***Dialectologia* 19** *(2017), 67-91.*
ISSN: 2013-2247

Received 7 December 2015.
Accepted 27 April 2016.

ON DOCUMENTING LOW RESOURCED INDIAN LANGUAGES INSIGHTS FROM KANAUJI SPEECH CORPUS

Pankaj DWIVEDI & Somdev KAR
Indian Institute of Technology Ropar*
pankajd@iitrpr.ac.in / somdev.kar@iitrpr.ac.in

Abstract

Well-designed and well-developed corpora can considerably be helpful in bridging the gap between theory and practice in language documentation and revitalization process, in building language technology applications, in testing language hypothesis and in numerous other important areas. Developing a corpus for an under-resourced or endangered language encounters several problems and issues. The present study starts with an overview of the role that corpora (speech corpora in particular) can play in language documentation and revitalization process. It then provides a brief account of the situation of endangered languages and corpora development efforts in India. Thereafter, it discusses the various issues involved in the construction of a speech corpus for low resourced languages. Insights are followed from speech database of Kanauji of Kanpur, an endangered variety of Western Hindi, spoken in Uttar Pradesh. Kanauji speech database is being developed at Indian Institute of Technology Ropar, Punjab.

Keywords

speech corpus, Kanauji, language documentation, endangered language, Western Hindi

DOCUMENTACIÓN DE OBSERVACIONES SOBRE LENGUAS HINDIS DE POCOS RECURSOS A PARTIR DE UN CORPUS ORAL DE KANAUJI

Resumen

Los corpus bien diseñados y bien desarrollados pueden ser considerablemente útiles para salvar la brecha entre la teoría y la práctica en la documentación de la lengua y los procesos de revitalización, en la

* Indian Institute Of Technology Ropar (IIT Ropar), Rupnagar 140001, Punjab, India.

construcción de aplicaciones de tecnología lingüística, en la prueba de hipótesis lingüísticas y en muchas otras áreas importantes. El desarrollo de un corpus para una lengua con pocos recursos o amenazada encuentra varios problemas. El presente estudio se inicia con una perspectiva general sobre el papel que los corpus (los corpus orales en particular) pueden desempeñar en la documentación del lenguaje y en los procesos de revitalización. A continuación, ofrece una breve reseña de la situación de las lenguas amenazadas y de los esfuerzos de desarrollo de corpus en la India. Posteriormente, se analizan las diversas cuestiones relacionadas con la construcción de un corpus oral para lenguas con pocos recursos. Las observaciones se extraen de la base de datos de Kanauji de Kanpur, una variedad amenazada del Hindi occidental, hablada en Uttar Pradesh. La base de datos del discurso de Kanauji se está desarrollando en el ndian Institute of Technology Ropar, de Punjab.

Palabras clave
corpus oral, lengua kanauji, documentación de la lengua, lenguas amenazadas, hindi occidental

1. Introduction

> The old quip attributed to Uriel Weinreich, that a language is a dialect with an army and a navy, is being replaced in these progressive days: a language is a dialect with a dictionary, grammar, parser and a multi-million-word corpus of texts –and they would better all be computer tractable. When you have got all of those, get yourself a speech database,[1] and your language will be poised to complete on terms of equality in the new Information Society (Nicholas Ostler *apud* Borin 2006).

What Nicholas Ostler meant here is that a language will not be able to survive in the today's world, the world which looks as if it has turned into a technological park, if it is not used in language technology applications. Mentioning the importance of speech database, Ostler says: "…parser and a multi-million-world corpus of texts [...] get yourself a speech database" (p. 317). Trosterud (2006: 293) mentions "Languages may

[1] In this paper the term 'speech corpus' refers to an organized collection of spoken data of a language in a way so that it is used as a base for the various purposes ranging from phonetic analysis to multipurpose speech applications. Hence, the corpus will further need to be customized depending upon the focus of inquiry. It should be taken as a step, which, if appropriately dealt with, can act as platform for product development. Therefore, the terms speech or spoken corpora, speech database and digital spoken archives are more or less equivalent in nature.

live on without orthography. But no language will be able to function as administrative language in a modern society without a developed language technology applications".

