is known as Matra or headline. No Western alphabet has such characteristic. In continuous handwriting from left to right, the Matra of one character joins with the Matra of the previous or next character of the same word. In this fashion, multiple characters in a word appear as a single connected component joined through the common Matra.

The Brahmic script family are all unicameral, i.e., there is no upper and lower case. They are all also abugida writing systems, which means that they are partially alphabetic and partially syllable-based. This is different from a full alphabet (such as the Latin alphabet), where vowels have equal status to consonants. The guiding principle is basically that the displayed shape of a syllable can be generated using certain rules that combine the shapes of individual consonants and the vowel the syllable is made of. When a syllable is pronounced, it is strictly in the order of the consonants terminated by the vowel that always combines with the last consonant in the syllable. Each syllable is made up of one or more consonants and a vowel. A pure vowel is also seen as a syllable.

We can therefore represent a syllable as:

V
CV
CCV
CCCV
CCC----V

While this may imply that arbitrary syllables may be formed, it is impractical to pronounce such combinations. So, in reality, there is a finite set of CC--V combinations that are meaningful in any Indian language. The majority of the syllables in Indian languages are known to consist of just two or three consonants. The total number of combinations in use in each language usually runs into hundreds.

Transliteration

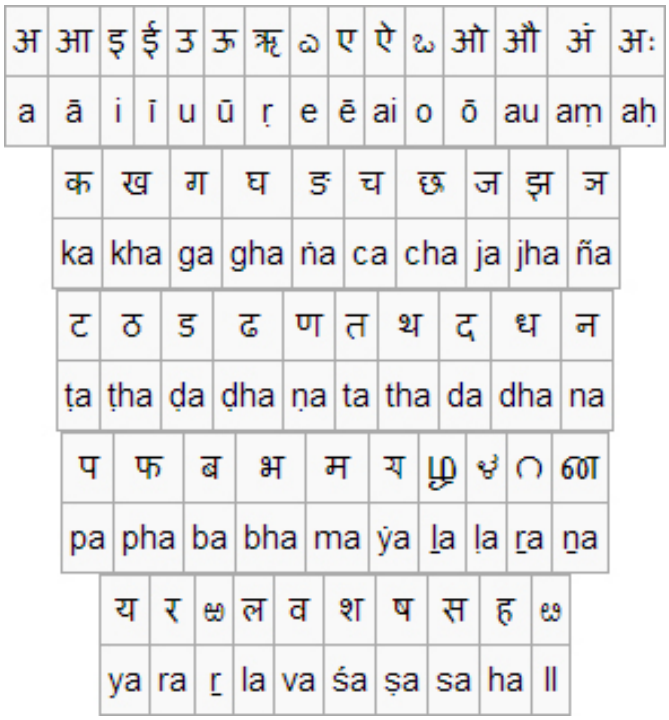
The syllables may be represented in the Latin alphabet if appropriate diacritic marks are used. For several decades now, Romanization has been used to represent texts of Indian languages, especially Sanskrit. In many printed books, a key to Romanization would be printed at the beginning in the form of a table. Since it is difficult to represent all syllables of Sanskrit using just the 26 letters of the Latin alphabet, scholars used varying schemes to accommodate sounds that could not be correctly indicated using a Latin letter. The different schemes have been somewhat arbitrary in the choice of letters. The International Phonetic Alphabet (IPA) also provides symbols for writing Indian languages.

The primary difficulty in data entry of the phonetic symbols is that there is no provision to input the symbols directly using a standard ASCII keyboard. Desktop publishing and word processing programs provide a means by which the glyph code of the symbol is input using the numeric keypad. While this is acceptable, it is not a natural approach. Romanization methods that use only the displayable ASCII symbols do not have this problem because the ASCII letters are directly typed. A special computer program is required to interpret the input string to produce the Indian language displayed or for a hardcopy. This is precisely what current Romanization schemes attempt. These schemes allow multiple representations for certain syllables and long vowels, but the processing program handles them well. The two most commonly used Romanization systems are the International Alphabet of Sanskrit Transliteration (IAST) and The National Library at Kolkata transliteration system.

IAST is a Romanization scheme that allows for lossless transliteration of Indian scripts that come from Sanskrit. It is also used to transliterate Pali and other scripts. IAST is commonly used for books about ancient Sanskrit and Pali in relation to Indian religions. It is based on a standard established by the International Congress of Orientalists in Geneva in 1894.

The National Library at Kolkata transliteration system is the most widely used Romanization scheme for dictionaries and grammars of Indian languages. This scheme is also known as the Library of Congress system and is nearly identical to one of many possible ISO-15919 variants. The table below mostly uses Devanagari, but includes letters from Kannada, Tamil, Malayalam, and Bengali to illustrate the Romanization of non-Devanagari characters. The scheme is an extension of the IAST scheme described above.

National Library at Kolkata Transliteration Scheme



Source: Wikipedia

While Romanization-based data input is useful, the schemes themselves vary for a given language. Consequently, the data entry procedures will change depending on the scheme or a given string will produce different outputs for different languages/ scripts. Transliteration schemes have to face the problem of letters present in one language and not in another. Thus, unless a superset of letters from all the Indian Languages is formed, uniform Romanization is not possible. Even if such a superset were identified, it turns out that unique Latin letter combinations are not easily identified for all complex syllables. Moreover, the large numbers of vowels in Indian scripts' add to the difficulty.

Encoding

Indian scripts form the most diversified and complicated script group in the world and are among the most difficult to handle with computers. While it is true that all Indian languages have a phonetic base built on top of a fixed number of vowels and consonants, the writing systems permit a great many different shapes to be generated depending on the syllables in the text.

Indian Script Code for Information Interchange (ISCII) was proposed in the 1980s and a suitable standard had evolved by 1991. More recently, multilingual text representations have been made possible through Unicode, the scheme that supports representation for a variety of scripts worldwide. Although Unicode allows for very large number of codes to represent all symbols in a language (11,000 are used for Korean Hangul and almost 20,000 for Chinese and Japanese scripts), both ISCII and Unicode have taken an approach of representing only the consonants and vowels of the Indic scripts, rather than all possible combinations. This means that displaying syllables is left to the processing software, which results in challenges with display, widths and heights of fonts, and copying-and-pasting between applications.

Unicode Standard is a 16-bit storage encoding standard that is used internationally by the industry for the development of multilingual software. TDIL is a voting member of the Unicode Consortium to ensure an adequate representation of Indian scripts in the Unicode Standards. While many scripts are now supported by Unicode, there is still work to do for the less-commonly used scripts and to solve the aforementioned issues.

List of South Asian Scripts Represented in Unicode 7.0

Assamese	Kannada	Modi	Syloti Nagri
Bengali	Kharoshthi	Mro	Takri
Brahmi	Khojki	Ol Chiki	Tamil
Chakma	Khudawadi	Oriya	Telugu
Devanagari	Lepcha	Saurashtra	Thaana
Devanagari Extended	Limbu	Sharada	Tirhuta
Grantha	Mahajani	Siddham	Vedic Extensions
Gujarati	Malayalam	Sinhala	Warang Citi
Gurmukhi	Meetei Mayek	Sinhala Archaic Numbers	
Kaithi	Meetei Mayek Extensions	Sora Sompeng	

Source: Unicode Consortium

Bengali Script

The Bengali script is also known as Bangla. With a few small modifications, it is also used for writing Assamese and in Assam, the preferred name of the script is Asamiya or Assamese. Other related languages in the region also make use of the Bangla alphabet. One example is Meitei, which has been written in the Bangla script for centuries, though Meitei Mayek script has been promoted in recent times.

It is a Brahmic script although its exact derivation is disputed. Bengali writing shares similarities with the Dravidian-language scripts, particularly in the shapes of some vowel letters, but it has greater similarity with the Indo-Aryan scripts, in particular Devanagari.

Bengali is an abugida and is written from left to right. There are thirty-five consonant letters and eleven independent vowel letters, two of which represent diphthongs. Each vowel letter also has a diacritic form that combines with a consonant to modify the inherent vowel. These can be written to the left or the right of, or above or below the consonant. Some are digraphs, written with part of the letter before and part after the consonant.

Devanagari Script

The official language of India, Hindi is written in the Devanagari script. Devanagari is also used for writing over 120 other Indo-Aryan languages, including Nepali, Marathi, Maithili, Awadhi, Newari, and Bhojpuri. It can be used for writing Classical Sanskrit texts, and is the official script of Nepal. The word "Devanagari" is a compound word consisting of: deva meaning "deity", and nagari meaning "city". Together it implies a script that is both religious and urban. Devanagari is the most commonly used script in India.

It is related to many other South Asian scripts including Gujarati, Bengali, and Gurmukhi; and more distantly to a number of Southeast Asian scripts including Thai, Balinese, and Baybayin. It has 33 consonants and 12 vowels. They are called basic characters. Vowels can be written as independent letters or by using a variety of diacritical marks that are written above, below, before, or after the consonant they belong to. When vowels are written in this way, they are known as modifiers. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters. These types of basic characters, compound characters, and modifiers are present in Devanagari and in other scripts, as well.

Devanagari is an abugida and each letter represents a consonant with an inherent vowel that can be modified using vowel diacritics. They can be written above, below, to the left or to the right of the consonant. The vowel signs represent long and short forms for five vowel sounds. Vowel sounds that are not preceded by a consonant are written with a vowel letter or indicated by a vowel diacritic. The script is written from left to right. Letters hang from a headstroke (matra), which is generally continuous throughout the length of the word, except for a few letters. In handwriting, the headstroke is sometimes omitted.

Gujarati Script

The Gujarati script is used to write the Gujarati and Kutchi languages. It is a variant of Devanagari script differentiated by the loss of the characteristic horizontal line running above the letters and by a small number of modifications in the remaining characters. The loss of the headstroke reflects the script's origins in informal writing. Until the mid-19th century, it was used primarily for bookkeeping and personal correspondence. Since printing facilities have become widely available to Gujarati speakers, the script is used in schools, for printing books and newspapers, in government offices, and public signage.

With a few additional characters, added for this purpose, the Gujarati script is often used to write Sanskrit and Hindi. It is also used alongside the Devanagari script for writing a number of languages used by the Bhil people, one of India's largest indigenous groups.

Gujarati is written from left to right. It is an abugida, i.e., every consonant letter contains an inherent vowel. There are 45 basic symbols, 34 consonants, and 11 vowels. Despite the loss of the headstroke, letters are still aligned as from a hanging baseline.

Gurmukhi Script

Gurmukhi script is used in Northwestern India, mainly in Punjab region. It is used primarily by followers of the Sikh religion to write the Punjabi language. It has little literature and shows little borrowing from Persian, Arabic, or Sanskrit. Gurmukhi was devised in the 16th century by the Sikh Guru Angad (1539-52), successor to the founder of the Sikh religion, Guru Nanak, so that sacred literature can be accurately recorded. The word Gurmukhi means “from the mouth of the guru”. Muslims in the Pakistani Punjab write Punjabi in a Persian script called Shahmukhi, which means “from the mouth of the king.”

Gurmukhi script alphabet consists of 41 consonants and 12 vowels. In addition to these, so-called “half characters” are present in the feet of characters. Gurmukhi is written from left to right using letters that hang from a horizontal headstroke. The script is an abugida, i.e., each letter represents a consonant with an inherent vowel.

Kannada Script

The Kannada alphabet is an abugida of the Brahmic family, used primarily to write the Kannada language, one of the Dravidian languages of Southern India. Kannada script is derived from the Old Kannada script, which evolved around the 10th century. Kannada is second only to the Devanagari in the number of languages using it on the Indian subcontinent. Apart from Kannada, several minor languages, such as Tulu, Konkani, Kodava, and Beary also use Kannada script. Kannada and Telugu scripts share high mutual intelligibility. They are often considered regional variants of a single script, because both languages were written using the Old Kanarese script in the past until it diverged into two distinct varieties. Similarly, Goykanadi, a variant of Old Kannada has been historically used to write Konkani in the state of Goa. The script was standardized in the 19th century under the influence of Christian mission organizations, who established printing presses for printing books in the Kannada language.

Kannada letters are rounded in shape, which is typical of South Indian scripts. They are generally written with a headstroke running along the top, although it is not continuous throughout words as in Devanagari, and is omitted from some letters. There are 49 basic letters in the Kannada inventory. Of these, 34 are consonants, 13 are vowels, and two are non-alphabetic symbols.

Malayalam Script

The Malayalam script is a Brahmic script used commonly to write the Malayalam language, the principal language of the Indian state of Kerala. Like many other Indic scripts, it is an abugida. The script is also used to write several minority languages such as Paniya, Betta Kurumba, and Ravula. The Malayalam language itself was historically written in several different scripts and it was developed in the 16th century from the Vatteluttu script.

Malayalam is written from left to right. There are 53 letters, called aksaras; 37 of these represent full syllables consisting of a consonant and the vowel [a], and 16 which represent independent vowels. Independent vowel letters are only used where a vowel appears at the beginning of a word. Vowels that follow a consonant are written with a diacritic above, below, to the left or right of, or flanking either side of the consonant letter. The script also contains a subset of letters called “letter fragments” to represent sonorants followed by a pause, for example, at the end of a word.

Meetei Mayek Script

The name "Meetei Mayek" is used on official documentation in Manipur. Many linguists also use the name "Meitei Mayek", and other Romanization variants occur. It is an abugida that was used for the Meetei language (Manipuri), one of the official languages of the Indian state of Manipur, until the 18th century, when it was replaced by the Bengali script. The origins of the script are controversial, because most of the early documents were destroyed. Some sources claim it has been used for almost 4,000 years, while others suggest it was derived from the Bengali script as recently as the 17th century. In the last 100 years, the script has experienced a resurgence.

Meetei Mayek is an abugida written from left to right with spaces between words. There are two related forms of the script, the form used before the 18th century, and the modern form. Unmarked letters contain an inherent [ə] vowel, and other vowels are represented by writing one of seven (in the modern orthography - older texts use twelve) diacritics above, below, to the left, or to the right of the base letter. Since the Meetei language does not have voiced consonants, there are only 15 consonant letters used for native words, plus 3 letters for pure vowels. Nine additional consonant letters inherited from other Indian languages are available for borrowings. There are seven vowel diacritics and one final consonant diacritic.

Oriya/ Odia Script

The Oriya script, also called Utkala Lipi or Utkalakshara is used to write the Oriya/ Odia language, as well as a number of Dravidian and Munda minority languages spoken in the region. It is also used in Odisha state for transcribing Sanskrit texts. The earliest inscriptions in the Odia language have been dated to 1051 CE, written in the Kalinga script from which modern Odia writing is derived. As of 2012, the name "Oriya" for this script and language is officially spelled "Odia" in India. There are noticeable similarities between Oriya and Thai scripts, most likely due to traders who traveled between the two regions. The curved appearance of the Oriya script is a result of the practice of writing on palm leaves, which have a tendency to tear when straight lines are written on them.

Oriya is an abugida with all consonants having an inherent embedded vowel. Diacritics, which can appear above, below, before, or after the consonant they belong to, are used to change the form of the inherent vowel. When vowels appear at the beginning of a syllable, they are written as independent letters. Also, when certain consonants occur together, special conjunct symbols are used to combine the essential parts of each consonant symbol. Like the other Brahmic scripts used in India, the Odia script is written from left to right and is based on the orthographic syllable called aksara. An aksara represents either a lone vowel or a consonant with a vowel attached. Each vowel sound can be written with one of letters. When used at the beginning of a word it is written with an independent vowel letter or when it follows a consonant it is written with a dependent vowel diacritic that attaches above, below, beside, or flanking both sides of the consonant letter.

