Indian Subcontinent Language Vitalization

András Kornai, Pushpak Bhattacharyya

Department of Computer Science and Engineering, Department of Algebra
Indian Institute of Technology, Budapest Institute of Technology
kornai@math.bme.hu, pb@cse.iitb.ac.in

Abstract

We describe the planned Indian Subcontinent Language Vitalization (ISLV) project, which aims at turning as many languages and dialects of the subcontinent into digitally viable languages as feasible.

digital vitality, language vitalization, Indian subcontinent

In this position paper we describe the planned Indian Subcontinent Language Vitalization (ISLV) project. In Section 1 we provide the rationale why such a project is called for and some background on the language situation on the subcontinent. Sections 2-5 describe the main phases of the planned project: Survey, Triage, Build, and Apply, offering some preliminary estimates of the difficulties at each phase.

1. Background

The linguistic diversity of the Indian Subcontinent is remarkable, and in what follows we include here not just the Indo-Aryan family, but all other families like Dravidian and individual languages spoken in the broad geographic area, ranging from Kannada and Telugu with tens of millions of speakers to the languages of scheduled tribes which may be spoken by only a few hundred people. We define the Subcontinent broadly, so as to include not just India, Pakistan, and Bangladesh, but also Nepal, Bhutan, Sri Lanka, Afghanistan, and the Maldives, because the languages spoken in this geographic area often form cross-border continua. The simple question of exactly how many languages/dialects we need to consider is already fraught with difficulty, with estimates ranging from over 1,600 in the 1961 Census, see http://www.languageinindia.com/aug2002/indianmothertongues1961aug2002.html, to less than 500 in the *Ethnologue* (Lewis et al. 2013).

Kornai (2013) divided languages in four major categories: digitally Thriving, Vital, Heritage, and Still. Without prejudging matters, it is clear that on the subcontinent all four possibilities obtain: English is thriving, Hindi is vital, Sanskrit is heritage, and Bagata (the language of a scheduled tribe in Andhra Pradesh, not even listed in the Ethnologue) is still. ISLV puts the emphasis on the borderline cases between digitally viable (T and V) languages on the one hand, and digitally dead (H and S) languages on the other. The goal is not just to enhance scholarly knowledge in this area, but also to inform decisionmakers where the limited resources available to language vitalization are best applied. This requires not just a detailed *survey* of the languages in question (see Section 2) but also an objective *triage* mechanism (see Section 3).

We will be paying considerably less attention to languages like English and Hindi that are thriving or nearly so, suggesting that efforts aimed at *building* language technology (see Section 4) are best concentrated on the less vital (but still vital or at the very least borderline) cases at the expense of the obviously moribund ones. To find this borderline we need to distinguish the heritage class of languages, typically understood only by priests and scholars, from the still class, which is understood by native speakers from all walks of life. For heritage language like Sanskrit considerable digital resources already exist, both in terms of online available material (in translations as well as in the original) and in terms of lexicographical and grammatical resources of which we single out the Köln Sanskrit Lexicon at http://www.sanskrit-lexicon.uni-koeln.de/monier and the INRIA Sanskrit Heritage site at http://sanskrit.inria.fr. For still languages, there is practically nothing, and conservation efforts are very justified.

We emphasize at the outset that we do not advocate the wholesale abandonment of still languages. Unlike in a field hospital, where triage really means the abandonment of the likely fatally wounded so that those who can still be saved get a better chance, here still languages can receive a different kind of treatment, heritage preservation. This is a very worthy goal, and there are already significant societal efforts in this direction such as the Endangered Languages Project at http://www.endangeredlanguages.com. This should be kept in mind as we *apply* our findings, especially as the preservation effort is in a substantively different direction, requiring very different resources, than vitalization proper. As we shall see, preservation is primarily the work of anthropologists and linguists trained in fieldwork, while digital vitalization requires machine learning techniques.

2. Survey

The purpose of the first stage of ISLV is to collect a broad range of facts and opinion that covers not just branches of Indic in the strict sense but also languages and cultures deeply influenced by Indic vocabulary and script on the subcontinent. We use the directed crawling technique described in Zséder et al. (2012) to collect as much data for each dialect as possible. An important intermediate result of this stage is the development of robust dialect-identification models along the lines of the well known `TextCat` (see

http://odur.let.rug.nl/~vannoord/TextCat) and CLD2 (see https://code.google.com/p/cld2) models, taking into account various encodings ranging from legacy schemes such as ISCII to varieties of Unicode and even latinized writing (still common in text messaging) and scholarly systems such as IPA.