First half of the 19th century saw a gradual shift from rule-based approach to corpus-based approach. Before it, a linguist's typical job involved to sit at a table and think of a language. Until 1950s it was realized that it is easier to formulate a hypothesis, compare results and perform a recheck if you have larger data at your disposal, which grew into concept what we today know as corpus linguistics. However, use of electronic corpora was not very clear in the picture till then. The first motivation in the field of electronic corpora came from the work of Jesuit priest Roberto Busa, who created an electronic lemmatised index of the complete works of St Thomas Aquinas, Index Thomisticus, beginning in the 1950s and completing it in the late 1970s (Tognini-Bonelli 2010). Since then both electronic text and speech corpora came to be used widely in all areas of language research including language documentation.

Language Corpora, including archives, can play a significant role in the documentation of endangered languages in following ways: a) help to preserve the dying languages for the future generations; b) facilitate the use of primary materials, such as filed notes, audio and video recordings; and thereby helping language maintenance and revitalization; anthropological, typological, historical and comparative studies; c) provide a platform for the products and deliverables not only to the researchers but also to the communities speaking endangered languages.

2. Language documentation and corpus linguistics

Documentation of a language often ends with a lot of material as it aims to collect the examples of the full spectrum of language forms and uses that the language community employs (Johnson 2004). Till the first quarter of the 20th century, language documentation meant collection of data on paper and therefore each documentation project resulted into large bodies of texts, that is, text corpora. Task such as dictation, transcription, translations, elicitation, analysis, etc. were all done only on papers. Linguists and other language researchers such as anthropologists worked painstakingly to collect and preserve these collections of texts. However, lately, technology entered into the scene and revamped the entire process all through. It provided the documentary linguists with the useful tools and resources leading to better organization and long-term preservation of the language data apart from significantly reducing the manual efforts. It became possible to store the text data in electronic form using floppy disks, CD ROMs, cassettes etc. According to Bird & Simons (2003) the process of documenting and describing the world's language is undergoing radical transformation with the rapid uptake of new digital technologies for capture, storage, annotation and dissemination. With the further advancement in the technology, linguists could also make good quality recordings of the spoken from itself and store them very longer time and thereby making spoken corpus.

The best thing about a carefully designed corpus is that one can perform almost any kind of linguistics without the help of language consultants/informants or any linguistic fieldwork. Large collection of language data lets you formulate a potential hypothesis and later to cross verify it. Corpus linguistics as well as language documentation share many common points of interaction. For example, one of the major points of interaction is data collection. McEnery & Ostler (2000: 410) define corpus linguistics as a methodological toolkit dealing with construction and analysis of consistent collection of data, whereas Jonhson's (2004) take on language documentation is 'effort to produce permanent and reusable *collections of diverse linguistic data*'. Cox (2009) opines that corpus linguistics and language documentation complements each other in the following ways:

> For corpus linguistics, language documentation offers a diverse and well-catalogued data, and 'raw material' for corpus construction for under-represented languages-which is a standing challenge to corpus linguistics, whereas for language documentation, corpus linguistics offers a methodological perspective on the documentary record and a set of new tools for analysis. To language documentation corpus linguistics also offers another means of rendering the documentary record available, both to academic and non-academic communities (Cox 2009: 4).

3. Language endangerment

Language endangerment is a serious concern to which linguists and language planners have turned their attention in the last several decades. The famous linguist and professor David Crystal (see Jansen & Sørensen 2005)[2] says "When a language is lost, a vision to the world is lost". For a variety of reasons, speakers of many smaller, less dominant languages stop using their heritage language and begin using another. For example, parents begin to use only the second language with their children and gradually the intergenerational transmission of the heritage language is reduced and may even cease. As a consequence there may be no speakers who use the language as their first or primary language and eventually the language may no longer be used at all (Paul, Simons & Fennig 2013). Crystal (2000) claims that the rate of language disappearance is as high as two languages each month (cf. Allwood 2006). Larger languages are held responsible on the endangered situation for the other languages. Writing about as to how English is devouring other smaller languages, McWhorter (2009) writes:

> There are about 20 languages that are slowly eating up the other 6,000. That's essentially because of how England developed a global presence starting in the 1600s, and the language they happened to carry with them was English. What we're seeing is an increasingly Anglophone world and an increasingly oral, rather than written world. So many of the other languages are falling by the wayside that we may lose 90 percent of the languages we have now by the year 2100 (John McWhorter *apud* Stephenson 2013).