Perso-Arabic Script

The Persian or Perso-Arabic alphabet is a writing system based on the Arabic script. Originally exclusively used for the Arabic language, the Arabic alphabet was adapted to the Persian language by adding four letters. Many languages that use the Perso-Arabic script add other letters. In addition to the Persian alphabet itself, the Perso-Arabic script has been applied to the Urdu alphabet, Sindhi, Saraiki, Kurdish Sorani, Lurish (Luri), Ottoman Turkish, Balochi, Shahmukhi, Tatar, Azeri, and several other alphabets. In order to represent non-Arabic sounds, new letters were created by adding dots, lines, and other shapes to existing letters.

The Perso-Arabic script is an abjad, meaning that each symbol stands for a consonant and the readers have to add the appropriate vowel themselves. In texts for beginners, three floating symbols are used in the place of vowels to help with pronunciation. These are, in addition to six other symbols, nine secondary symbols of the Perso-Arabic alphabet. Without the extra symbols, correct pronunciation of the words is difficult and only possible with prior knowledge. It is also exclusively written in cursive, and the majority of letters in a word connect to each other. This is also implemented on computers. Whenever the Perso-Arabic script is typed, the computer connects the letters to each other. Unconnected letters are not widely accepted. In Perso-Arabic, as in Arabic, words are written from right to left, while numbers and foreign words (for example in Latin alphabet) are written from left to right.

Sinhala Script

The Sinhalese alphabet is an abugida used by the Sinhala people in India, Sri Lanka, and elsewhere to write Sinhala language, as well as Pali and Sanskrit. As a member of the Brahmic family of scripts, the Sinhalese script traces its origins back more than 2000 years.

Sinhalese is often considered two alphabets, or an alphabet within an alphabet, due to the presence of two sets of letters. The core set, known as the “śuddha simhala” (pure Sinhalese) or “elu hōdiya” (Elu alphabet), contains 20 consonant and 20 vowel letters and can represent all native phonemes. However, in order to render Sanskrit, Pali, Hindi, and English loanwords, an extended set, the “miśra simhala” (mixed Sinhalese), which contains additional 18 consonant letters is used. Sinhala is a diglossic language, i.e., the spoken and written forms show considerable variation. Additionally, spelling conventions do not always reflect current pronunciation.

Tamil Script

The Tamil script, also called “tamiz ezhuthu”, is an abugida used by the Tamil people in India, Sri Lanka, Malaysia, and elsewhere to write the Tamil language. It can also be used to write the liturgical language Sanskrit, using consonants and diacritics not represented in the Tamil alphabet. Certain minority languages are also written in the Tamil script. The script is thought to have evolved from the Brahmi script. The use of palm leaves as the primary medium for writing meant that the scribe had to be careful not to pierce the leaves with the stylus. That was because a leaf with a hole was more likely to tear and decay faster. This resulted in more rounded shapes of letters. The forms of some of the letters were simplified in the 19th and 20th centuries in a series of reforms. They standardized the vowel markers used with consonants by eliminating special markers and most irregular forms.

Tamil is written from left to right using an abugida containing 18 consonants (called "body letters"), and 12 vowels ("soul letters"). The consonant inventory is much smaller than for many other Brahmic scripts. This is because voiced and unvoiced counterparts of a sound are represented using a single letter and the pronunciation is dictated by context. Traditional Tamil grammars contain detailed rules, observed in formal speech, for when a stop is to be pronounced with and without voice. These rules are not always followed in daily speech.

Telugu Script

Telugu script, an abugida from the Brahmic family of scripts, is used to write the Telugu language. The Telugu script is also used for writing a number of minority languages in Southern India. The so-called Bhattiprolu Brahmi script evolved into the Telugu script by the 5th century. The script is closely related to the Kannada script, and a person familiar with one can normally read the other. The two scripts developed from a common Brahmic source but diverged around the 13th century. From that time until the early 20th century, Telugu was a literary language reflecting an archaic spoken form. Modern standard Telugu only began to be written during the second half of the 20th century.

Like most Brahmic-derived scripts, Telugu is an abugida written from left to right. Visually, it differs from many of the North Indian scripts in that the letters have a rounded base and the characteristic North Indic headstroke has been replaced by a hook on the top left of each letter. An [a] vowel is inherent in each of the 32 consonant symbols. Vowels other than [a] are written using diacritics attached above, below, or to the right of the consonant symbol. These vowel diacritics override the inherent vowel so that the syllable is read with the correct vowel sound. Where a vowel occurs at the start of a word, there is no preceding consonant symbol by which a diacritic may be attached, so that one of the 16 independent vowel letters is used.

Languages of India

It is beyond the scope of this white paper to attempt to describe all or even most of the hundreds of languages spoken and written in India. Therefore, we have used the following criteria:

1. Is the language included in the Eights Schedule in the Constitution?
2. Is it an official language for any state or union territory?
3. Does it have a Classical Language of India or a secondary official language status in any state or union territory?

Using these characteristics, we have shortlisted 29 languages to describe in more detail and to look at their localization potential. On the following pages, the languages will be introduced one-by-one in the following format, with the exception of English, which will receive a more holistic overview:

Introduction: A few words about the language and any important information;

Language family: For simplicity's sake, we have used the following families: Indo-Aryan, Dravidian, Austroasiatic, and Tibeto-Burman;

Status: What is the official status of the language in India, if any;

Script: The script(s) that the language is written in;

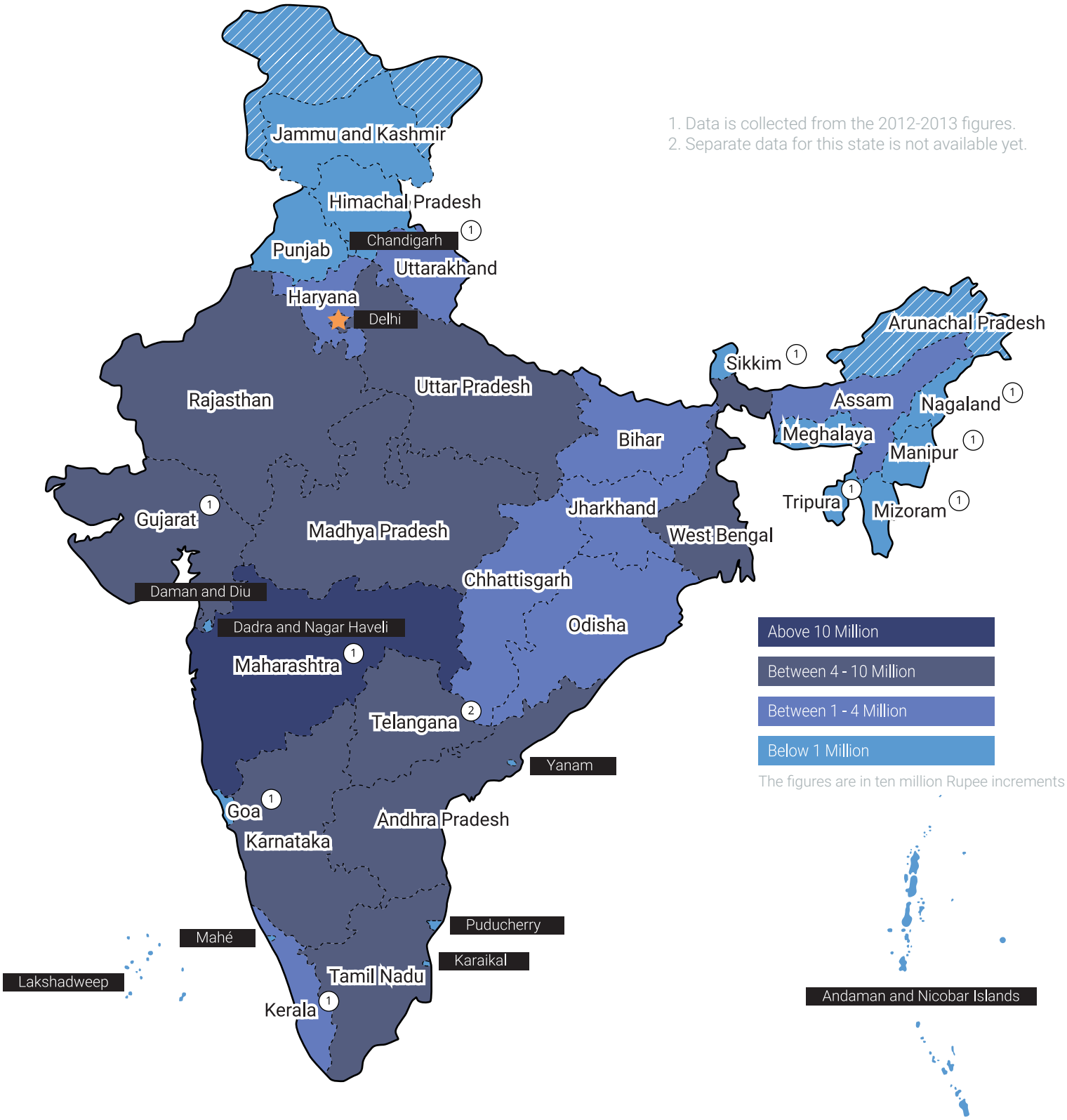
Distribution: Where in India and worldwide the language is spoken and how many speakers there are in each location;

Per-capita income of native speakers: The Indian government does not make this information public and there is no direct data available, but an attempt to estimate the relative income of the native speakers of each language who reside in India has been made by taking the average income of states into consideration;

Localization grade: This is Andovar's ranking of the languages in terms of their importance for localization, which takes into account such factors as the number of speakers, their distribution, and average incomes;

Dialects: What are the major dialects of the language and which one is considered standard?

Map of GDP Per-Capita by State



Source: Ministry of Statistics and Programme Implementation of India. 2013/14 Data

The status of Arunachal Pradesh and parts of Jammu and Kashmir states is disputed by Pakistan, China and India, as marked by diagonal lines on the map. The borders shown on this map are not meant as a statement in this dispute by Andovar.

Andovar

English

India is world's second largest English-speaking country. The Constitution of India designates a bilingual approach to official language of the Government of India employing usage of Hindi and English. English finds everyday use for important official purposes such as parliamentary proceedings, the judiciary, and communications between the Central Government and a State Government. English is also the language of business, IT and education.

It is also the "common" language among most educated Indians today. When two Indians from different states meet, they invariably communicate with each other in English. English speakers in India are estimated to be around 9% of the total population. It can be estimated that India has over 350 million English users and about 100 million English speakers.

Public instruction of English began in India in the 1830s during the rule of the East India Company. Originally, UK English was seen to be more popular, but there is a sizeable US influence from international corporations and BPO giants that have set up in India. There is also “Hinglish”, a version of English that is neither UK nor US. Hinglish (the name is a combination of the words "Hindi" and "English") is a macaronic language, a hybrid of English and South Asian languages (not only Hindi), whereby words are freely interchanged within a sentence or between sentences. While the name is based on the Hindi language, it does not refer exclusively to Hindi, but is used everywhere in the country, with English words blending with Indian languages used locally.

While it used to be seen as the language of the street and the uneducated, Hinglish has now become the lingua franca of India’s young urban middle class. One high-profile example is Pepsi’s slogan “Yeh Dil Maange More!” (The heart wants more!), a Hinglish version of its international “Ask for more!” campaign. Hinglish is commonly used in movies, advertisements, media, and almost everywhere else. It is common not only in India, but also widely spoken by Indians abroad (NRIs) in various countries. There are some good reasons for the explosion of English words. They can be badges of honor in a society intent on becoming modern. Even if you do not speak English fluently, you might be able to use the odd word to impress your neighbor.

In some states, a pure form of Hindi is used, which has more of Sanskrit vocabulary. Nevertheless, in states like Delhi, Hinglish is more prevalent. There is a debate between using a pure form of language and using Hinglish. In Government departments, pure Hindi was widely used and appreciated. However, recently, English alternatives in Devanagari script have started to appear.

Hindi

हिन्दी
(Devanagari Script)

Introduction: Hindi is the second most commonly spoken language in the world. Only Mandarin Chinese has a greater number of speakers. It is mutually intelligible with Urdu, and indeed both languages used to be treated as one: Hindustani. Hindi is a direct descendant of Sanskrit through Prakrit and Apabhramsha languages. Over the years, it has been influenced and enriched by Dravidian, Turkish, Farsi, Arabic, Portuguese and English.

In Hindi, unlike in English, all nouns have one of two genders, either masculine or feminine, and adjectives and verbs change according to gender. Learning the gender aspect of Hindi grammar is usually one of the most difficult steps in learning Hindi for Westerners. There are many words in English which are either Hindi or of Hindi origin. For example: guru, jungle, karma, yoga, bungalow, cheetah, looting, thug and avatar.

Hindi got its name from the Persian word "Hind", meaning "land of the Indus River". Persian-speaking Turks who invaded Punjab and Gangetic plains in the early 11th century named the language of the region Hindi, "language of the land of the Indus River".

Status: Included in the Eighth Schedule to the Constitution, official language of the states of Bihar, Chhattisgarh, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand, and Uttar Pradesh and in the union territory of Chandigarh and the capital Delhi. It is also an official secondary language in several other states.

Script: Devanagari script

Distribution: More than 180 million people in India regard Hindi as their mother tongue. Another 300 million use it as a second language. If one includes Bhojpuri and Chhattisgarhi, over 40% of the population of India can be considered Hindi speakers.

Outside of India, there are hundreds of thousands Hindi speakers in the United States; South Africa; Mauritius; Yemen; Uganda; and tens of thousands in Western Europe, Singapore, Australia, and New Zealand.

260M
native speakers worldwide

Aあ
Indo-Aryan
language family

Devanagari
script

Average
per-capita income
of native speakers

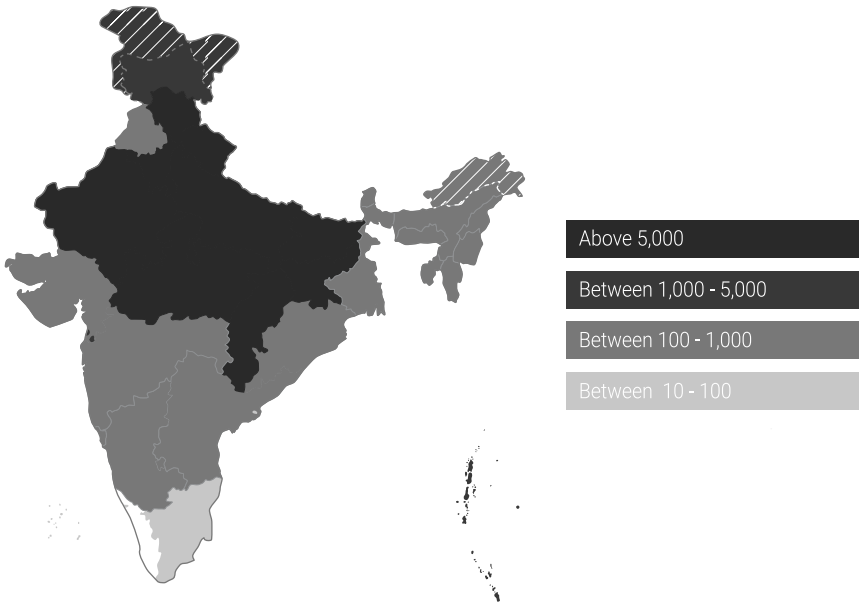
High
localization grade

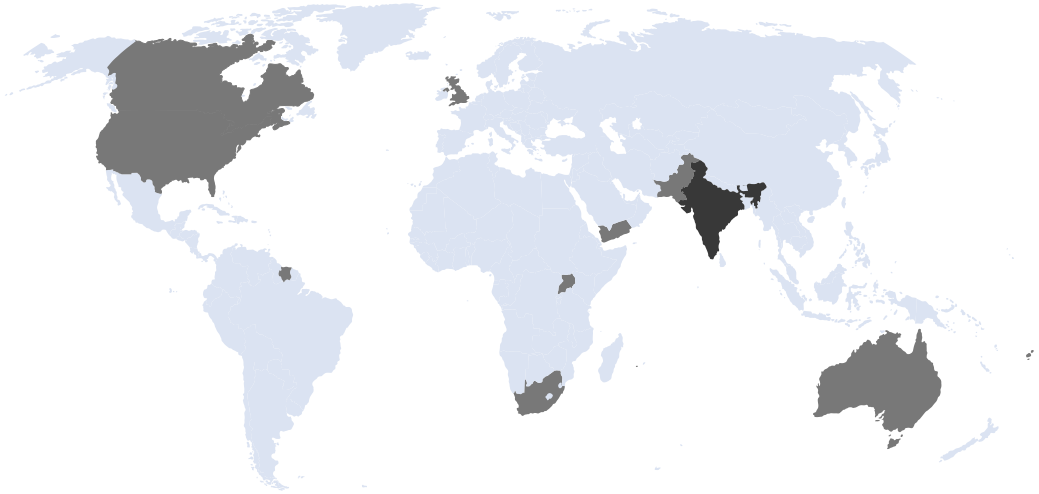
Andovar

Dialects: What is a dialect and what is a separate language is a debatable issue and seen very starkly in the case of Hindi. First or all, there is the Hindustani/ Hindi /Urdu debate already explained earlier. Secondly, there are many languages that are considered dialects of Hindi by some, but not others. For example, the 2001 official government census counted 422 million Hindi speakers (41% of the population) by including people who identified their language as Awadhi, Bagheli, Bhojpuri (Bihari), Bundeli, Chhattisgarhi, Garhwali, Harauti, Haryanvi, Khortha (Khotta), Kumauni, Lamani (Lambadi), Magadhi (Bihari), Malvi, Marwari, Mewari, Nimadi, Pahari, Rajasthani, and Sadan (Sadri), as well as numerous other languages with fewer than 2 million self-identified speakers.