At the current stage, our database covers 634 languages and dialects of the subcontinent, excluding English. Table 1 gives the breakdown per primary country.

Country	Language
Afghanistan	30
Bangladesh	16
Bhutan	20
India	397
Maldives	1
Nepal	107
Pakistan	59
Sri Lanka	4

Table 1: Current coverage

It is, of course, quite debatable whether languages of Nepal or Afghanistan should all be all included, and we welcome any cogent argument in this regard, especially as it is unclear whether the same funding sources that support efforts in India would be equally available in other countries. That said, we lean toward inclusion, rather than exclusion.

3. Triage

Once the data is collected, we apply the methodology of Kornai (2013) to decide which varieties can be classed as vital, heritage, or still. There are no digitally thriving languages as defined originally, though we acknowledge that Hindi may be classified as such. Table 2 summarizes the breakdown is as follows:

Status	Language
vital	36
borderline	21
heritage	1
still	576

Table 2: Main classes

Only one language, Pali, is listed as heritage, since a considerable number of people (over 10,000) are listed as native (L1) speakers of Sanskrit. Be it as it may, the data is dominated by digitally still languages (over 90% of the languages considered, see the Appendix), and we are left with some 50-60 languages that have a chance to take root digitally. It should be noted that our classification of vital vs. borderline was already highly optimistic (see Kornai 2013 for a more detailed description of the conservative methodology chosen so as not to raise false alarms), with languages like Dogri (dgo) listed as digitally vital, which is quite debatable.

In the full study, we are likely to retain the positive outlook that characterized the earlier work, so as not to hinder the digital ascent of any language that has a fighting chance. With five million speakers, Dogri may very well not be a lost cause. A good first step toward demonstrating the vitality of a language could be the collection of a BLARK (Krauwer 2003).

Since these lists are so small, we provide them in full below, but with the clear understanding that *our results are preliminary*, and more sophisticated data gathering in Phase 1 may still change them.

Vital: Angika, Assamese, Bengali, Bishnupriya, Brahui, Chakma, Dogri, Maldivian, Dzongkha, Gujari, Goan Konkani, Gujarati, Hindi, Kannada, Kashmiri, Kachchi, Khasi, Khowar, Lushai, Maithili, Malayalam, Marathi, Nepali, Newari, Oriya, Western Panjabi, Dari, Rangpuri, Sanskrit, Sinhala, Seraiki, Sindhi, Tamil, Tulu, Telugu, Urdu.

Borderline: Awadhi, Baluchi, Southern Balochi, Badaga, Bhojpuri, Halbi, Chhattisgarhi, Kukna, Konkani, Manipuri, Naga Pidgin, Ao Naga, Adivasi Oriya, Punjabi, Southern Pashto, Pashto, Rajasthani, Santali, Saurashtra, Sylheti, Kok Borok.

We welcome criticisms both of the data (if a language of the subcontinent that you are working with does not appear on any of the above lists nor in the Appendix this obviously points at an error in the data gathering process) and of the classification. We are particularly interested in cases where languages should evidently be moved from the still to the vital category, or the other way round. Again, we emphasize that these results are preliminary, and we welcome scholarly debate and discussion.

The ISLV plan is to make the final results, and the data it is based on, publicly accessible, either hosted directly at a dedicated website or by means of pointing to or replicating data already available elsewhere.

We should add here that a policy recommendation based on an assessment of digital death should go beyond a simple exhortation to concentrate all effort regarding this language on heritage conservation. As we already noted in Kornai (2013), such efforts, while obviously necessary for preserving the cultural heritage of humankind, contribute practically nothing to language vitality. To mitigate the human cost of digital language death it is therefore suggested that we expend effort on identifying, if at all feasible, for each digitally still language a vital 'champion' of similar vocabulary, script, and grammar, with the idea that this champion can become a medium of access to the digital realm that is easier to acquire than English.

In certain cases, such as Andaman Creole (hca), bilingualism is so strong that the choice of the champion is obvious, but in many cases the task is far from trivial. This effort, which should be undertaken primarily by scholars intimately familiar with the language and the sociolinguistic/dialect situation, requires only user-level knowledge of language technology. This is very different from our recommendations for vital/borderline languages, to which we turn now.

4. Build

The build stage no longer considers all dialects and languages of the Indian Subcontinent, just those deemed vi-

tal/thriving in Stage 2. These we can hope to endow with a full computational toolchain composed of the following stages.

