NLP For Indian Languages

Niyati Bafna, May 2023

A broad categorisation of languages in India based on their status and needs

Although such a categorisation naturally has fuzzy boundaries, we can try to list groups of languages on the basis of their existing vitality and functional load, the grassroots drive for their maintenance and promotion, as well as their vulnerability in the context of the country's constitution and governance. Currently, there are 22 scheduled languages of India, and the census recognizes about 92 unscheduled languages.

0. English: Numerical minority, but high transparency in the legal, administrative (between states), media, science and technology, and medical domains.
1. Hindi: High functional load in public civic life in large swathes of North India, official language in 8 states. Attrition in scientific and technical domains.
2. Scheduled languages I: Languages such as Tamil, Marathi, Gujarati, Bengali, Telugu, Kannada, Malayalam. State funds available for their promotion, existing film and music industries, used widely in their kin state in public life. Target of schemes such as Bhashini, and most likely to benefit from them. Such languages also are relatively high-resourced, and have considerable public backing for their maintenance.
3. Scheduled languages II: Languages such as Assamese, Punjabi, Odia, Konkani, Sindhi, Nepali. These languages, while they have official status, are somewhat lower-resourced than the above languages, and are spoken in smaller states. However, they generally have a community and/or state enthusiastic about their promotion. They are also the target of CIIL intervention and Bhashini-type schemes, and would benefit from a more expansive web resources, better technologies (typing, printing, etc.)
4. Scheduled languages III: Languages such as Metei, Maithili, Bodo, Dogri, Santhali, Kashmiri. Although they have official status (some of them recently so), they are much lower-resourced than other scheduled languages, and are 'trailing' targets of government schemes and funding. However, most of them are backed by a speaker base that is enthusiastic about their maintenance (perhaps with the exception of Kashmiri and Dogri). Although Khasi is not a scheduled language of India, it has official status in Meghalaya, although a lack of both grassroots activism and government attention.
5. Non-scheduled languages with a reasonably big native speaker base: Languages such as Bhojpuri, Bhili, Rajasthani languages, Tulu, Gondi. These languages are natively spoken by a reasonably large community, but lack official status, and have varying degrees of grassroots activism and research interest, if any.
6. Vulnerable non-scheduled languages: All other languages, most of them tribal languages, such as Kudukh, Ao, Angami, and others, especially those spoken in the central parts of India (Jharkhand, Madhya Pradesh), as opposed to North-Eastern tribal languages. These languages are facing attrition and language death even in the home domain, because of reducing language loyalties, and the incentive to shift to more dominant regional languages and/or Hindi and English.

Note that 97% of the Indian population records one of the scheduled language as their 'mother-tongue', or speaks at least scheduled language. However, dialectical varieties of these languages are often ignored, and the speaker base specifically of Hindi (40% of population) is inflated for the political Hindi agenda. Therefore, there is an additional category that spans C1-5, namely, the dialect continua associated with certain languages of those listed. This is a related concern with which to approach NLP for C1-7 languages, apart from building support for their standardized variants.

What are the broad domains of target problems for these languages?

Although as mentioned, different groups of languages have different sets of problems, here are some of the general potential areas for further development. * indicates (clearly) out-of-scope for the purposes of NLP.

1. Educational materials and resources. Facilitation of mother-tongue education via bilingual and monolingual resources. *Original resources.
2. Science and technology. Coining and integration of technical terminology. Translation of technical textbooks.
3. Other professional spheres: medicine, law, perhaps business. Same as (2).
4. Internet. Making the web available in different languages, including translation from other languages as well as *initiatives for contribution.
5. Administration. Translation tools for minority languages, better OCR.
6. ASR, speech bots, chatbots. Handling dialectical variations (at least understanding if not responding in), code-switching, and technical terminology.
7. Script standardization or technologies to handle script variation. Useful for certain C4-6 languages with a strong oral tradition but without a formal script, or without standard orthography. This is especially relevant when the lack of script and standardization is used as a rationale for not funding education or further resources.
8. Typing technology in Latin script. Better autocomplete/correct for Indian languages typed in the Latin script, code-switching, Indian named entities.
9. Typing technology in associated script. Well-designed easy-to-use keyboards, autocomplete/correct.
10. Improving general NLP applications such as search/IR, QA, language modelling, machine translation, text generation, etc.

Some broad problems to be solved emerge.

We need good **translation** tools *into* minority or regional languages (C2-6) from link or official languages (C0-2). This entails solving many subproblems, primarily the translation of technical terminology, cross-domain generalization, etc.