Much of the Hindi spoken outside of the subcontinent is quite distinct from the India-Pakistan standard language as well. This includes varieties spoken in Mauritius, Suriname, Fiji, Trinidad and South Africa.

Finally, from the point of view of linguistic classification, Hindi can be split into two major groups: Western Hindi, which evolved from the Apabhramsa form of Sauraseni Prakrit and Eastern Hindi evolved from Ardhamagadhi.





Assamese

অসমীয়া
(Assamese Script)

Introduction: Assamese, also known as Asambe, Asamiya, or Ôxômiya is the easternmost Indo-Aryan language. Assamese is closely related to Bengali. The word "Assamese" is an English formation built on the same principle as Japanese or Vietnamese. It is based on the English word "Assam", name of a geographical area consisting of the Brahmaputra Valley and its adjoining areas. The people call their state Ôxôm; hence, the name Ôxômiya for the language.

13M
native speakers worldwide

The present standard Assamese script is identical to the Bengali script except for three letters.

Aあ
Indo-Aryan
language family

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Assam. Secondary official language in Arunachal Pradesh state.

Script: Modern Assamese uses the Bengali script with a few small modifications, and, historically, Kamrupi script.

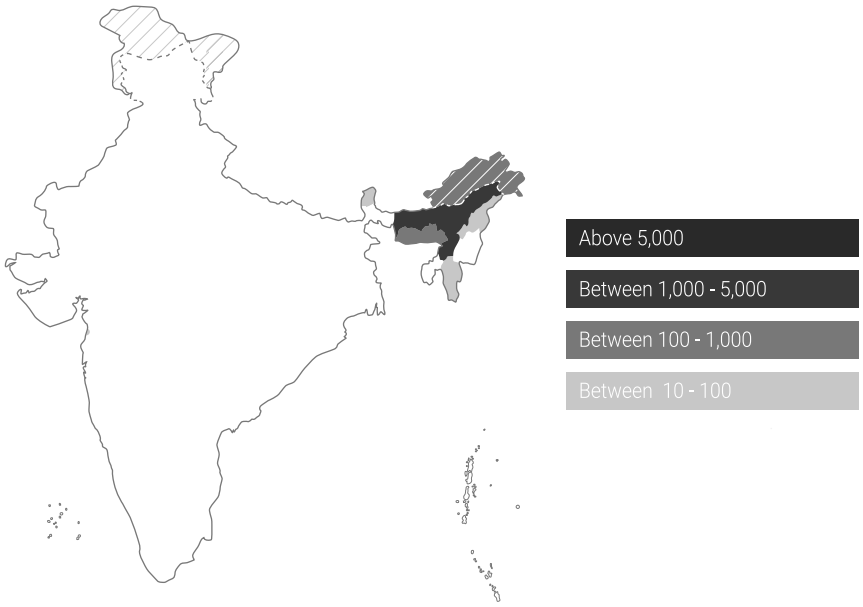
Assamese
script

Distribution: 13 million native speakers in India, mainly in the state of Assam. It is also spoken in parts of Arunachal Pradesh and other northeast Indian states. Small pockets of Assamese speakers can be found in Bangladesh.

High
per-capita income
of native speakers

Dialects: Assamese has a number of regional dialects, with an Eastern dialect centered in and around Sivasagar District that is considered standard. Nagamese spoken in Nagaland and Nefamese spoken in Arunachal Pradesh are closely related to Assamese.

Low
localization grade



Bengali

বাংলা
(Bengali Script)

Introduction: Bengali, also known as Bangla is native to the region of Bengal that comprises present-day Bangladesh and the Indian states of West Bengal, Tripura, and Southern Assam. The National Anthems of Bangladesh, of India, and of Sri Lanka as well as the national song of India were first composed in the Bengali language.

193M
native speakers worldwide

UNESCO's International Mother Language Day on the 21st February was established in 1999 to promote linguistic and cultural diversity and multilingualism. The date represents the day in 1952 when students demonstrating for recognition of Bengali as one of the two national languages of Pakistan (when Bangladesh was still part of Pakistan) were shot and killed by police.

Aあ
Indo-Aryan
language family

Status: Included in the Eighth Schedule to the Constitution, official language of the state of West Bengal. Official secondary language in Tripura, Assam, Jharkhand states, and the union territory of the Andaman and Nicobar Islands.

Bengali
script

Script: Bengali script. There are various ways of Romanization systems of Bengali created in recent years, but none accurately represents true Bengali phonetic sounds.

High
per-capita income
of native speakers

Distribution: 82 million native speakers in India, mainly in West Bengal and neighboring states, as well as in the Indian union territory of Andaman and Nicobar Islands.

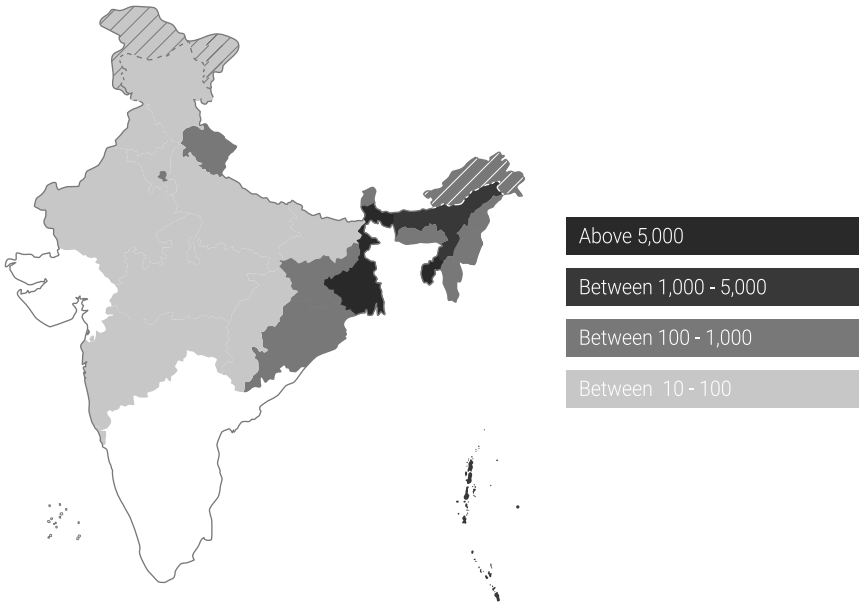
It is also the national and official language of Bangladesh. Additionally, there are significant Bengali-speaking communities in the Middle East, Japan, United States, Pakistan, Singapore, Malaysia, Maldives, Australia, Canada, and United Kingdom. It is estimated that worldwide there are about 220 million speakers.

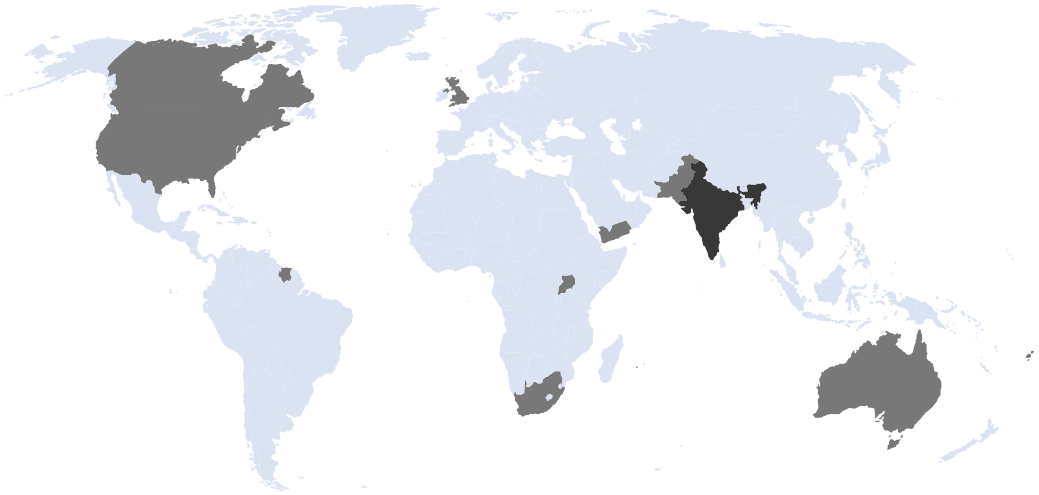
High
localization grade

Andovar

Dialects: Bengali presents a strong case of diglossia, with the literary and standard form differing greatly from the colloquial speech of the regions that identify with the language. There are two standard styles in Bengali: the Sadhubhasa (archaic dialect) and the Chaltibhasa (modern or colloquial dialect). The former was largely shaped by the language of early Bengali poetical works.

What is accepted as the standard form today in both West Bengal and Bangladesh is based on the West-Central dialect of Nadia District, located next to the border of Bangladesh. However, standard Bengali in West Bengal and Bangladesh are marked by some differences in usage, accent, and phonetics and different voice-over artists should be used for these markets.





Bhojpuri

भोजपुरी
(Devanagari Script)

Introduction: Bhojpuri or Bojpuri is a language from the Hindi belt and for cultural reasons; it is usually seen as a dialect of Hindi. Many Bhojpuri-language magazines and papers are published in Bihar and Uttar Pradesh. The term "Bihari" (a pejorative alternate name for Bhojpuri) is also used to other languages such as Maithili and Magahi.

40M
native speakers worldwide

The Bhojpuri-speaking region has a rich tradition of creating leaders for building post-independence India such as the first President, Dr. Rajendra Prasad.

Aあ
Indo-Aryan
language family

Status: Unofficial secondary language of the state of Bihar. Proposed for inclusion as official language in the Eighth Schedule to the Constitution in 2012.

Script: Historically written in Kaithi scripts, but since 1894, the primary script has been Devanagari script.

Devanagari
script

Distribution: 33 million native speakers in India. Bhojpuri is also a major language spoken in Nepal with official status and is one of the national languages of Guyana, Fiji, and Suriname. A few hundred thousand speakers Bhojpuri are also found in Pakistan and Mauritius. Total number of speakers is estimated at 40 million worldwide.

Average
per-capita income
of native speakers

Dialects: Known dialects are as follows: Bhojpuri Tharu, Domra, Madhesi, Musahari, Northern Standard Bhojpuri (Basti, Gorakhpuri, and Sarawaria), Southern Standard Bhojpuri (Kharwari), and Western Standard Bhojpuri (Benarsi, Purbi). The variant of Bhojpuri of the Indo-Surinamese is also referred to as Sarnami Hindustani, Sarnami Hindi, or just Sarnami and is considerably influenced by Creole and Dutch lexically.

Low
localization grade

Bodo

बोड़ो
(Devanagari Script)

Introduction: Bodo, Boro, or Mech is closely related to the Dimasa language of Assam, the Garo language of Meghalaya, and Kok Borok language spoken in Tripura. The Boros are sometimes known as "Mech".

Although the Bodo language is a rich and ancient language, it did not have written literature until the second decade of the 20th century. Christian missionaries, who entered Bodo speaking areas with an intent to preach their religion, published books on religion, tales, rhymes, and songs using Latin alphabet.

Bodo women are expert in rearing the Eri and Muga worms and weave different kinds of clothes from their threads. This is known as Assam silk.

Status: Included in the Eighth Schedule to the Constitution as official language. Official secondary language of the state of Assam.

Script: Earlier it was written using Latin and Bengali scripts, but since 1963, it has used Devanagari script.

Distribution: Over 1 million in India, mostly in Assam and neighboring states. Some in Nepal.

Dialects: The dialects of Bodo can be broadly sub-divided into Western, Eastern, and Southern groups. Western Bodo dialect has gained the status of a standard dialect and has developed a written form.

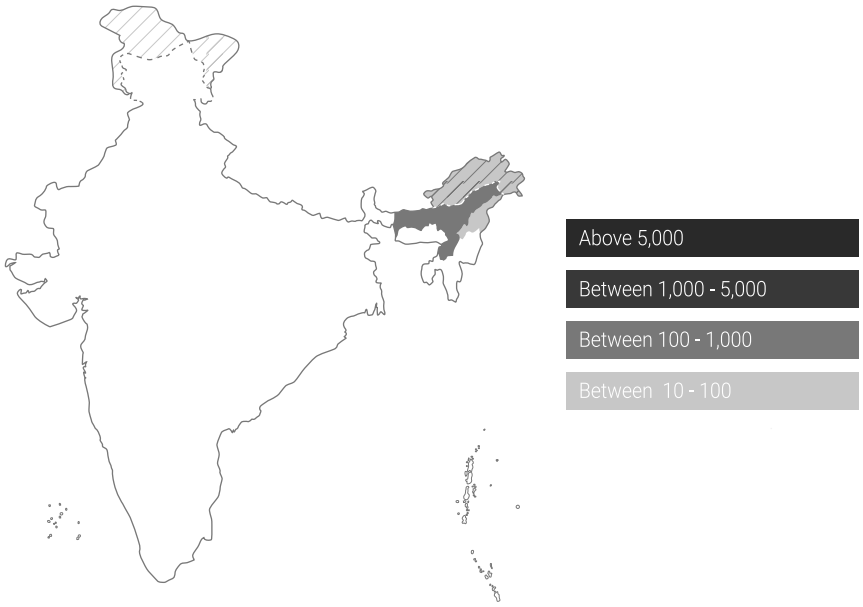
1M
native speakers worldwide

Aあ
Tibeto-Burman
language family

Devanagari
script

Average
per-capita income
of native speakers

Low
localization grade



Dogri

डोगरी (Devanagari Script)

ڈوگری (Perso-Arabic Script)

Introduction: Dogri or Dongri speakers are called Dogras or Duggars. At various times, Dogri has been claimed to be a dialect of Punjabi. The language is referred to as Pahari in Pakistan. Unusually for an Indo-European language, it is tonal. This makes it difficult for speakers of other languages in the region to learn Dogri.

2M
native speakers worldwide

Oldest reference to Dogri comes from the Greek astrologer Pulomi who accompanied Alexander in his 323 B.C. military campaign to the Indian subcontinent.