Tool	Effort
script	0.1
normalization	1
language detector	1
word list	2
bilingual dictionary	6
morphology	12
spellchecker	6

Table 3: Word-level tools

Table 3 lists the effort (in person-months) associated with building the tool or resource in question *after* the data-gathering phase is complete, but assuming significant online material is found, for if there is no material online the digital vitality of the language is in grave doubt. It is not assumed that the tools so obtained will be of quality comparable to those available for English (or for MT, say English-French). Nevertheless, such tools are already useful for a broad variety of purposes, and their incremental refinement and they higher level tools built on top of them are left to the last Phase of the ISLV project (see Section 5 below).

Special funds may be obtained from the NSF Documenting Endangered Languages program, the Endangered Languages Project, or other similar preservation efforts, but the only direct contribution of ISLV in this regard would be the recommendation that these dialects are indeed in need of such an effort. The main focus of building would initially be on the word-level technologies, including spellchecking (standardized orthography), stemming (prefix- and suffix-removal, but not necessarily deeper morphological analysis), glyph analysis, and building a common multilingual dictionary of basic vocabulary similar to Ács et al (2013). Such efforts obviously require better, engineer-level understanding of language technology, and may serve as a training ground for a new generation of native computational linguists.

It is the task of this phase to determine to what extent standard (two-tape) finite-state transducer technology is usable for providing cross-transliteration among the vital varieties on the one hand, and between the vital champions and their satellites on the other. Only a limited amount of parallel (synchronized) grammar writing is envisioned at this stage.

5. Apply

The main applications we envision extend the text and image-based work to speech (and if resources permit, to sign languages). Of particular interest are cross-language speech translators which do not assume users to have great familiarity with standard Hindi, text to speech systems capable of synthesizing speech in any vital language, and perhaps captioning systems that would extend the reach of broadcast operations. As can be seen from the following Table 4, the effort of building these tools is considerably larger, and it is only with computational linguists who are both native speakers and already skilled in the design and application of the word-level tools that most of these can be attempted.

Tool	Effort
light parser	12
NER	6
OCR	12
ASR	12
MT	12

Table 4: Higher tools

For a significant subset (over half) of the thriving/vital languages, in particular Assamese, Bengali, Bodo, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Panjabi, Sanskrit, Tamil, Telugu, and Urdu, there is already a concentrated effort under way, see the IndoWordNet site at http://www.cfilt.iitb.ac.in/indowordnet, and that a practical application with font transcoding and other critical parallelization technology is available in a 5-way parallel tourism site (Bengali, Hindi, Marathi, Tamil, Telugu), see http://www.tdil-dc.in/sandhan. The build phase can increment these systems for the languages not yet considered.

It is expected that orphaned dialects without a near champion can only be documented in the sense of heritage preservation, while dialects close to champions will participate in (sub)koine formation. The beneficial effects may even extend to some of the languages and dialects outside the Subcontinent. It will require a great deal of care to select the champion dialects, a matter of particular importance in Hungary, where several Roma dialects, some obviously close to main Indian languages, some less visibly so, are spoken. It is not expected that such dialects would be vital in and of themselves, but finding a champion they could attach to would significantly enhance their chances of digital survival.

6. Acknowledgement

Work supported in part by the European Union and the European Social Fund through project FuturICT.hu grant #TÁMOP-4.2.2.C-11/1/KONV-2012-0013.

7. References

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August. Association for Computational Linguistics.

András Kornai. 2013. Digital language death. *PloS one*, 8(10):DOI 10.1371/journal.pone.0077056.

Paul Lewis, Gary Simons, and Charles Fennig, editors. 2013. *The Ethnologue*. Summer Institute of Linguistics.

Attila Zséder, Gábor Recski, Dániel Varga, and András Kornai. 2012. Rapid creation of large-scale corpora and frequency dictionaries. In *Proceedings to LREC 2012*, pages 1462–1465.