Speech tools are also useful to be integrated into other technologies such as chat assistants and search, and can have wide applications ranging from public administration, general

search and QA technologies, assistive technologies in regions with high illiteracy, as well educational tools to mitigate illiteracy.

Then, there is a set of **other basic software** such as better typing interfaces, and OCR, including applications of language modelling like autocomplete/correct.

Finally, there is **assisting at the level of grassroots activism**, involving help with actually scripting and evolving a standard variant of a language, coming up with new terminology, as well as responding to the specific needs of the concerned community.

Many of the above NLP problems will be jointly benefited by certain shared research goals, such as developing *better language modelling*. These problems also all need to be solved in *low-resource settings*, although, of course, C0-6 languages lie on a spectrum of low-resourcedness, roughly in order from higher to lower.

There is also an orthogonal dimension to the above problems, i.e. *better handling code-switching*, which is a highly relevant issue. Code-switching can be addressed as a sub-problem of any of the above general problems.

Discussing language in education

→Background

In theory, according to the 1968 National Policy Resolution, Indian schools are supposed to implement the Three Language Formula (TLF) in education. This refers to the medium of education or languages taken as subjects, and stipulates the following that the state language, Hindi, and English be taught as first, second, and third languages. In the case that the state language is Hindi, a third modern Indian language, preferably a Dravidian/South Indian language, should be taught instead.

A (relative or absolute) minority community has the right to open schools with their minority language as the medium of education, although they are not allowed to deny admission to people outside their community as per discrimination laws.

The NEP 2020, applicable to both public and private schools, reinforces the TLF policy and tried (and failed) to make Hindi mandatory in all schools. It also recommended education in the mother tongue language, usually the state language.

In practice, the most preferable medium of education in Indian primary and secondary schools is English, since it is perceived as the language of upward mobility. According to the UDISE+ report[1] in 2021, 42% children study in Hindi medium schools, 26% in English medium schools, 6% in Bengali, 5% in Marathi, and 2% in Tamil. Karnataka and West Bengal are the only two states where a considerable percentage of students have a preference for studying in the state language as opposed to English, although the number of students

[1] https://www.ndtv.com/education/udise-report-more-42-children-study-in-hindi-over-26-in-english-2478232

studying in English medium schools is steady increasing.[2] A negligible number of students study in Bishnupriya Manipuri and Dogri medium schools. On the other hand, there is an initiative to introduce education in mother tongue tribal languages in Jharkhand, including Santali, Mundari, Ho, and Kudukh,[3] and there are several hundred Bodo-medium schools in Assam.[4]

Further, the above figures refer to the medium of instruction in primary education; the percentage of English-medium schools is drastically higher for upper primary (middle) schools and especially for secondary schooling.[5]

Schools in minority and/or tribal languages generally aim to use the minority language as a transition into using either the state language or Hindi/English as a full medium of instruction, i.e. they aim to phase out the usage of the minority language gradually over the later years of primary school (in the second or third standard). The idea is to prevent early drop-outs at a pre-primary or primary stage due to the well-studied negative impacts of education in a non-mother-tongue language, especially in already vulnerable communities with high illiteracy.

→ What are the needs of the new education policies, and of students?

(My understanding and opinion from MLII[i] and contemporary articles and news)

There is a clear shift towards English especially in middle and upper class families,[6] and there are vanishingly few (if any) degrees offered in state languages (other than degrees in e.g. literatures in regional languages). There is an ongoing national debate about the best policies to pursue regarding medium of education at each level of education.

The primary argument in favour of introducing education at all levels, including university degrees, in regional languages, is to make education more accessible to those students from non-privileged backgrounds, whose families probably cannot afford an English medium school. English is increasingly a barrier to entry to any kind of professional or corporate job, or an indispensable 'skill'. Proponents of regional language education at all levels present it as a counter to this problem.

On the other hand, opponents and critics of this idea (might) argue that India needs a link language/languages to enable national as well as global integration of regional economies and future job seekers. India is vastly multilingual; a degree in a regional language would effectively not only restrict its holder to her home state, but also hinder relevant professional communication and collaboration with across states (or even internationally). This is

---

[2] https://timesofindia.indiatimes.com/india/50-of-children-complete-class-x-in-vernacular-nso/articleshow/78049695.cms
[3] https://www.telegraphindia.com/jharkhand/tribal-languages-at-centre-of-jharkhands-new-teaching-model/cid/1832200
[4] https://www.sentinelassam.com/topheadlines/bodo-medium-notified-for-higher-secondary-level-616069
[5] https://timesofindia.indiatimes.com/india/50-of-children-complete-class-x-in-vernacular-nso/articleshow/78049695.cms
[6] There's a movie about this: https://en.wikipedia.org/wiki/Hindi_Medium

especially relevant for degrees in subjects such as law, medicine, and STEM, where professionals of these fields eventually use, contribute to, and participate in, national or international systems of databases, textbooks/materials, research, conferences, and lectures. In the case of entrepreneurship and IT, people similarly benefit from relocating (if necessary) to national technology hubs, and potentially hiring non-local employees from all over the country.