Aあ
Indo-Aryan
language family

Status: Included in the Eighth Schedule to the Constitution. Official secondary language of the state of Jammu and Kashmir.

Script: Devanagari script in India and written from right to left using Nastaleeq form of Perso-Arabic script in Pakistan.

Devanagari
Perso-Arabic
script

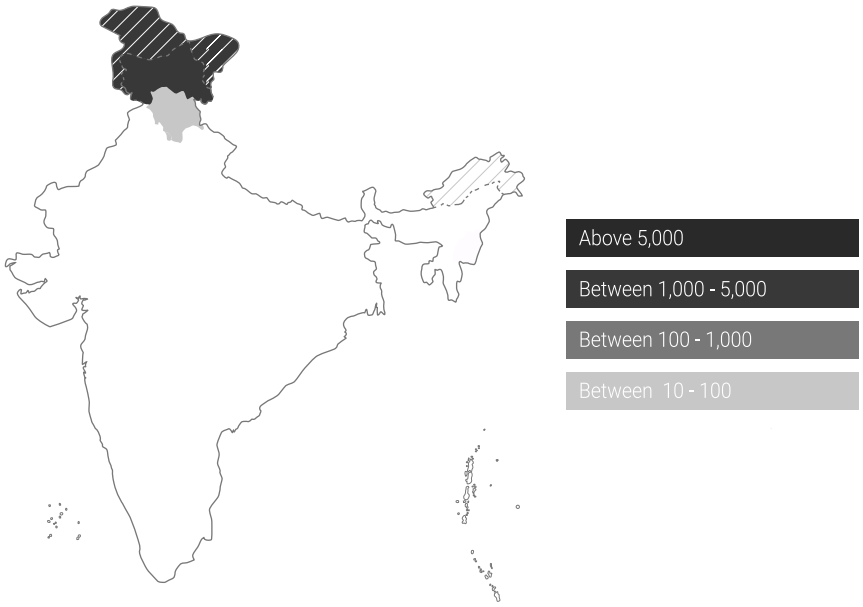
Distribution: Over 2 million native speakers in India, chiefly in the Jammu region of Jammu and Kashmir and Himachal Pradesh, and another 2 million in Pakistan.

Average
per-capita income
of native speakers

Dialects: Dogri speakers understand each other well. Department of Dogri at Jammu University designated Samba as the standard dialect and published textbooks based on this variety.

Low
localization grade

Andovar



French

français indien

Introduction: Indian French is a dialect of French spoken by Indians in the former colonies of Puducherry and Chandannagar. Currently, the French language owes its presence in India to both an established network of Alliances Françaises cultural institutes and a solid basis for the language itself in secondary schools, where French is the first foreign language studied. The reason why French enjoys a privileged status in India is probably due to the successful decolonization, in the 1950s, of the five trading centers, which France had owned in India since the 17th century. The best known of these is Pondicherry, which the first prime minister of India, Jawaharlal Nehru, called "an open window on French culture".

? "French India" is the name commonly used to refer to the French possessions acquired by the French East India Company from the second half of the 17th century onward. They included Pondicherry, Karikal, and Yanaon on the Coromandel Coast, Mahé on the Malabar Coast, and Chandannagar in Bengal. They were incorporated into the Union of India in 1947 and 1954.

Status: Not included in the Eighth Schedule to the Constitution, official language of the union territory of Puducherry.

Script: Latin script

Distribution: Around 10,000 in Puducherry, some in Chandannagar, a small city and former French colony located 30km north of Kolkata, in West Bengal; over 60,000 in France.

Dialects: There are several varieties of Indian French, corresponding to the former French colonies: Tamil (Pondicherry Tamil dialect), Telugu (Yanam Telugu dialect), and Malayali (Mahe Malayalam dialect). None has been recognized as a standard variety.

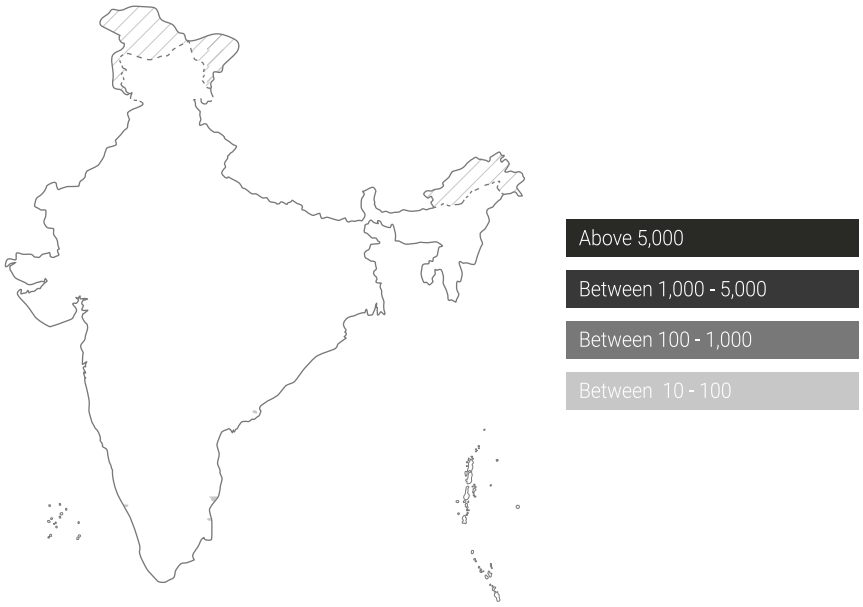
Under 1M
native speakers worldwide

Aあ
Indo-European
language family

Latin
script

Average to High
per-capita income
of native speakers

Low
localization grade



Garo

গারো
(Bengali Script)

Introduction: Garo also knows as A·chik has a close affinity to Bodo, the language of one of the dominant communities of the neighboring state of Assam. In 1996, the Department of Garo was established by popular demand to document in audio and videotapes parts of Garo folktales, folksongs, and traditional oral poetry.

Garos are mainly Christians. The Bible was translated into Garo language in 1924.

Status: Included in the Eighth Schedule to the Constitution. Secondary official language of the state of Meghalaya.

Script: Garo uses the Bengali script as well as Latin script.

Distribution: Number of native speakers is estimated at under 1 million. Garo is the language of the majority of the people of the Garo Hills in the Indian state of Meghalaya. Garo is also used in several districts of Assam, as well as in neighboring Bangladesh.

Dialects: The list of "dialects" of Garo is based on the list of sub-tribes of a people who share the same clan names. There are no sharp boundaries between different "dialects" and not all of them are mutually intelligible. Therefore, from a Western Linguistic point of view, they would be considered as separate languages. However, in India the term "language" is reserved for officially recognized speech varieties; therefore, the term “dialect” is used. Some of the dialects of Garo such as A’tong and Dual are spoken in both India and Bangladesh.

1M
native speakers worldwide

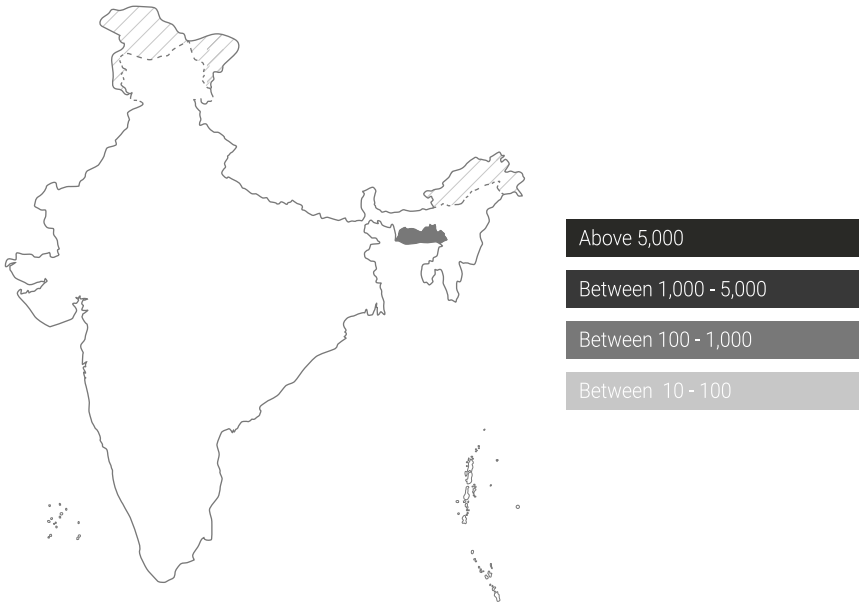
Aあ
Indo-Aryan
language family

Bengali
Latin
script

Average
per-capita income
of native speakers

Low
localization grade

Andovar



Gujarati

ગુજરાતી
(Gujarati Script)

Introduction: Gujarati is a modern Indo-Aryan language evolved from Sanskrit with history that can be traced back to the 12th century. Gujarati is a direct descendant from Old Gujarati (c. 1100 – 1500 BCE).

Gujarati was the first language of Mahatma Gandhi.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Gujarat. Official secondary language of two union territories: Dadra and Nagar Haveli and Daman and Diu.

Script: Gujarati script

Distribution: 46 million native speakers in India, mainly in Gujarat and Mumbai, and another few million in other countries spread out worldwide, especially in the United Kingdom and the United States. Smaller numbers can be found in Uganda, Tanzania, Kenya, Mauritius, Réunion Island, and in Karachi, Pakistan.

Dialects: The accepted standard dialect is the speech of the area from Baroda (Vadodari dialect) to Ahmedabad and north. Kutchi language is often referred to as a dialect of Gujarati, but most linguists consider it closer to Sindhi.

46M
native speakers worldwide

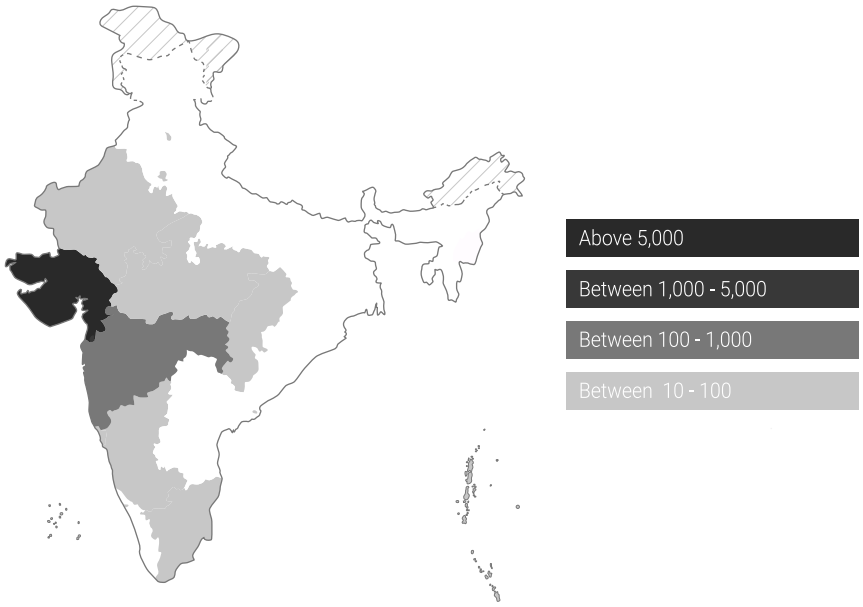
Aあ
Indo-Aryan
language family

Gujarati
script

High
per-capita income
of native speakers

Medium
localization grade

Andovar



Kannada

ಕನ್ನಡ
(Kannada Script)

Introduction: Native speakers of Kannada are called Kannadigas. Kannada is considered to be at least 1,500 years old and has an unbroken literary history of over 1,000 years.

> The language uses 49 phonemic letters, divided into three groups: swaragalu (vowels – 13); vyanjanagalu (consonants – 34 letters); and yogavaahakagalu (neither vowel nor consonant – two letters).

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Karnataka. One of the Classical Languages of India.

Script: Kannada script

Distribution: 38 million native speakers in India and another 9–10 million thought to speak it as a second language. A significant number of Kannada-speaking people can also be found in the United States, United Arab Emirates, Singapore, Australia, and the United Kingdom.

Dialects: There is a considerable difference between the spoken and written forms of the language. Spoken Kannada tends to vary from region to region. The written form is more or less consistent throughout Karnataka. Three regional varieties of Kannada are:

1. Southern (cities of Mysore and Bangalore);
2. Northern (Hubli-Dharwad);
3. Coastal (Mangalore).

The Mysore-Bangalore variety is considered the most prestigious. Social varieties are characterized by education and class or caste, resulting in at least three distinct social dialects: Brahman, non-Brahman, and Dalit (formerly known as “untouchables”).

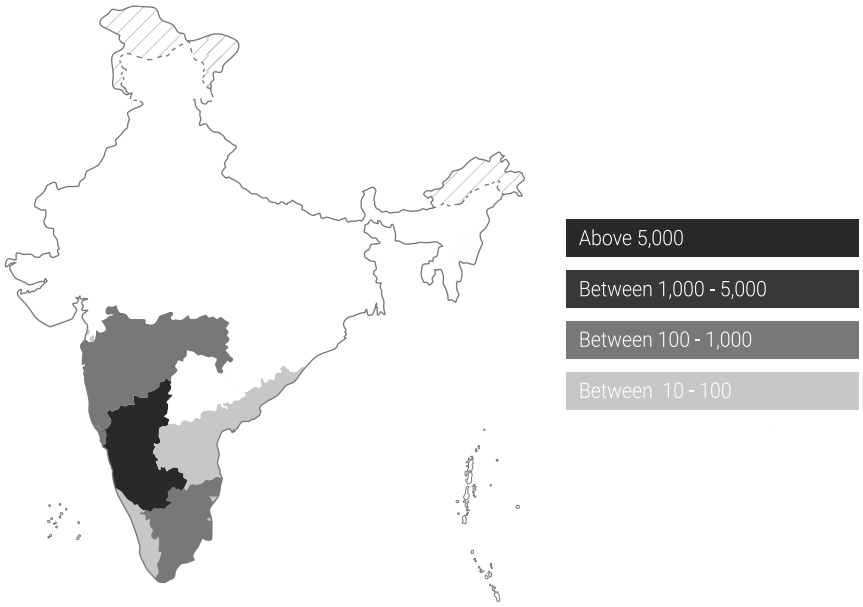
38M
native speakers worldwide

Aあ
Dravidian
language family

Kannada
script

High
per-capita income
of native speakers

Medium
localization grade



Kashmiri

कॉशुर
(Devanagari Script)

كأ شُر
(Perso-Arabic Script)

Introduction: Kashmiri is spoken primarily in the Kashmir Valley, in Jammu and Kashmir. Since November 2008, it has been made a compulsory subject in all schools in the valley up to the secondary level. Reflecting the history of the area, the Kashmiri vocabulary is mixed, containing Dardic, Sanskrit, Punjabi, and Persian elements. Like other North Indian languages, Kashmiri branched off from the Indo-Aryan Sanskrit, but had another ancestor before that, the Shina languages of the Indo-Iranian family. When mighty Sanskrit came, Shina was quickly pushed aside. Additionally, from about the 14th century, medieval Persian started creeping into Kashmiri. With such foreign influences, the Kashmiri language boasts of peculiarities, which no other Indian language has, such as certain vowel and consonant sounds.

Kashmiri literature dates back over 750 years, which is more-or-less the same as literature in modern English.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Jammu and Kashmir.

Script: After the 8th century, Kashmiri was traditionally written in the Sharada script. This script is not in common use today, except for religious ceremonies, and modern Kashmiri is written using the Devanagari script by Hindus and right to left in the Perso-Arabic script by Muslims.

Distribution: 6 million native speakers in India. Over 100,000 speakers in Pakistan, who are mostly emigrants from the Kashmir Valley after the Partition.

Dialects: There is a minor difference between the Kashmiri spoken by Hindus and Muslims.