4. Language endangerment in India: an overview

Times of India (TOI, August 9, 2013), a reputed English newspaper, reads (via People Linguistic Survey of India (PLSI) conducted by Bhasha Research & Publication

[2] http://www.dfi.dk/Service/English/News-and-publications/FILM-Magazine/Artikler-fra-tidsskriftet-FILM/60/ Voices-of-the-World-Language-is-the-House-of-Being.aspx

Centre, Gujrat, India) that India currently speaks more than 880 languages. Up till 1961, about 1,100 languages would be spoken in India. Of these, 220 languages have already disappeared in the last 50 years, that is, we have lost about 20% of total languages spoken in India during the period. TOI further reports that more than 150 out of 880 languages currently spoken are going to become extinct in the next 50 years, that is, about 17.05% are going to die. The situation seems to be awfully alarming.

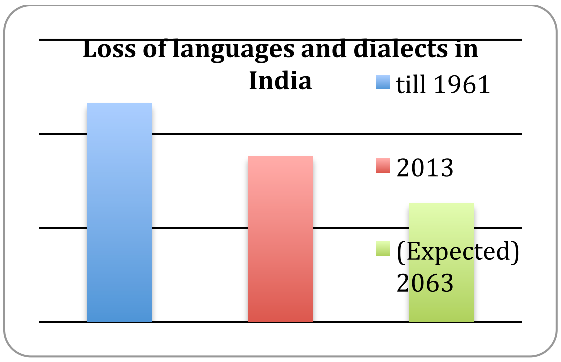


Figure 1. Loss of languages and dialects in India

According to 2001 census and language report,[3] 122 languages have been recognized as by the government. Of these 122 languages, 22 languages[4] have been classified as scheduled languages and 100 languages[5] have been categorized as non-scheduled languages. Statement 9 of the report reads that 99.82% of the population, i.e. 1,028,610,328, recognize their mother tongue among of these 122 languages. It further adds that only 0.17%, i.e. 1,762, 388, of the total population speaks other languages or dialects. These languages were not identifiable for they were returned by less than 10, 000 people at all India level as their mother tongue. If we assume that no language loss has occurred between 2001-2013, 768 languages were spoken by only 0.17 of the population of the India in 2001. And if we take PLSI statistics in account, about 32 languages are dying per decade in India. About 790 languages were spoken by merely

[3] Latest census was done in the year 2011 by the govt. of India. However, the govt. released no data pertaining to the no. of language spoken in India. According to news and reports, govt. plans to conduct a separate linguistic survey of the country.

[4] In the eight schedule, following languages have been recognized as scheduled languages:- Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santhali, Sindhi, Tamil, Telugu and Urdu.

[5] List of non-scheduled languages can be retrieved from <http://censusindia.gov.in/> (accessed 9th August 2015).

0.17% population of the country in 2001, which is extremely alarming for linguistic diversity of a country such as India.

There are various interrelated factors which together lead to endangerment of a language. Ethnologue uses Expanded Graded Intergenerational Disruption Scale (hence EGDIS) to estimate the vitality of the languages in the countries where they are primarily spoken. For India, Ethnologue lists following languages in the category of 'Threatened' (group-1), shifting (group-2), Nearly Extinct (group-3), and Dormant (group-4) on EDGIS scale:

Group I: Aiton, Allar, Andh, Apatani, Aranadan,Bantawa, Byangsi, Chamling, Chaudangsi, Chin-Bawm, Darmiya, Dubli, Gadaba, Bodo, Gahri, Gata', Godwari, Gurung-Western, Hruso, Indo-Portuguese, Kachari, Kadar, *Kanauji,* Khamti, Koraga, Korra, Koraga-Mudu, Koro, Kui, Kulung, Kumbaran, Kupia, Kurichiya, Magar-Eastern, Mahali, Mal Paharia, Malavedan, Mannan, Nihali, Öñge, Pardhan, Powari, Sentinel, Thachanadan, Thangmi and Vishavan. Group II: A'tong, Bazigar, Bellari, Majhi, Majhwar, Manna-Dora, Ralte, Rawat, Sansi, Yakkha, and Zakhring. Group III: Turi, Great Andamanese, Khamyang, Nefamese, Parenga, and Ruga. Group IV: Malaryan, Rangkas, Ullatan, Urali, Ahom, Pali.