Both sides assume implicitly (and correctly) that cheap, quick, and high quality automatic translation is not available as an easy solve to this problem. Note that this would have to include translation for text as well as speech (audio + real time subtitling, for, say, lectures at an in-person conference), for all of these domains and their technical terminology. Not only would such systems have to theoretically exist (as the output of research), but they would have to be implemented and scaled into cheap, easy-to-use and widely available products, affordable for each event at each gathering, translating from the source language to, theoretically, at least the 22 scheduled languages of India (excluding Sanskrit). Machine translation, therefore, currently holds no place in this debate.

Until this MT utopia has been achieved, scaled, and productionized, it seems inevitable for people studying certain subjects, or with potential aspirations outside the borders of their state, to have to learn English at some point. This inevitability is more significant the smaller the economy or native speaker base of the regional language of the state under question (leaving aside the question of minority languages within a state).[7]

However, this still leaves several potential points of intervention for MT and other NLP applications in the current situation, that can alleviate problems people face due to the English barrier.

→ Mother-tongue education

MTE has been shown to be, theoretically, an effective scheme for preventing early drop-outs from primary school. As mentioned, schools implementing MTE also phase in state and national link languages gradually, to prepare students for higher education and opportunities.

In practice, MTE faces several problems in India. Many of these problems are intertwined with the general sociopolitics of marginalized communities and in general out of scope for NLP solutions, but one pressing matter is the lack of educational materials in the target language, often cited as a reason not to fund or recognize such schools.

Textbook translation for primary schooling, therefore, is likely to be helpful to language communities who are looking to start schools in their languages, or communities that have already done so for updating and maintaining relevant materials. MLII mentions that another

[7] However, note that the *desire* and demand to study in the state official language may be influenced by other factors than its vitality in professional spheres, such as the strength of the 'sub-nationalist' movement within that state or community. We see this, for example, with the Assamese, Bodo and Santali movements.

problem, i.e., for communities with significantly distinct cultures from the mainstream, primary education needs to engage with culturally relevant themes, vocabulary, and socioecological context. While this work is of course best done under the supervision of native speakers and people who culturally identify with the language community under question, it is possible that **'creative' translation models**, that accept natural language instructions and explanations,[8] can aid in the process of translation with a cultural shift.

→ What about education at higher levels?

While the share of English in workplaces and academia in India is certainly increasing, increasing the pressure for people seeking the concerned jobs to *eventually* learn English, it does not follow that all university degrees, lectures, and materials, must be in English from the get-go, with no support for students who might struggle. At the same time, as discussed, it may be important for students to learn the technical parlance of their chosen subject in English, and people in India today clearly show the desire to be fluent in English as a marketable skill.

A possible response to this problem may be **MT with the special treatment of technical terms**. This might mean different things based on the needs of the user – the user might want to:

1. Only highlight technical terms and leave the text as is (in English)
2. Translate the text to the user's language, but maintain English technical terms via lexical and phrasal code-switching, so that the user can learn their meaning by exposure. Such code-switching is extremely common in technical conversations among students as well as existing educational materials, tutorials, and videos.
3. Offer translations for technical terms in the target language, and leave the text as is.
4. Translate everything.

In essence, the user may choose all combinations of translating/not translating x text, technical terms, including functionalities for highlighting, and/or translations only for hovered-on words, etc. Note that such a system deals closely in **controllable code-switching**, and will face all the entailed problems, such as maintaining grammaticality under the language-specific morphosyntactic constraints of code-switching with English.

The user can use such technology to comfortably consume educational materials while simultaneously learning English parlance in her field (and presumably, improving general English skills using other facilities depending on individual needs).

Also note that having flexible technologies such as the above will allow people to control the extent to and pace with which they learn English. Vocational jobs, for example, in a given state, or courses in education to be, say, a public school teacher in the state language, or other such subjects of study may not really require fluency in English beyond comprehension of

[8] Existing generative tools, like ChatGPT, already show the ability to modify their output according to specific instructions by the user, who can use the iterative collaborative process to create the required output.

some jargon, and such professionals may not have the desire to learn English (and should have the option not to).