6M
native speakers worldwide

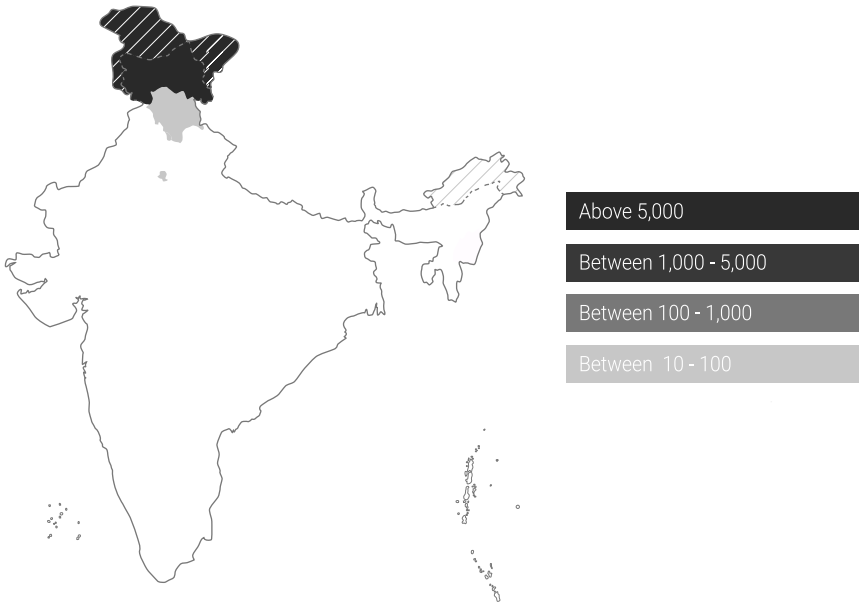
Aあ
Indo-Aryan
language family

Devanagari
Perso-Arabic
script

High
per-capita income
of native speakers

Low
localization grade

Andovar



Khasi

Introduction: Khasi is an Austroasiatic language spoken primarily in Meghalaya state in India by the Khasi people. Khasi is part of the Austroasiatic language family, and is closely related to the Munda branch of that family, which is spoken in East-Central India. Khasi is rich in folklore and folktale as the basis of stories for most of the names of hills, mountains, rivers, waterfalls, birds, flowers, and animals.

Under 1M
native speakers worldwide

Until 2012, Khasi was considered an endangered language by UNESCO.

Aあ

Austroasiatic
language family

Status: Not included in the Eighth Schedule to the Constitution. Official secondary language of the state of Meghalaya.

Latin
script

Script: In the past, the Khasi had no script of its own. A large number of Khasi books were initially written in the Assamese script, but owing to a Welsh missionary, since 1841 a slightly modified Latin script is used.

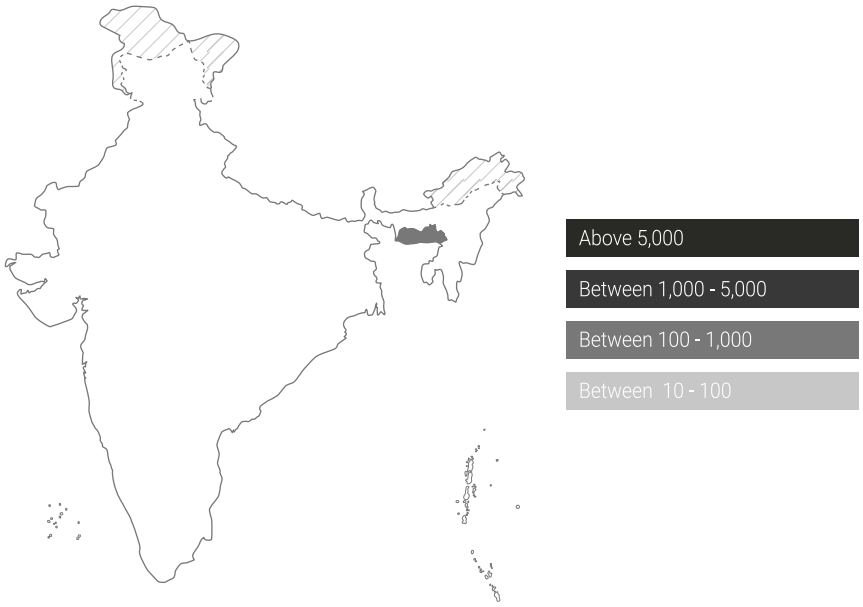
Average
per-capita income
of native speakers

Distribution: Most of the 865,000 Khasi speakers are found in Meghalaya state and the language is also spoken by a number of people in the hill districts of Assam bordering with Meghalaya and by a sizable population of people living in Bangladesh, close to the Indian border.

Low
localization grade

Dialects: Khasi has significant dialectal variation. Several dialects are only partially mutually intelligible, and some are distinct enough to be sometimes considered separate languages. Cherrapunji is considered the standard dialect.

Andovar



Kok Borok

কোকবরক
(Bengali Script)

Introduction: The Borok language, or Kok Borok (kók "language" and borok "people"), also known as Tripuri, means any of the native languages of the Tripuri people of the Indian state of Tripura and neighboring areas of Bangladesh. Kok Borok has existed in its various forms since at least the 1st century. During the rule of the Borok Kings in the Kingdom of Tripura from the 14th until the 20th century, Kok Borok was relegated to a common people's dialect, in contrast to Bengali, which was given higher status. It was finally recognized as an official language of the Tripura state in 1979 along with Bengali.

Kok Borok is not a single language, but a collective name for the several languages and dialects spoken in Tripura.

Status: Not included in the Eighth Schedule to the Constitution. Official language of Tripura state. Proposed for inclusion in the Eighth Schedule in 2004.

Script: Kok Borok had a script known as Koloma, which has disappeared. Since the 19th century, the Kingdom of Tripura has used the Bengali script for writing in Kok Borok. Since Independence and Tripura's merger with India, a Latin script has been promoted by NGOs. This is a highly politicized issue. At present, both scripts are used in the state for education as well as in literary and cultural circles.

Distribution: Under 1 million, mostly in Indian state of Tripura and neighboring areas of Bangladesh.

Dialects: The three main dialects are not mutually intelligible, though the Debbarma dialect of the royal family is a prestige dialect understood by everyone.

1M
native speakers worldwide

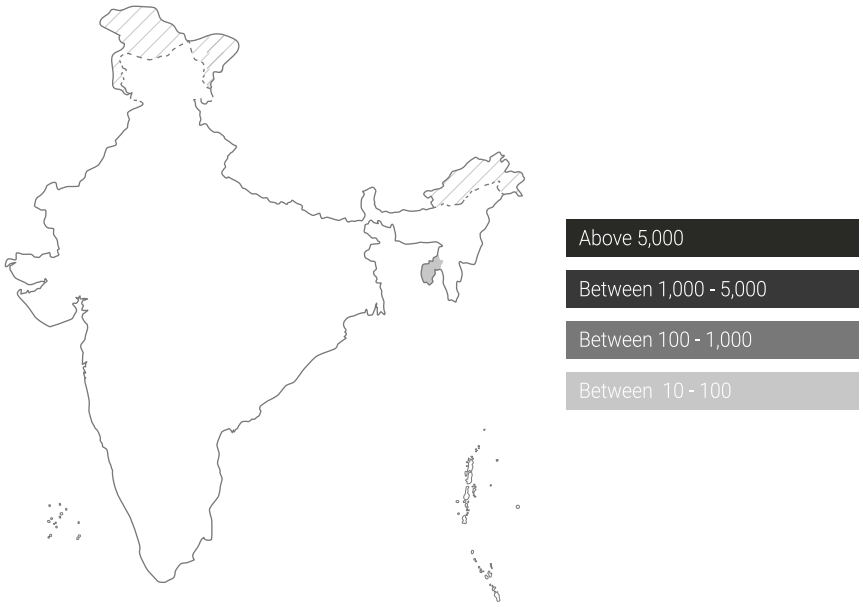
Aあ
Tibeto-Burman
language family

Bengali
script

Low
per-capita income
of native speakers

Low
localization grade

Andovar



Konkani

कोंकणी
(Devanagari Script)

ಕೊಂಕಣಿ
(Kannada Script)

Introduction: Konkani or Goani is similar to Marathi in some aspects, since their speakers have lived in the same area, but Konkani is the older of the two. The first known Konkani inscription dates from 1187. It has also been influenced by Gujarati, Arabic, and Persian as well as Portuguese (in dialect spoken by Goan Catholics).

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Goa and the union territory of Daman and Diu.

Vaman Raghunath Varde Valaulikar was credited with a revival of Konkani after it was nearly made obsolete by the more powerful languages of the region. His death anniversary on the 9th of April is celebrated as World Konkani Day.

Script: Konkani is written in five scripts today. The Goan Hindus use the Devanagari script in their writings while the Goan Catholics use a Latin script. The Saraswats of Karnataka use the Devanagari script in the North Kanara district, but those in Udupi and South Kanara use the Kannada script. The Karnataka Christians also use the Kannada script. The Malayalam script was used by the Konkani community in Kerala, but now there is a move to use the Devanagari script instead. Konkani Muslims around Bhatkal taluk of Karnataka use Perso-Arabic script to write Konkani from right to left.

Distribution: 2.5 million native speakers living along the West coast of India, mainly in Goa, Dadra and Nagar Haveli, Daman and Diu, Maharashtra, Karnataka and northern Kerala. A significant number of Konkani speakers are found in Kenya, Uganda, Pakistan, the Persian Gulf, and Portugal, because during Portuguese rule many Goans had migrated to these countries.

Dialects: Many immigrant families still continue to speak different dialects that their ancestors spoke, which are now highly influenced by the native languages. Similarly, within India Konkani has several dialects influenced by the dominant language of the region, but the Goan Antruz dialect in Devanagari script is considered Standard Konkani.

6M
native speakers worldwide

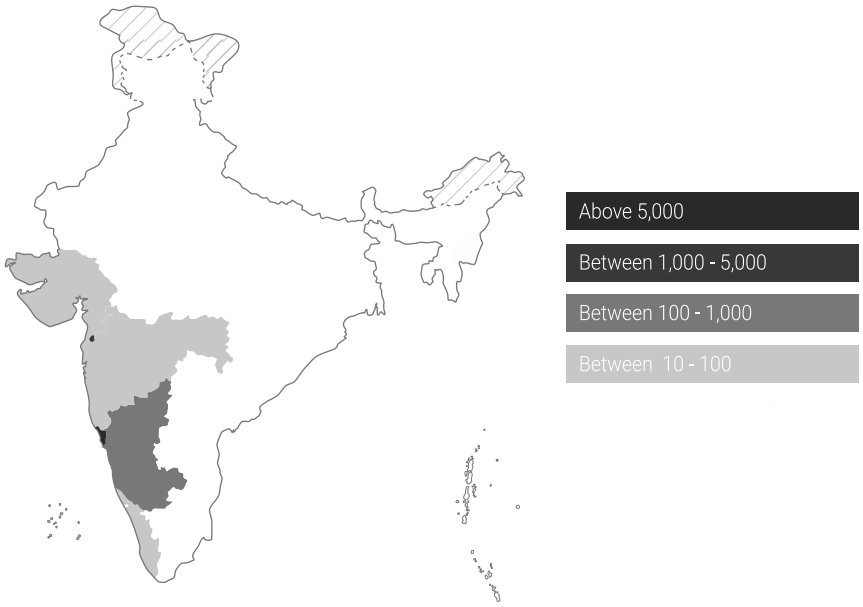
Aあ
Indo-Aryan
language family

Devanagari
Kannada
script

High
per-capita income
of native speakers

Low
localization grade

Andovar



Maithili

मैथिली
(Devanagari Script)

𑒧𑒻𑒟𑒱𑒪𑓂
(Maithili/ Tirhuta Script)

Introduction: The name Maithili is derived from the word Mithila, which was also the name of the ancient kingdom ruled by King Janaka from the Ramayana. Scholars in Mithila used Sanskrit for their literary work and Maithili was the language of the common folk. The earliest work in Maithili is dated to about the 13th century.

It used to be considered a dialect of Hindi and Bengali. However, Maithili achieved an official language status in India in the year 2003 thanks to a mass movement that called for inclusion of Maithili in the Eighth Schedule of the Indian Constitution.

> **?** Creation of a new state called Mithila comprising Maithili speakers in North Bihar and surrounding areas was proposed, but has not gained enough political momentum so far.

Status: Included in the Eighth Schedule to the Constitution since 2003. Official language of the state of Bihar.

Script: The Tirhuta or Maithili script (similar to Bengali) was traditionally used and has a rich history spanning a millennium, but years of neglect by the Bihar government have taken their toll on the use of Tirhuta and nowadays most speakers of Maithili language have switched to using the Devanagari script. As a result, the number of people with a working knowledge of Tirhuta has dropped considerably in recent years.

Distribution: 12 million native speakers in India, mostly in Bihar state. It is also the second most spoken language of Nepal with a few million speakers in southeastern part of the country. Total population of all speakers worldwide is estimated at around 34 million.

Dialects: Tharuwat dialect is spoken exclusively in Nepal. Jolaha Maithili is a dialect spoken by Muslims. Central Maithili is the standard form, in which books are written.

34M
native speakers worldwide

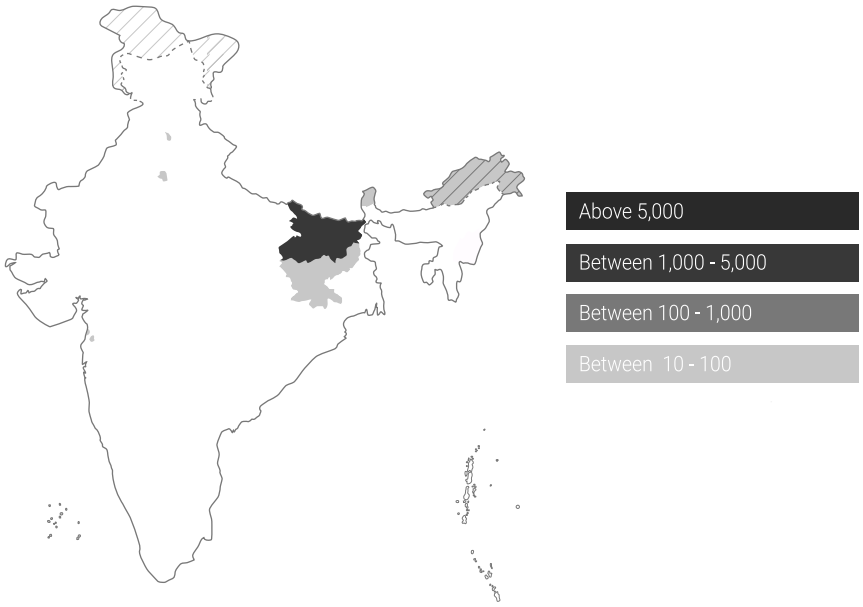
Aあ
Indo-Aryan
language family

Devanagari
Maithili/ Tirhuta
script

Average
per-capita income
of native speakers

Low
localization grade

Andovar



Malayalam

മലയാളം
(Malayalam Script)

Introduction: Linguists are not sure whether Malayalam (sometimes called Malabar) originated as a dialect of Tamil or an independent offshoot of a Proto Dravidian language. Either way, it is generally agreed that by the end of the 13th century, a written form of the language emerged that was distinct from Tamil.

The word Malayalam probably originated from the Malayalam/ Tamil words “mala” meaning hill, and “elam” meaning region. Malayalam translates as "hill region" and used to refer to the land itself, and only later became the name of the language.

The word “Malayalam” is one of the longest palindromes that exist in English (word that reads the same forward and backward).

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Kerala. Official secondary language in union territories of Lakshadweep and Puducherry. One of the Classical Languages of India.