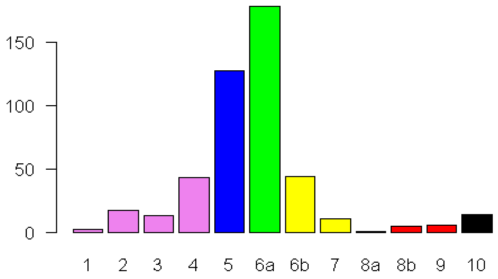
Total Languages	Living languages	Extinct Languages	
461	447	14	
Institutional Languages	75		
Developing languages	127		
Vigorous Languages	178		
In Trouble	55		
Dying	12		

Figure 2. Language vigor based on EDGIS scale (Source: http://www.ethnologue.com/country/IN)

This graph shows status of language endangerment vs. language development in India. The horizontal axis represents the estimated level of development or endangerment as measured on the EGIDS scale. The height of each bar indicates the number of languages that are estimated to be at the given level.

5. Corpora development efforts for Indian Languages

In India considerable progress has been made in terms of the generation of the text corpus with reference to Indian languages, this is true only for the widely spoken languages however. And work on spoken corpora is still in its primary stage even for the widely spoken above languages. Work on lesser-spoken languages is still in its infancy even for text corpus, leave alone speech – it is true at least in the open literature. If we compare the number of corpora with number of endangered languages in India, situation seems to be gloomy. And of course many of these corpora, especially speech corpora, are still in their raw form.

Excluding some foreign bodies, some of the main Indian organizations that are engaged in the corpora work for Indian languages are: CIIL Mysore, Karnataka; TDIL, New Delhi; IIIT Hyderabad; IIT Kanpur; IIT Mumbai; IIT Guwahati; ISI Kolkata; JNU, New Delhi; University of Hyderabad, Hyderabad; Annamalai University, Chennai; MS University, Baroda; Thapar University, Patiala; Punjabi University, Patiala; ER&DCI, Trivandrum; C-DAC, Pune & Noida; Utkal University, Bhubaneswar.

Available Text Corpora: Assamese, Bengali, Hindi, Marathi, Punjabi, Urdu, Tamil, Telugu, Oriya, Malayalam, Kashmiri, Kannada, Gujarati, Maithili, Manipuri, Indian English, English, Sanskrit, Nepali, Konkani, Kodava, Yarava, Dogri, Bodo.

Available Speech Data: Assamese, Bengali, Hindi, Marathi, Punjabi, Urdu, Tamil, Telugu, Oriya, Malayalam, Kashmiri, Kannada, Gujarati, Maithili, Manipuri, Some varieties of Indian English, Nepali, Konkani, Kodava, Dogri, and Bodo.

6. Planning data collection and corpus design

Before starting speech data collection process for any endangered language, it is advisable to keep following points in consideration:

a) *Level of endangerment***:** If linguistic variety in question qualifies to III group – which means it, is extremely endangered and almost beyond revival – then it is necessary to record as much data as possible. Quantity becomes very important in such cases. A little delay in such cases can leave you with void that will never be filled.

b) *Speaking population*: How many people speak this language? As life of a language is usually directly proportional to the number of its speakers. Is this language being passed on to the younger generation or has become limited to the older ones.

c) *Embedded foreign language*: An endangered or under-resourced language is likely to be influenced by one or more neighboring linguistic varieties. It is more often the variety, which is considered as standard or institutionalized, and is used for the purpose of imparting education in schools and other modes of official communication.

d) *Orthography*: Whether the language has its own script for writing or it is written in some other language's orthography?

e) *Literature***:** Is there some popular literature on this linguistic variety? If yes then is it available in form of manuscripts? These manuscripts may be available with the local administration, libraries, and museums, or with the town's head, or with the person/people who oversee religious ceremonies.

f) *Presence over the World Wide Web***:** Do you know if there are some websites, blogs, e-portals, etc. which are exclusively written in this language or dedicated to the people using this linguistic variety.

f) *Versatility of use***:** For corpus to be fully representative of the language, versatility of communication is another point to be kept in mind. Since an endangered language is usually limited in its domains and scope of use, you may have to provide a little training to your language consultants on what you want them to speak during recording.