Appendix: digitally still languages

A'tong, A-Pucikwar, Adap, Adi, Adiwasi Garasia, Aer, Afghan Sign Language, Agariya, Ahirani, Ahom, Aimaq, Aimol, Aiton, Aka-Bea, Aka-Bo, Aka-Cari, Aka-Jeru, Aka-Kede, Aka-Kol, Aka-Kora, Akar-Bale, Allar, Alu Kurumba, Amri Karbi, Anal, Andaman Creole Hindi, Andh, Angami Naga, Ange, Apatani, Aranadan, Ashkun, Asuri, Athpariya, Attapady Kurumba, Badeshi, Bagheli, Bagri, Bahing, Balti, Bantawa, Baraamu, Bateri, Bauria, Bawm Chin, Bazigar, Belhariya, Bellari, Betta Kurumba, Bhadrawahi, Bhalay, Bharia, Bhatola, Bhatri, Bhattiyali, Bhaya, Bhilali, Bhili, Bhoti Kinnauri, Bhujel, Bhunjia, Biete, Bijori, Bilaspuri, Birhor, Bodo Gadaba, Bodo Parja, Bodo, Bondo, Bote-Majhi, Braj, Brokkat, Brokpake, Brokskat, Bugun, Buksa, Bumthangkha, Bundeli, Burushaski, Byangsi, Camling, Car Nicobarese, Central Nicobarese, Central Pashto, Chalikha, Chamari, Chambeali, Chang Naga, Changthang, Chantyal, Chaudangsi, Chaura, Chenchu, Chepang, Chhintange, Chhulung, Chilisso, Chinali, Chiru, Chitkuli Kinnauri, Chittagonian, Chitwania Tharu, Chocangacakha, Chodri, Chokri Naga, Chothe Naga, Chug, Chukwa, Churahi, Dakpakha, Dameli, Dandami Maria, Dangaura Tharu, Darai, Darlong, Darmiya, Deccan, Degaru, Dehwari, Deori, Desiya, Dhanki, Dhanwar, Dhatki, Dhimal, Dhodia, Dhundari, Digaro-Mishmi, Dimasa, Dolpo, Domaaki, Dotyali, Dubli, Dumi, Dungmali, Dungra Bhil, Dura, Duruwa, Dzalakha, Eastern Balochi, Eastern Gorkha Tamang, Eastern Gurung, Eastern Magar, Eastern Meohang, Eastern Muria, Eastern Parbate Kham, Eastern Tamang, Eravallan, Far Western Muria, Gadaba, Gadaba, Gaddi, Gade Lohar, Gahri, Galo, Gamale Kham, Gamit, Gangte, Garhwali, Garo, Gata', Gawar-Bati, Ghandruk Sign Language, Ghera, Goaria, Godwari, Gondi, Gongduk, Gowlan, Gowli, Gowro, Grangali, Gurgula, Hajong, Harijan Kinnauri, Haroti, Haryanvi, Hazaragi, Helambu Sherpa, Hinduri, Hmar, Ho, Holiya, Hrangkhol, Hruso, Humla, Idu-Mishmi, Indian Sign Language, Indo-Portuguese, Indus Kohistani, Inpui Naga, Irula, Ishkashimi, Jad, Jadgali, Jandavra, Jangshung, Jarawa, Jatapu, Jaunsari, Jennu Kurumba, Jerung, Jhankot Sign Language, Jirel, Juang, Jumla Sign Language, Jumli, Juray, Kabutra, Kachari, Kachi Koli, Kadar, Kagate, Kaikadi, Kaike, Kalaktang Monpa, Kalami, Kalanadi, Kalasha, Kalkoti, Kamar, Kamviri, Kanashi, Kanauji, Kangri, Kanikkaran, Kanjari, Kannada Kurumba, Karbi, Kathoriya Tharu, Kati, Katkari, Kayort, Khaling, Khamba, Khamyang, Khandesi, Kharam Naga, Kharia Thar, Kharia, Khengkha, Khetrani, Khezha Naga, Khiamniungan Naga, Khirwar, Khoibu Naga, Kinnauri, Koch, Kochila Tharu, Koda, Kodaku, Kodava, Kohistani Shina, Koi, Koireng, Kol, Kom, Konyak Naga, Korku, Korlai Creole Portuguese, Koro, Korra Koraga, Korwa, Kota, Koya, Kudiya, Kudmali, Kui, Kullu Pahari, Kulung, Kumaoni, Kumarbhag Paharia, Kumbaran, Kumhali, Kundal Shahi, Kunduvadi, Kupia, Kurichiya, Kurmukar, Kurtokha, Kurukh, Kusunda, Kutang Ghale, Ladakhi, Lahnda, Lahul Lohar, Lakha, Lambadi, Lambichhong, Lamkang, Lasi, Layakha, Lepcha, Lhokpu, Lhomi, Liangmai