Script: The Grantha script was used in the past and it later developed into the modern Malayalam script used today. It is also written from right to left with a version of the Perso-Arabic script by Muslims in Singapore and Malaysia, and occasionally by Muslims in Kerala.

Distribution: 33 million native speakers in India, over 90% of which live in Kerala. Malayalam is also spoken in the neighboring states of Tamil Nadu and Karnataka. A large number of Malayalis have also emigrated to the Middle East, the United States, and Europe, with an estimated few hundred thousand speakers in each region.

Dialects: Malayalam has three important regional dialects and a number of smaller ones. There is some difference in dialect along social, particularly caste, lines.

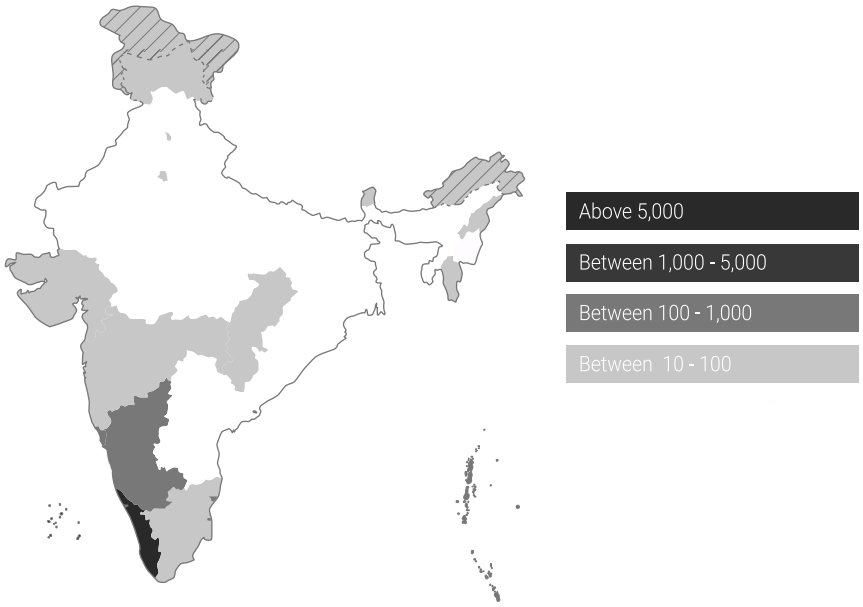
33M
native speakers worldwide

Aあ
Dravidian
language family

Malayalam
script

High
per-capita income
of native speakers

Medium
localization grade



Manipuri

মৈতৈলোন্ (Meitei Mayek Script)

মণিপুরী (Bengali Script)

Introduction: Manipuri, also known as Meithei, Meetei, Meitei, Meithei-lon, Meithei-lol, and pangal-lol, is a Tibeto-Burman language whose exact classification remains unclear. Meithei has proven to be an integrating factor among all distinct ethnic groups in Manipur who use it to communicate among themselves. It has been recognized (as Manipuri), by the Indian government and included in the list of scheduled languages.

Manipuri is a tonal language with two tones, namely: high and low.

> Many old Meithei documents were destroyed at the beginning of the 18th century during the reign of Hindu-converted King Pamheiba, under the instigation of Bengali Hindu missionaries.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Manipur. Official secondary language in Tripura state.

Script: The language has its own script, known as Meetei Mayek script, which was originally used until the 18th century. Since the advent of British rule in 1891, the Bengali script has been used to write Meithei. Nowadays, the Bengali script is being gradually replaced by the Meetei Mayek script in schools.

Distribution: 1.5 million native speakers in India, predominantly in Manipur state. Smaller speech communities exist in the Indian states of Assam, Mizoram, and Tripura, as well as in Bangladesh and Myanmar.

Dialects: Those in Bangladesh may understand Indian Meithei better than vice-versa possibly due to more language changes in Bangladesh over the years. Indian Meithei is more standard.

1.5M
native speakers worldwide

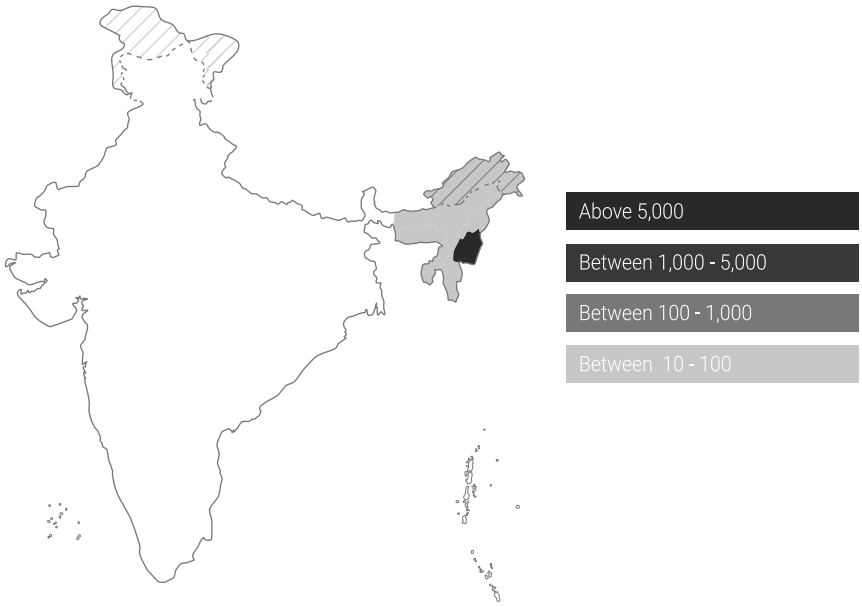
Tibeto-Burman
language family

Meitei Mayek
Bengali
script

Average
per-capita income
of native speakers

Low
localization grade

Andovar



Marathi

मराठी
(Devanagari Script)

Introduction: The odyssey of written Marathi begins in the 11th century from stone inscriptions and copper plates. Since the 17th century, Marathi started gaining prominence with the rise of the Maratha Empire beginning with the reign of Chhatrapati Shivaji. His name now graces the international airport and main train terminal in Mumbai, the capital of Maharashtra state. Marathi has the fourth largest number of native speakers in India.

73M
native speakers worldwide

The most comprehensive Marathi-English dictionary was compiled by Captain James Thomas Molesworth in 1831. The book is still in print nearly two centuries after its publication.

Aあ
Indo-Aryan
language family

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Maharashtra and union territories of Daman and Diu and Dadra and Nagar Haveli. Secondary language in the state of Goa.

Devanagari
script

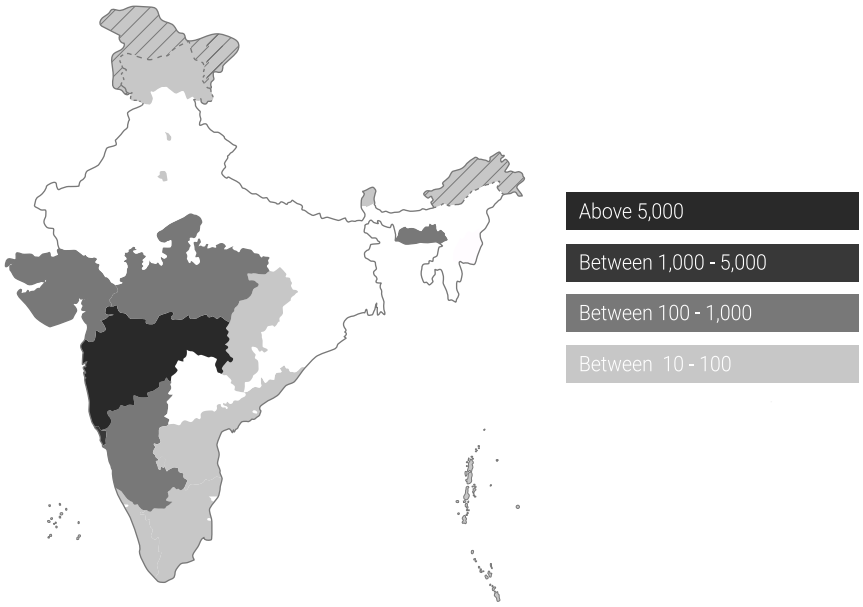
Script: From the 13th century until the mid-20th century, Marathi was written in the Modi script. Since 1950, it has been written in Devanagari script.

Very High
per-capita income
of native speakers

Distribution: 73 million native speakers in India, chiefly in Maharashtra and parts of neighboring states of Gujarat, Madhya Pradesh, Goa, Karnataka, Chhattisgarh and Andhra Pradesh, union territories of Daman and Diu and Dadra and Nagar Haveli. Marathi is also spoken by Maharashtrian emigrants worldwide, especially in the United States, Israel, Mauritius, and Canada.

High
localization grade

Dialects: The major dialects of Marathi are called Standard Marathi and Warhadi Marathi. There are a few other sub-dialects like Ahirani, Dangi, Vadvali, Samavedi, Khandeshi, and Malwani. Standard Marathi is the official language of the state of Maharashtra.



Mizo

Introduction: The Mizo language is spoken natively by the Mizo people. The language is also known as "Lushai", a colonial term still commonly used, although considered incorrect by the Mizo themselves. It is a tonal language with eight distinct tones. There is no gender for nouns and no articles.

Mizo has a thriving literature with Mizo departments in Mizoram University and Manipur University. The governing body is the Mizo Academy of Letters, which awards the annual literary prize MAL Book of the Year since 1989.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Mizoram.

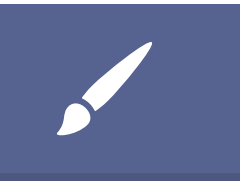
Script: Christian missionaries started developing an alphabet for the language by adapting it to a Latin alphabet. The result is 25 letters used for writing the Mizo language.

Distribution: Less than 1 million native speakers in India, spoken by the Mizo people in the Mizoram state of India. Some also in Chin State in Burma, and the Chittagong Hill Tracts of Bangladesh.

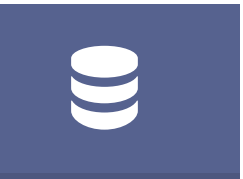
Dialects: The various clans of the Mizo peoples had respective dialects, among which the Lushei dialect was most common, and which subsequently became the Mizo language and the standard form, due to its extensive and exclusive use by the Christian missionaries.

1M
native speakers worldwide

Tibeto-Burman
language family

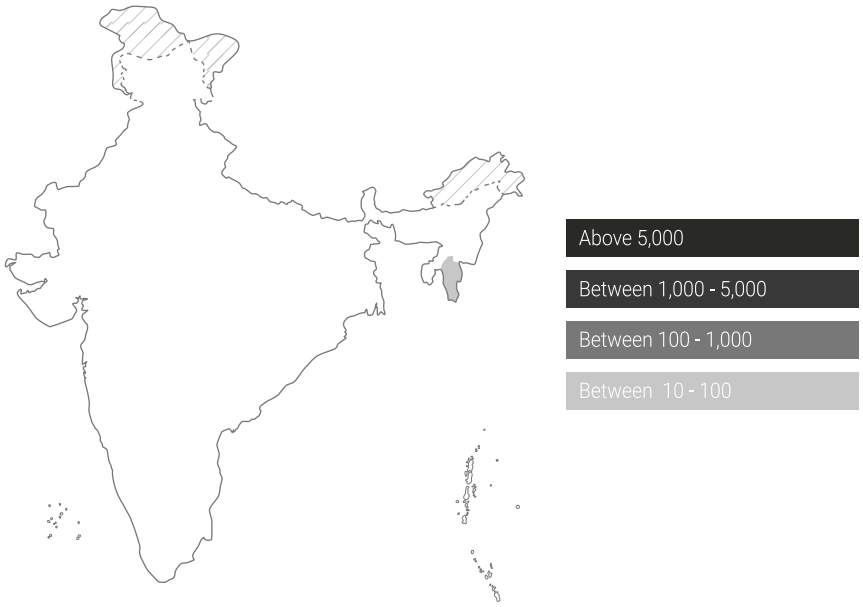


Latin
script



Average
per-capita income
of native speakers

Low
localization grade



Nepali

नेपाली
(Devanagari Script)

Introduction: Nepali or Nepalese is the official language and de facto lingua franca of Nepal. While Nepali is technically from the same family as languages like Hindi and Bengali, it has taken many loan words. Nowadays, it shares a 40% lexical similarity with the Bengali language and has many similarities with Hindi.

15M
native speakers worldwide

In the past, Nepali was known as Gorkhali or Gurkhali (language of the Gorkha Kingdom).

Aあ
Indo-Aryan
language family

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Sikkim. Official secondary language of the state of West Bengal.

Script: Many scripts have been used to write Nepali in the past, but nowadays Devanagari script is considered the standard.

Devanagari
script

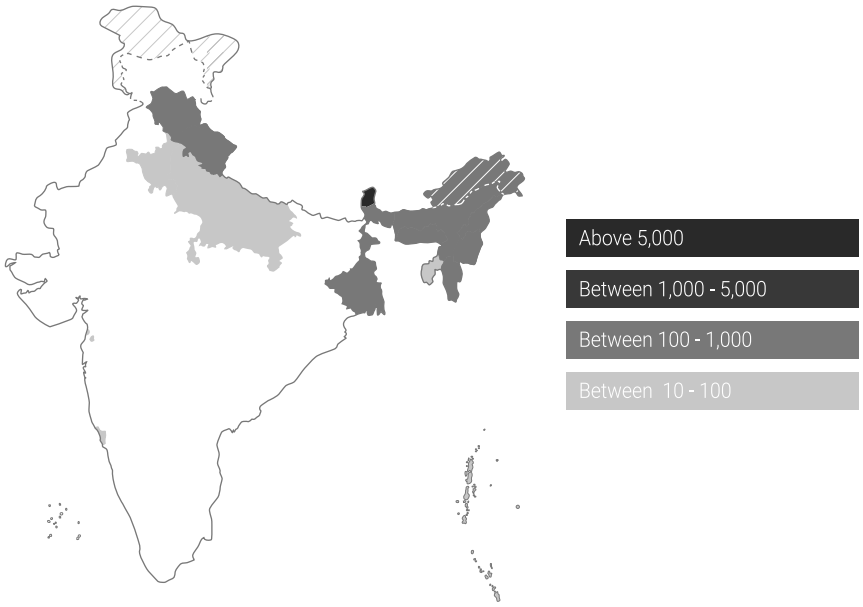
Distribution: 3 million native speakers in India, mostly in Sikkim, Darjeeling district of West Bengal and Assam, as well as in the whole North-East India region and most major cities of the country. Outside of India, it's the official language of Nepal, spoken as the mother tongue by nearly half of the population. Smaller communities exist in Bhutan, Brunei, and Myanmar. Estimated total number of speakers worldwide is over 15 million.

High
per-capita income
of native speakers

Dialects: The standard Nepali is as spoken in Nepal. There are three major dialects distinguished: eastern, central, and western with little variation in phonology from one to another.

Low
localization grade

Andovar



Oriya/ Odia

ଓଡ଼ିଆ
(Oriya Script)

Introduction: Oriya, Odissa, or Odisha is officially spelled Odia and dates back to around the 10th century. It is billed as the first language from the Indo-Aryan linguistic group and the case for making it a classical language was premised on the fact that it has no resemblance to Hindi, Sanskrit, Bengali, and Telugu.

Odia was the first modern language from the Indo-Aryan family to be added to the Classical Languages of India list in 2014.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Odisha. Official second language of Jharkhand. One of the Classical Languages of India.

Script: Oriya script

Distribution: 33 million native speakers in India, mostly in the state of Odisha, where native speakers comprise 80% of the population, and in parts of West Bengal, Jharkhand, Chhattisgarh and Andhra Pradesh states. Due to the increasing migration of labor, the West Indian state of Gujarat also has a significant Oriya-speaking population. Additionally, there are many speakers in Bangladesh and Indonesia, and in Western countries such as the United States, Canada, Australia, and the United Kingdom.