Kanauji is under-resourced[6] variety of Hindi and falls within the group I – which signifies that it is rapidly going out of use. We collected 18 hours of speech data for it. Kanauji, including its all sub-dialects, has approximately six million speakers. Hindi is so embedded with Kanauji for younger generation that it does not even realize if there is any difference between Hindi and Kanauji. Most of them confirmed that they did not even know if they spoke Kanauji. The level of mixing may differ from one area to another depending on its proximity to city and towns, exposure to the education, economic status of people, etc. Like other varieties of Western Hindi, Kanujai uses Devnagari script. There is little literature available in Kanauji language; only to mention of mid 17th century authors, from Tikampur/Tikawanpur town of Kanpur district, such as *Chintamani Tripathi, Matiram Tripathi, Bhushan Tripathi and Nilkanth Tripathi* (see Keay 1920; Upadhyaya 1934). Folk Kanauji songs, which are based on famous valor stories of two brothers named Alha and Udal, are sung during festive seasons (Russell 1916). Kanauji has little presence over web. Some blogs and website are found which either talks about the history and culture of the region. Nearly all younger Kanauji speakers speak Hindi in their day to day communication with persons outside of their family.

For our corpus we collected the data from 12-15 domains of daily common use of language (Table 1). The Consultants' age, education and their contact with the other languages are also very important factors to be considered, as they together significantly influence the choice of vocabulary, mixing of codes and pronunciation factors. Most often young and educated speakers tend to use the language (pronunciation and words) according to neighboring standard form of the language. In contrast old and uneducated speakers usually retain the original form of the language.

No.	Domain	Explanation
1	Basic wordlist	equivalent words from Swadesh list and variety of domains.
2	Demographic description	regarding population, educational status, income etc.
3	Cuisine	relating to various dishes and cooking procedure
4	Family communication	day-to-day family communication among family members
5	Games	games played by children, teens and adults
6	Culture and Traditions	temples, festivals, marriages, ceremonies etc.

[6] Under-resourced languages, in our case, are the languages which don't have a wide presence on the web and/or don't have sufficient speech transcribed data.

7	Flora and Fauna	farmlands, gardens, types of crops, seasons etc.
8	Mythological stories	mythological stories, local deities, beliefs etc.
9	Daily life activities	usual working of children, teens and men and women
10	Children stories	popular stories among the children
11	Number systems	number systems of Kanauji
12	Free discourse	free discourse from variety of life activities
13	Minimal pairs	minimal pairs
14	Representative sentences	sentences providing information about tense, gender & aspect
15	Group conversation	group conversation on a given topic

Table 1. List of recorded items

The participants who were recorded had Kanauji as their first language/mother tongue. Since, like it happens with speakers of an endangered language, younger generation (12-25 years) is rapidly shifting to official standard variety, i.e., Hindi, most of them grow to become fluent bilingual. In contrast, older people (40-60 years) preserved the original form. Mid-aged people (25-40 years), as expected, fell somewhere between these two extremes. The factors that catalyze this process are education, employment and biased linguistic attitude. Information regarding these factors has been summarized in the given Table 2.

Male 15	**Female** 15	**Age**	**Education**	**Bilingualism (Kanauji-Hindi)**
8	8	12-25 years	10th-undergraduate	Fluent bilingual
5	5	25-40 years	Illiterate to undergraduate	Intermediate to Fluent
2	2	40-60 years	Illiterate	Mostly monolingual, speaking Kanauji only

Table 2. Language consultants based on age, education and gender

Here are three samples extracted from the corpus that can give you the clue as to how Kanauji is getting overshadowed by the Hindi.