Naga, Limbu, Lingkhim, Lish, Loarki, Lodhi, Lohorung, Loke, Lotha Naga, Lui, Lumba-Yakkha, Lunanakha, Lyngngam, Magahi, Mahali, Mahasu Pahari, Majhi, Majhwar, Mal Paharia, Mala Malasar, Malankuravan, Malapandaram, Malaryan, Malavedan, Malvi, Manangba, Manda, Mandeali, Manna-Dora, Mannan, Mao Naga, Mara Chin, Maram Naga, Maria, Maring Naga, Marma, Marwari, Marwari, Marwari, Mawchi, Megam, Memoni, Merwari, Mewari, Mewati, Miju-Mishmi, Mina, Mirgan, Mirpur Panjabi, Mising, Mixed Great Andamanese, Mogholi, Monsang Naga, Moyon Naga, Mru, Mudu Koraga, Muduga, Mugom, Mukha-Dora, Mullu Kurumba, Munda, Mundari, Munji, Musasa, Muthuvan, Mzieme Naga, Na, Naaba, Nachering, Nagarchal, Nahali, Nahari, Nar Phu, Nefamese, Nepalese Sign Language, Nepali Kurux, Nepali), Nihali, Nimadi, Nocte Naga, Noiri, Norra, Northeast Pashayi, Northern Ghale, Northern Gondi, Northern Hindko, Northern Pashto, Northern Rengma Naga, Northwest Pashayi, Northwestern Kolami, Northwestern Tamang, Nubri, Nupbikha, Nyenkha, Nyishi, Od, Oko-Juwoi, Olekha, Oraon Sadri, Ormuri, Pahari-Potwari, Pahlavani, Paite Chin, Pakistan Sign Language, Paliyan, Palpa, Palya Bareli, Panchpargania, Pangwali, Paniya, Pankhu, Pao, Parachi, Pardhan, Pardhi, Parenga, Parkari Koli, Parsi, Pathiya, Pattani, Pauri Bareli, Pengo, Phake, Phalura, Phangduwali, Phom Naga, Phudagi, Pnar, Pochuri Naga, Porja, Poumei Naga, Powari, Prasuni, Puimei Naga, Puma, Purik, Puroik, Purum Naga, Purum, Rabha, Rajbanshi, Raji, Rajput Garasia, Ralte, Rana Tharu, Rangkas, Ranglong, Rathawi, Rathwi Bareli, Raute, Ravula, Rawat, Reli, Riang, Rongmei Naga, Rongpo, Ruga, Saam, Sadri, Sajalong, Sakachep, Sambalpuri, Sampang, Samvedi, Sanglechi, Sangtam Naga, Sansi, Sartang, Sauria Paharia, Savara, Savi, Seke, Sentinel, Shekhawati, Shendu, Sherdukpen, Sherpa, Sheshi Kham, Shina, Sholaga, Shom Peng, Shumashti, Shumcho, Sikkimese, Simte, Sindhi Bhil, Singpho, Sirmauri, Sonha, Sora, Southeast Pashayi, Southeastern Kolami, Southern Ghale, Southern Gondi, Southern Hindko, Southern Nicobarese, Southern Rengma Naga, Southern Uzbek, Southern Yamphu, Southwest Pashayi, Southwestern Tamang, Spiti Bhoti, Sri Lankan Creole Malay, Sri Lankan Sign Language, Stod Bhoti, Sumi Naga, Sunam, Sunwar, Surgujia, Surjapuri, Tagin, Tangchangya, Tangkhul Naga, Tarao Naga, Tawang Monpa, Teressa, Thachanadan, Thado Chin, Thakali, Thangal Naga, Thangmi, Thudam, Thulung, Tichurong, Tilung, Tinani, Tippera, Tirahi, Tiwa, Toda, Torwali, Toto, Tregami, Tshangla, Tsum, Tukpa, Turi, Turung, Tutsa Naga, Ullatan, Urali, Ushojo, Usui, Vaagri Booli, Vaghri, Vaiphei, Varhadi-Nagpuri, Varli, Vasavi, Veddah, Vishavan, Waddar, Wadiyara Koli, Wagdi, Waigali, Wakhi, Waling, Walungge, Wambule, Wancho Naga, Waneci, War-Jaintia, Warduji, Wayanad Chetti, Wayu, Western Balochi, Western Gurung, Western Magar, Western Meohang, Western Muria, Western Parbate Kham, Western Tamang, Wotapuri-Katarqalai, Yakha, Yamphu, Yidgha, Zangskari, Zeme Naga.