Dialects: The language has several major dialects with Mughalbandi (Coastal Oriya) considered the standard dialect and the language of education.

33M
native speakers worldwide

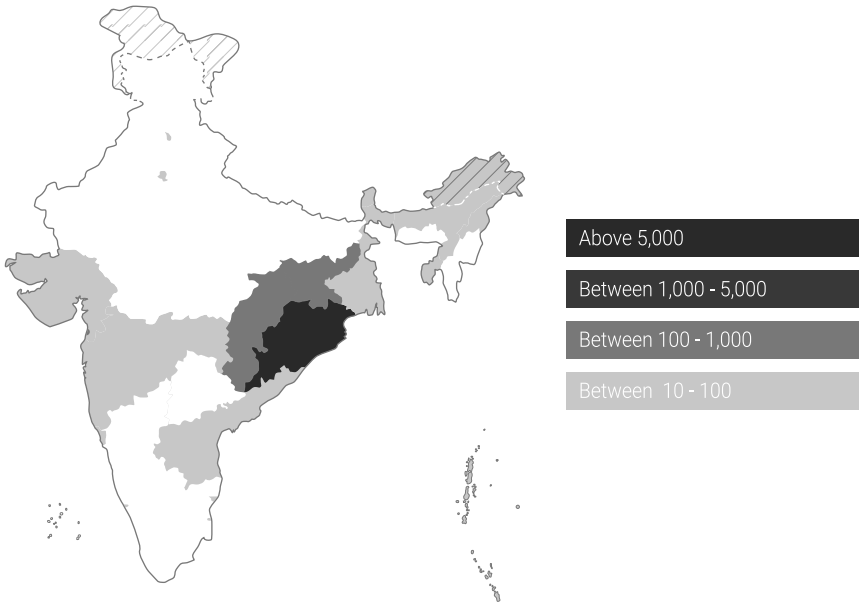
Aあ
Indo-Aryan
language family

Oriya
script

High
per-capita income
of native speakers

Medium
localization grade

Andovar



Punjabi

ਪੰਜਾਬੀ
(Gurmukhi Script)

پنجابی
(Shahmukhi Script)

Introduction: Punjabi or Panjabi is the native language of the Punjabi people who inhabit the historical Punjab region of Pakistan and India. It emerged as an independent language in the 12th century and is the only tonal language among the major Indo-Aryan languages. The Sikh religion originated in the 15th century in the Punjab region and Punjabi is the predominant language spoken by Sikhs.

Although, it is the most widely spoken language in Pakistan, it has no official recognition by the government. It is also the 11th most widely spoken in India, the 4th most spoken language in England and Wales, and the 3rd most spoken in Canada. The influence of Punjabi as a cultural language on the Indian subcontinent is increasing day by day due to Bollywood. Most Bollywood movies now have Punjabi vocabulary mixed in, along with a few songs fully sung in Punjabi. At any point in time, Punjabi songs in Bollywood movies now account for more than 50% of the top of the charts listings.

The word Punjabi is derived from the word Punjab, Persian for "Five Waters", which refers to five major eastern tributaries of the Indus River.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Punjab and the union territories of Chandigarh and Delhi. Official secondary language of Haryana state.

Script: In India, Punjabi is written in either Gurmukhi script or Devanagari script. The Muslims in the region later created the Shahmukhī script based on Persian Nastaleeq script.

Distribution: 29 million in India and 70 million in Pakistan. There are also important overseas communities of Punjabi speakers, particularly in the United Kingdom, the United States, and Canada.

100M
native speakers worldwide

Aあ
Indo-Aryan
language family

Gurmukhi
Shahmukhi
script

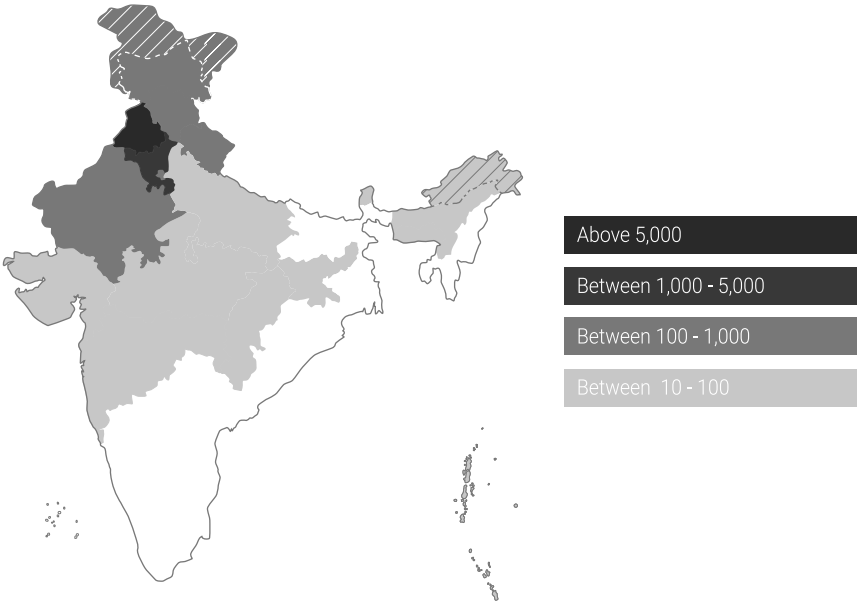
Very High
per-capita income
of native speakers

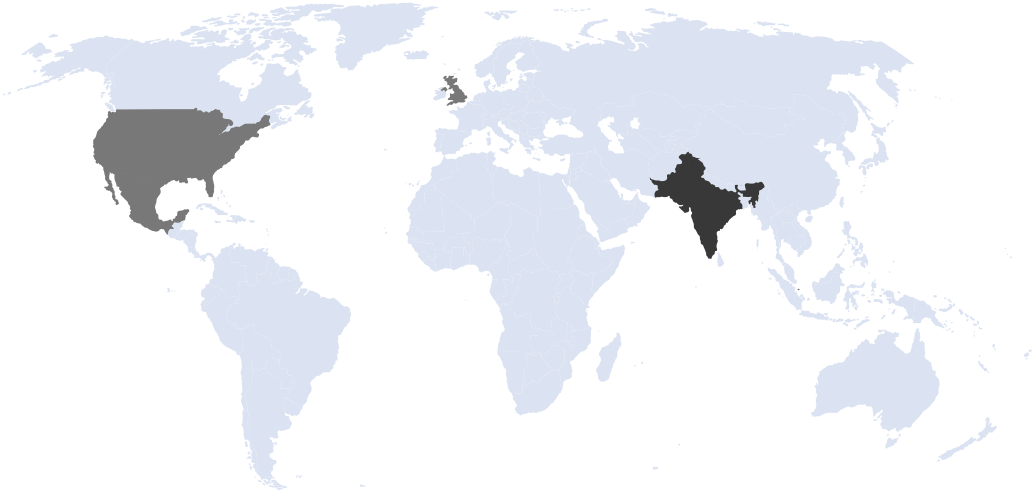
High
localization grade

Andovar

Dialects: The major dialects of Punjabi include Majhi, Doabi, Malwai, Powadhi, Pothohari, and Multani. The dialects in the Lahnda dialect continuum, including Saraiki and Hindko, are considered as dialects of Punjabi by many linguists but as distinct languages by others. The Majhi dialect spoken around Amritsar is Punjabi's prestige dialect because it is the standard of written Punjabi. The Majhi (and Lahnda) spoken in Pakistan is more Persianized in vocabulary and somewhat different in pronunciation.

Punjabi spoken in India is sometimes referred to as Eastern Punjabi, while the Pakistani variant is called Western Punjabi.





Sanskrit

संस्कृतम्
(Devanagari Script)

Introduction: Sanskrit is the primary liturgical language of Hinduism, a philosophical language in Hinduism, Buddhism, and Jainism, and a scholarly literary language that was in use as a lingua franca for Indian culture. Although, nowadays, there are very few speakers of Sanskrit when compared to other languages used in India, it holds a special position in the country. You can compare it to Latin and Ancient Greek in Europe and it has significantly influenced most modern languages of the Indian subcontinent.

Despite being the linguistic progenitor of the major Indo-Aryan languages used in the north of India, Sanskrit retains a somewhat pan-Indian character and is held in high regard in both the north and the south. This is largely due to extensive borrowing from Sanskrit into the Dravidian languages spoken in the south.

Sanskrit has been a "dead" language for hundreds and hundreds of years, yet it has remained relatively intact and unchanged due to religious sanctions to retain its original character and pronunciation. Today, it is studied by many Indians for both religious and scholastic reasons as well as personal interest. It continues to be widely used as a ceremonial language in Hindu religious rituals and Buddhist practice in the forms of hymns and mantras. Spoken Sanskrit has been revived in some villages with traditional institutions, and there are efforts at further popularization.

In the Republic of India, Nepal, and Indonesia, Sanskrit phrases are widely used as mottos for various national, educational, and social organizations (much as Latin is used by some institutions in the West).

Status: Included in the Eighth Schedule to the Constitution. Secondary official language of the state of Uttarakhand. One of the Classical Languages of India.

Under 1M
native speakers worldwide

Aあ
Indo-Aryan
language family

Devanagari
script

High
per-capita income
of native speakers

Low
localization grade

Script: Sanskrit at first did not have a written form, but was maintained as part of oral tradition. When writing was introduced, the choice of writing system was influenced by the regional scripts of the scribes. Therefore, Sanskrit has no native script of its own and virtually all of the major writing systems of South Asia have been used in the past. However, since the late 19th century, Devanagari script has become the de facto standard writing system for Sanskrit publication. Sanskrit can also been written with a Latin alphabet using the International Alphabet of Sanskrit Transliteration (IAST).

Distribution: Around 50,000 people in India speak the Sanskrit language fluently, and about 200,000 as second language. Many Buddhist scholars in Japan, China, and Thailand use Sanskrit language apart from those in India, Sri Lanka, Bangladesh, Nepal, and other areas of South and Southeast Asia.

Dialects: Vedic Sanskrit is the oldest form of the language and can be traced as early as 1700–1200 BCE. It is the language of the Vedas, a large collection of hymns, incantations, and theological and philosophical discussions. Scholars often distinguish Vedic Sanskrit and Classical Sanskrit as separate dialects. Though they are quite similar, they differ in a number of essential points of phonology, vocabulary, grammar, and syntax.

Santali

संताली
(Devanagari Script)

সাঁওতালি
(Bengali Script)

Introduction: Santali or Santhali is a language in the Munda subfamily of Austroasiatic languages, related to Ho and Mundari.

The Santhal people are the largest tribal community in India.

Status: Included in the Eighth Schedule to the Constitution, official language of the Santhal tribes living in the states of Bihar, Chhattisgarh, Jharkhand, and Odisha. Official second language in Jharkhand state.

Script: During the British rule, it was written in Latin script. In 1925, a dedicated Santali alphabet was created. It is also known as Ol Chiki or Ol Cemet' script. Nowadays, Santali is often written in the Devanagari script or Bengali script.

Distribution: Nearly 6 million native speakers in India, mostly in the states of Jharkhand, Assam, Bihar, Odisha, Tripura, and West Bengal. It is also spoken in Bangladesh, Bhutan, and Nepal. Total world population estimated at over 6 million.

Dialects: Several dialects are in use related to the area where speakers reside.

6M
native speakers worldwide

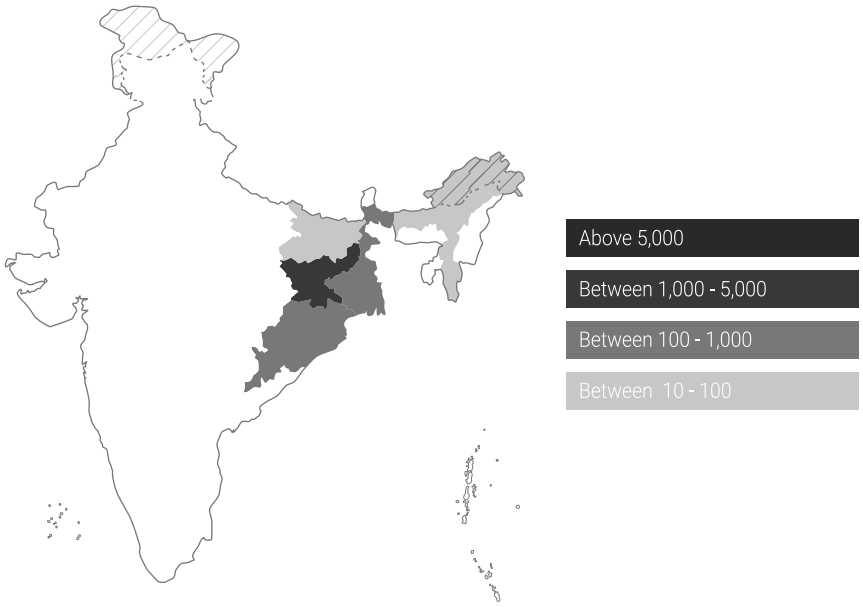
Austroasiatic
language family

Devanagari
Bengali
script

Low
per-capita income
of native speakers

Low
localization grade

Andovar



Sindhi

सिन्धी
(Devanagari Script)

سنڌي
(Arabic Script)

Introduction: Sindhi is an Indo-Aryan language of the Indo-European language family. It is the language of the historical Sindh region, spoken by the Sindhi people and the official language of the Pakistani province of Sindh. It has influences from Balochi spoken in the adjacent province of Balochistan.

Sindhi has a vast vocabulary and a very old literary tradition. This has made it a favorite of many writers and consequently a vast volume of literature and poetry has been written in Sindhi.

Sindhi means “of the Sindhu”. Sindhu was the proper name of the Indus River.

Status: Included in the Eighth Schedule to the Constitution, although it is not an official language of any state.

Script: In India, both the Sindhi-Arabic and Devanagari script are used. The Gujarati script is used to write the Kutchi dialect in India. In Pakistan, it is written with Perso-Arabic script.

Distribution: About 1.7 million native speakers in India concentrated in the states of Rajasthan, Gujarat, and Maharashtra. It is also spoken in Ulhasnagar near Mumbai. The remaining speakers in India are composed of the Hindu Sindhis who migrated from Sindh, after it became a part of Pakistan. In Pakistan, it is spoken by about 18 million people in the Sindh and Balochistan provinces.

Dialects: Several dialects exist with Vicholi forming the basis for standardized Sindhi.

20M
native speakers worldwide

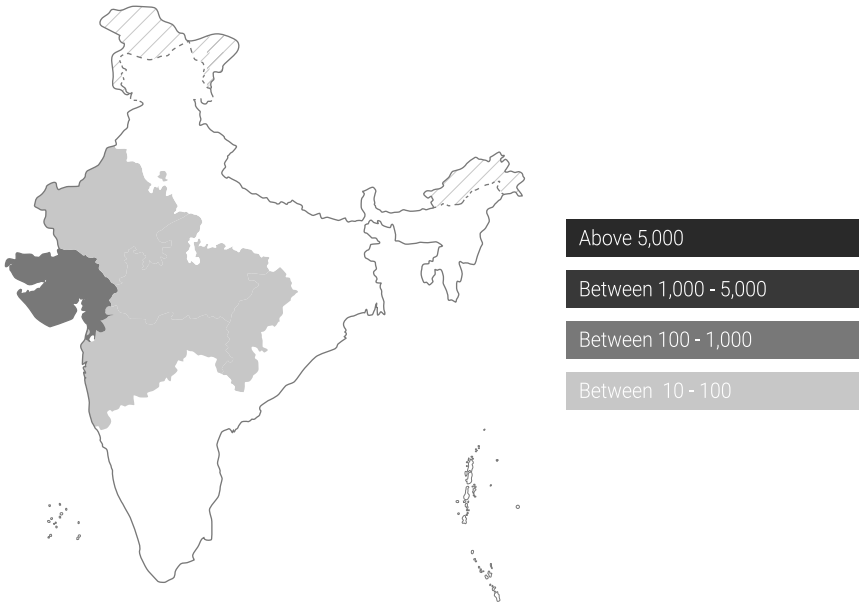
Aあ
Indo-Aryan
language family

Devanagari
Arabic
script

Average
per-capita income
of native speakers

Low
localization grade

Andovar



Tamil

தமிழ்
(Tamil Script)

Introduction: Tamil is one of the longest surviving classical languages in the world. It has been described as the only language of contemporary India, which is recognizably continuous with a classical past. Tamil literature has existed for over 2,000 years, which is the oldest extant literature amongst all Dravidian languages.