Audio transcription of the samples is done in Devanagari and each sentence is enclosed in a pair of angle brackets; IPA transcription of the sample is given and English translation of the sample is provided. Instances of three dots (…) between the pair of angle brackets <>…<> refer to the fact there are one or two sentences in between the presented sentences have been omitted for they did not fit our purpose and also to save space. The code *'bjj* and *Hindi'* are ISO 693-3 codes for Kanauji and Hindi, respectively. Its use in the beginning and end of the sample marks that entire conversation (syntax) has been in Kanauji.

Sample 1: Record Entry No.121215-03

In the first sample, language consultant 1 is a fifty-eight years old man who sells milk for the livelihood and language consultant 2 is a 38 years old man who runs a small shop in the village. Except for some occasional visits, both of these men usually don't go out of village. Language consultant 1 narrates a personal story about his ghost experience in a recording session with data collector. Language consultant second 2 provides some inputs as an active listener. Only few sentences have been presented here for our purpose.

Language Consultant 1:

< bjj < पीपा लदो दूध क्यार >

<bjj<pipa ləd̪o d̪udʰ kyar>

< Can of milk was loaded.>

Language Consultant 2:

<**bjj** <दूसर कउनो आदमी ह्वात तो फेंक के भाग **ठाढ़** ह्वात।> <कोउ न रुकी,कि कोउ पक्कल लीन्हेस> **bjj**>

<**bjj** < d̪usər kəʊno **ad̪əmi** hwat̪ t̪o pʰẽk ke bʰag **t̪ʰaɽʰ** hwat̪ >< **koʊ** nə rʊki, kɪ koʊ pəkkəl linʰes > **bjj**>

If there was someone other than you, he would run away. No one would dare stop thinking that something could have caught him.

Language Consultant 1:

<**bjj** < पीपा उठाओ औ **भुँई** मा सारे का पटको।>< तौ, **इस्टैंड** (बब्बी) तरे का चलो गा > **bjj**>

<**bjj** < pipa ʊʈʰao ɔ **bʰʊĩ** ma sare ka pəʈko >< t̪ɔ, **ɪsʈɛnɖ** t̪əre ka ʧəlo ga > bjj>

< (I) raised the can of milk and turned him down to earth.> < Then, stand (of bicycle) went down.>

In this piece of conversation, both of the language consultants fluently use Kanauji without any hint of Hindi coming in their way of communication. Only the word “/ɪsʈɛnɖ/ - <stand>” is borrowed from English via Hindi into Kanauji and has got adapted phonologically adapted to Kanauji by inserting a schwa between /d̪/ and /m/ sounds. This insertion is due to the fact that Kanauji disfavors st* cluster, a vowel is inserted in the beginning of the word to break the cluster. Words like / bʰʊĩ / and /t̪ʰaɽʰ/ are typical to Kanauji vocabulary and not found in Hindi.

Sample 2 (Recording Entry No. 121211-0024)

This sample is from a twenty years old girl who is in the first year of her university education. Data collector asks her to narrate him a story in Kanauji that she stills remembers from her childhood textbooks.

<**bjj** <एक राजा राहे> <तीन रानियाँ राहें>...<राजा सिकार ख्यालैं जात राहें>...<मतलब आइटम बना-बना खवाओ करिबे>…<उई सुन लीन्हेन, उई तीनो ना र.रानि से सादी करि लीन्हेन>... <तीनो रानिन से जब सादी करि लीन्हेन तो उनके एक लड़का औ एक बिटिया भे> **bjj**>

<**bjj** <**ɛk** raʤa rahe>. <t̪in **ranıjã** rahẽ>…<raʤa **sıkar** kʰjalɛ̃ jat̪ rahẽ>…<**mət̪ləb aıʈəm** bəna-bəna kʰəwao kərıbe>….<ʊi **sʊn** linʰen, ʊı t̪ino *na rə* ..rani **se sad̪ı** kərı linʰen>. …<t̪ino **ranin** se ʤəb **sad̪i** kərı linʰen **t̪o** unke ɛk **ləɽka** ɔ **ɛk** bıʈija bʰe> **bjj** >.