Discussing Public Administration

→ Background

The language of public administration, including official notices, court language, signs and notice boards, complaints and responses, and press communications, has historically been a controversial topic in India. After independence, it was stipulated that Hindi and English be used for the above purposes, including communication with any state of India from the center, or between any two states, and that after 15 years, Hindi become the sole language of the Union. However, when the time came, there were protests and demonstrations against Hindi imposition especially in the Dravidian-language speaking states, and the Official Languages Act 1963 (amended 1967) allowed the continued use of English as it was being used before, also establishing rules for state languages. Namely, each state can choose its own official language(s) and conduct its internal administration in it, but both the center and any state are required to provide an English translation in communication with any state that does not have Hindi as one of its official languages. The Languages Act also allows the optional use of Hindi or the state official languages in High Courts with the permission of the Chief Justice of India,[9] provided that it is accompanied by an English translation, and stipulates a Hindi translation of bills and acts at the center.

In general, states use their official language(s) for low-level administration at the district and taluka level; internal communication is performed in state official languages to a varying extent, and any communication with the center, or between two states where at least one is non-Hindi-speaking, is accompanied with an English translation.

Note that all of the above generally concerns C0-3 languages, and that there is very little support as yet for the rest. However, by law, any district where a (relative or absolute) minority language (i.e. different from the official languages of the state) is spoken by more than 60% of the population, can be administered in that language, including having official notices and reports, and accepting complaints and documents in that language. In the case where such population comprises at least 15% of the population, the officials are required to interact with speakers in their language. This is the case, for example, for Tamil-speaking regions of Karnataka, or Bengali-speaking regions of Assam. It is also applicable to absolute minorities such as C5-6 languages, in tribal regions of Jharkhand or the north-eastern states, although data on the languages of governance at the district is hard to find.

→ Moving forwards?

In general, the main need of smoother multilingual administration in the country is speedy and portable translation services that works well across several domains. This is most relevant where the administrative language is not spoken by the person seeking to interact

[9] Currently, Hindi, Tamil, Gujarati, Bengali, and Kannada, are authorized to be used as High Court languages in respective states, and there are demands to add Marathi to that list.

with the police, district offices, etc. This is not uncommon even in states with a strong language identity; e.g. only about 40% of the population in Bombay speaks Marathi.[10]

Government schemes for handling linguistic diversity

Central Institute for Indian Languages: The CIIL is the overarching institute that manages and monitors the following schemes. It collaborates with the center and state governments in their language-related initiatives, and also manages its own projects in documentation, preservation, and new research and technologies. It maintains a catalogue of books available or created in language learning and/or as a result of said projects. See a complete list of projects under the CIIL. A few of them are introduced below.

Linguistic Data Consortium for Indian Languages: The LDC-IL (managed by CIIL) is concerned with collection of corpora of different kinds, raw and annotated, for the scheduled languages of India, possibly expanding to non-scheduled languages as well. So far, they have collected raw text corpora, monolingual and comparable, speech corpora, as well as ISL corpora (listed here). However, these resources are not freely available to the public, and require a membership fee.

National Translation Mission: The NTM is mainly concerned with translating texts into scheduled languages of India, and as per 2023 has published 153 textbooks in linguistics, history, politics, and other subjects, and 72 e-books (listed here).

Bhashini was launched in 2020 as part of the NTM, dealing with making the internet available in 12 Indian languages. The scheme encompasses encouraging research in NLP technologies for these languages, funding start-ups and industries to deploy developed technologies, and crowdsourcing for data (although so far, less than 1000 people have contributed in the Bhasha Daan initiative).

Scheme for Protection and Preservation of Endangered Languages: The SPPEL, managed by the CIIL, lists 117 endangered languages of India, many with less than 10000 speakers, which is the minimum number of speakers required for a language for the census to report its existence. The scheme deals mainly with documentation and building dictionaries and writing grammars for these languages. The SPPEL will potentially handle another 500 languages in the future. So far, it has built dictionaries for 12 languages against Hindi and English, with about 1500-2000 words per language on average. There is ongoing work in about 3-4 languages.

There are different centres and offices of the CIIL in India, dealing in regional languages and problems. Other notable schemes include Bharatavani, which makes content in Indian languages as well as state language textbooks available, and Orthographic Development for Languages of North East India, which among other questions looks at how to create a script for an unwritten language.

[10] However, a much higher percentage speaks either Hindi, Marathi, or Gujarati.

[i] MLII: *Minority Languages in India,* Thomas Benedikter, 2004