The two earliest manuscripts from India to be acknowledged and registered by UNESCO Memory of the World register in 1997 and 2005 are in Tamil.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Tamil Nadu and union territory of Puducherry. Official secondary language in union territory of Andaman and Nicobar Islands. One of the Classical Languages of India.

Script: Tamil script

Distribution: 70 million Tamil is spoken predominantly by Tamil people and has about 60 million native speakers in India, mostly in the south of the country. It is also an official language in Sri Lanka and Singapore and has significant numbers of speakers in Malaysia, Mauritius, Fiji, and South Africa. Population of speakers in all countries is estimated at nearly 70 million.

Dialects: There are two separate registers varying by social status, a high register, and a low one. It is classified as being part of a Tamil language family, which alongside Tamil proper, also includes the languages of about 35 ethno-linguistic groups. Dialects used in India and Sri Lanka vary slightly.

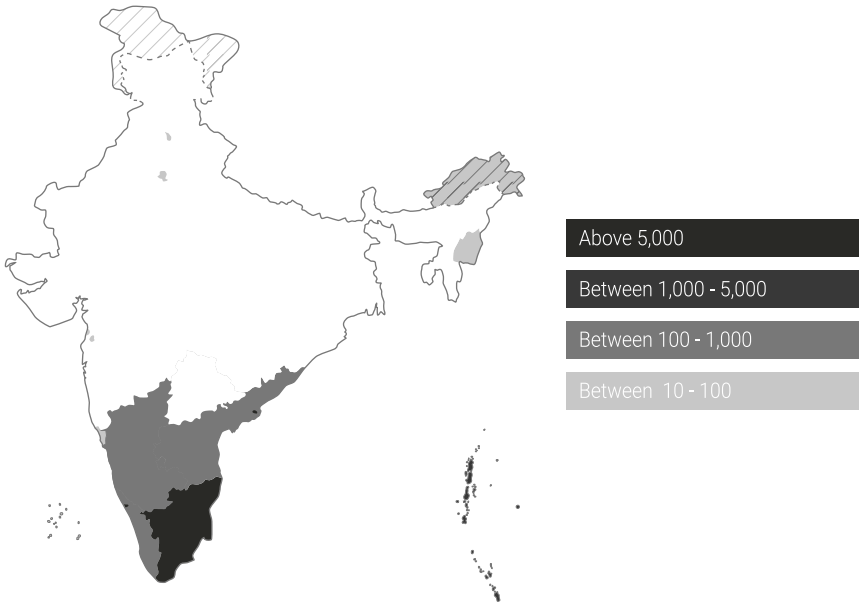
70M
native speakers worldwide

Aあ
Dravidian
language family

Tamil
script

High
per-capita income
of native speakers

High
localization grade



Telugu

తెలుగు
(Telugu Script)

Introduction: Telugu, Telangu, or Tenungu is the most widely spoken Dravidian language and the third most widely spoken language in India after Hindi and Bengali.

Status: Included in the Eighth Schedule to the Constitution, official language of the states of Andhra Pradesh and Telangana. Official secondary language in the union territory of Puducherry (in Yanam). One of the Classical Languages of India.

Script: Telugu script

Distribution: Over 70 million native speakers and a few million second-language speakers in India. Mainly spoken in the states of Andhra Pradesh, Telangana and Yanam district of Puducherry, as well as in the neighboring states of Tamil Nadu, Puducherry, Karnataka, Maharashtra, Odisha, Chhattisgarh, some parts of Jharkhand, and the Kharagpur region of West Bengal in India.

It is also spoken in the United States, where the Telugu diaspora numbers more than 800,000, as well as in Australia, New Zealand, Bahrain, Canada, Fiji, Malaysia, Singapore, Mauritius, Ireland, South Africa, Trinidad and Tobago, the United Arab Emirates, the United Kingdom, and other western European countries from the considerable Telugu diaspora.

Dialects: There are four distinct regional dialects in Telugu, as well as three social dialects that have developed around education, class, and caste. The formal, literary language is distinct from the spoken dialects, a situation known as diglossia.

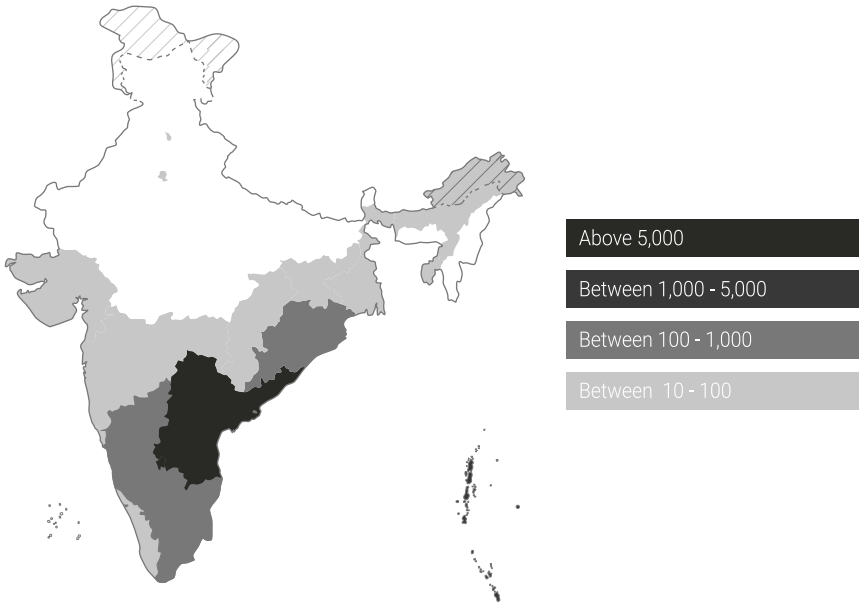
70M
native speakers worldwide

Aあ
Dravidian
language family

Telugu
script

High
per-capita income
of native speakers

High
localization grade



Urdu

اُردُو
(Urdu Script)

Introduction: Urdu is historically associated with the Muslims of the region of Hindustan. It is the national language and lingua franca of Pakistan and an official language of six Indian states. The importance of Urdu in the Muslim world is visible in the Islamic Holy cities of Mecca and Medina in Saudi Arabia, where most informational signage is written in Arabic, English, and Urdu.

From the establishment of the Delhi Sultanate and the Mughal Empire until the British Raj, Hindustani, written in the Urdu script, was the language of both Hindus and Muslims. The language was variously known as Hindi, Hindavi, and Dehlavi. The communal nature of the language lasted until it replaced Persian as the official language in 1837 and was made co-official, along with English. This triggered a Hindu backlash in North-Western India, which argued that the language should be written in the native Devanagari script. Thus a new literary register, called "Hindi", replaced traditional Hindustani as the official language of Bihar in 1881, establishing a sectarian divide of "Urdu" for Muslims and "Hindi" for Hindus, a divide that was formalized with the division of India and Pakistan after Independence. Urdu and Hindi are mutually intelligible because linguistically, they are the same language. If considered the same language, the population of Hindi-Urdu speakers is the fourth largest of the languages of the world, after Mandarin Chinese, English, and Spanish.

? The word Urdu is derived from the same Turkish word “ordu” (army) that has given English “horde”.

Status: Included in the Eighth Schedule to the Constitution, official language of the state of Jammu and Kashmir. Official secondary language in Andhra Pradesh, Bihar, Telangana, Uttar Pradesh, West Bengal, and the union territory of Delhi.

Script: Urdu is written right-to left and is an extension of the Perso-Arabic script.

64M
native speakers worldwide

Aあ
Indo-Aryan
language family

Urdu
script

Average
per-capita income
of native speakers

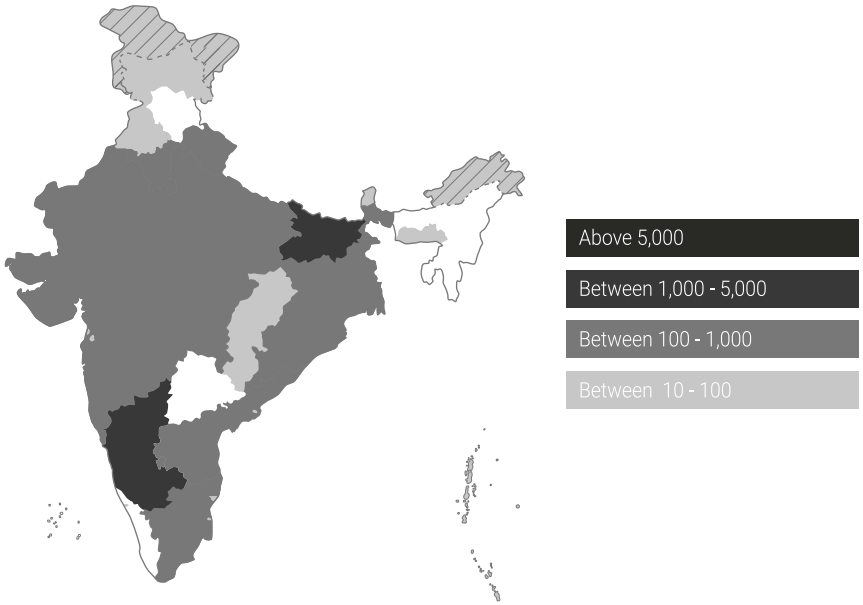
High
localization grade

Andovar

Distribution: Over 50 million native speakers of Urdu live in India, and 13 million in Pakistan (where it is the national and one of the two official languages), with the total population of speakers worldwide estimated at over 100 million. Significant speech communities exist in the United Arab Emirates, the United Kingdom, the United States, and Bangladesh, where it is called "Bihari".

Because of the difficulty in distinguishing between Urdu and Hindi speakers in India and Pakistan, as well as estimating the number of people for whom Urdu is a second language, the number of speakers is uncertain and controversial.

Dialects: Dakhini dialect of Urdu in India has fewer Persian and Arabic loanwords than Urdu. Rekhta is a form of Urdu used in poetry. Although the same language, there and socio-political movements in India, Pakistan, and Bangladesh pushing to make the local variants different from each other.



Conclusion

Addressing the localization needs of more than one billion people speaking many languages is a formidable task. In the Indian context, it becomes even more challenging when we take into consideration that 90% of the content available is in English and therefore, it fails to address the educational and informational requirements of more than 90% of the non-English-speaking population.

On one hand, the non-English-speaking population is not motivated to use computers because digital content in their mother tongue is unavailable. On the other hand, those who have the resources and knowledge to create digital content in regional Indian languages argue that only the English-speaking population is computer savvy. This is a vicious circle, which can be broken by making content available on the internet in more and more regional languages of India. This will then trigger a positive growth loop when availability of more localized content in regional languages will attract more of the non-English-speaking Indian population to use computers and open up new markets in the country.

The current state of the localization industry in India is far from perfect. It lags behind the international level as far as the development and application of localization tools and technologies are concerned. We can say that it is still very much an improvised translation market. The various localization tools developed and available in English have to be first localized and fine-tuned for use with Indian languages. Some of these will have to be customized to suit the scripts of the Indian languages.

At the same time, the advancement which has taken place in the West while progressing from dictionary-armed translators to those using translation memories, linguistic verification tools, machine translation, translation quality assurance and project management tools implies that there is a long way to go for the industry in India.

The benefits of information technology will percolate to every Indian only when computing interfaces are available in various local languages. This may happen sooner than expected. Computers, mobile phones, and connectivity are quickly spreading to every corner of India. This is expected to result in a rapid growth in the demand for multilingual content, localization tools and technologies as well as translation and localization services. Keeping in view the various initiatives underway in government, public, and private sectors, India is poised to become a major supplier of and a market for localized content in the near future.

Three Indian Markets

Segment	Size of market	Characteristics	Interests
Non-Resident Indians (NRIs)	20 million	▸ Mobile and PC-connected ▸ High PC and broadband penetration ▸ Well-off ▸ Early adopters	Entertainment services that offer Indian content (music, movies, news, etc.)
Developed (in India)	300 million	▸ Mobile and PC-connected ▸ Mostly urban, young middle-class ▸ Speakers of major languages	Lifestyle services that serve convenience and productivity
Emerging (in India)	600 million	▸ Limited connectivity (e.g. Kiosk-based) ▸ Mostly rural ▸ Speakers of both major and minor languages	Livelihood services that enable income generation and access to vital information (e.g. official paperwork, healthcare, etc.)

Adapted from: Web for All – Indic languages perspective

To reach the NRIs and the young, urban educated elites in India, English may suffice. However, to go for a more local presentation, it is necessary to consider adjusting the message to suit the market rather than simply re-use content that worked in Western English-speaking countries. The second most important language when trying to reach the highest number of people is Hindi, due to its sheer number of speakers. This one language will open the door to roughly 40% of the population of India and will give your communication efforts a much more localized feel than English.

What comes next? Here it is important to ask a few key questions about your product and target market. If you are trying to reach the highest number of people with no regard for the characteristics of their market segment, you should look at languages with the highest number of speakers after Hindi, such as Bengali, Marathi, Telugu, and Tamil. However, simply adding more languages can quickly become an expensive endeavor and it's more prudent to make a decision to segment the market according to whom your most valuable customers will be. For example, maybe you would rather target the smaller, but more affluent Punjabi-speaking community instead?

Another important factor to take into consideration is how difficult it will be to localize into a specific language. As explained above, nearly all languages used in India fall into one of the two major families: Indo-Aryan or Dravidian. You will find much more similarities both linguistically and culturally between different languages that belong to the same group, than when compared to a language from the other family. Additionally, native speakers of another language are more likely to understand the message even if it's in a different, but similar language of the same family.

India's economy boasts an ever-growing middle class, attracting entrepreneurs from all over the world trying to capture a share. The gross domestic product (GDP) has crossed USD$1 trillion, and by purchasing power parity, it is nearly USD$4 trillion. India is gradually, but steadily opening up sectors for foreign direct investment (FDI), which nearly tripled in recent years as more overseas investors flocked to the country. FDI into India was the second most important FDI destination (after China) for transnational corporations from 2010–2012. These numbers do not include the billions of dollars that have been coming into the stock and bond markets.

Localization is the future in a globalized world and it is not any different for India.

Sources

This is a list of the most important sources we drew on when writing this white paper. Whenever possible, a live direct link was added to allow you to read more about each topic.

- Acharya – Multilingual computing for Literacy and Education
- AKAMAI's State of the Internet – Q1 2014 Report
- Centre for Development of Advanced Computing
- Data on Language from 2001 Indian Census
- Eighth Schedule to the Indian Constitution
- Internationalization & Localization: Indian Perspective and Requirements, presentation by Swaran Lata of W3 India
- Internet in India 2013 Report
- Language and Translation Industry of India: A Historical and Cultural Perspective, presentation by Ravi Kumar at XVIII FIT World Congress 2008, Shanghai, China
- Languages of the Indian Subcontinent Map, by Dr. Stephen Huffman
- National Translation Mission
- Technology Development for Indian Languages (TDIL)
- Training needs for the Localization Industry in India, TechLink report
- Web for All – Indic languages perspective, presentation by Manish Bhargava, Google Inc. at W3C Conference, New Delhi - May 6th, 2010
- Wikipedia: Languages of India