<There was a king> <There were three queens>… <The king was on his way to hunt>…<Means, (I would) have (him) different types of cuisines prepared to eat>... <He heard this and married all three queens>...<After he married all three queens, one baby-boy and baby-girl were born to king>

This sample shows how Hindi words such as nouns (e.g., /sıkar/ ‘hunting’, /sad̪ı/ ‘marriage’, /rani/ ‘queen’, ləɽka ‘boy’), numbers (/ɛk/ ‘one’), verbs (sʊn ‘to hear’), conjunctions (se ‘with’, t̪o ‘then’); items from English (aıʈəm ‘item’) and Urdu (mət̪ləb ‘meaning’), etc. are frequently borrowed and used by young Kanauji speakers. In some instances borrowed items are phonologically adapted; others are borrowed in their original forms, however. For example, IPA transcriptions for Hindi words ‘marriage’ and ‘hunting’ is /ʃad̪ɪ/ and /ʃɪkar/, respectively. But in these words /ʃ/ is replaced by /s/ since /ʃ/ is not found in native Kanauji phonemic inventory. However, Hindi conjunction /t̪o/ ‘then’ and noun /ləɽka/ ‘son’ are adopted in their original forms despite of the fact that corresponding words for the same meanings in Kanauji have very similar phonotactics – /t̪ɔ/ and /lərɪka/, respectively.   One striking feature is that apart from the borrowing of lexical items from Hindi, a morphological process of ‘pluralization’ is also seemed to be borrowed. For example, *plural* of /rani/ is /ranɪn/ in Kanauji in contrast to /ranɪjã/, which is a Hindi *plural* form of /rani/.

Sample 3: Record Entry No. 12121-001

In this sample, language consultant is 28 years old man who works as a “Pandit (a kind of priest)”. He often has to visit city and towns for performing rituals. On being asked about his nature of work and how he makes money for living. Language consultant shares some details:

<**bjj** <जइसे हमारा जजमान है> <जइसे तुम हमार आदमी हो> <तुमका कुछ करवावें का है तो तुम हमते बात कीन्हेव> तुमका कुछ बता दीन अगर कि तुम अपन आदमी हो> ... <दूसर आदमी आओ, पइसे वाला है – ठीक ठाक> **bjj**>

<**bjj**<ʤəɪse **həmara ʤəʤman hɛ**> <ʤəɪse t̪ʊm həmar **ad̪əmi ho**> <t̪ʊmka **kuʧʰ** kərwawẽ ka hɛ t̪ʊm həmt̪e bat̪ kinʰew>. <ʊmka kuʧʰ bət̪a d̪in əgər kɪ t̪ʊm əpən **ad̪əmi** ho>… <dus̪ər **ad̪əmi** ao, pəɪse wala hɛ - ʈʰik ʈʰak**> bjj>**

<Like it’s my customer, like you are my man>. <And you want something to get done> < So, you talked to me>

<I would already tell you the things if you are an acquaintance>…<in case an unknown person visits and he is rich enough>…

Giving a closer look to this piece of conversation reveals that a few words and phrases are either borrowed from Hindi or are those which are very close to Hindi. For example, Hindi phrase /həmara ʤəʤman hɛ/ is used in place of Kanauji phrase /həmar ʤəʤman ajeː/. Words like /ad̪əmi/ and /kuʧʰ/ and eco-reduplicated form like /ʈʰik-ʈʰak/ is also from Hindi. This conversation does not have a word that is exclusive to Kanauji and not found in Hindi some modified form.

Sample 4: Record Entry No. (121215-0031 and 121214-0028)

This sample is from a 15 years language consultant who studies in high school. Like any other children of the current generation, girl has come out be fluent bilingual speaking Hindi and Kanauji. While children learn Kanauji from their parents and society, Hindi is the main language of education and cinema. English is also taught in the schools but due to lack of proper training, most of them don't gain more than basic skills. In this piece of recording, language consultant is first asked to tell something about her family. Language consultant answers in Hindi. Then, language consultant is asked to share some story that she could have ever heard from her grandmother. And language consultant retells an entire story in Kanauji without a hitch. Only few sentences from both parts have been taken for our purpose.

(Part-I, introduction)

< **hin** <हमार नाम X[7] है > <हमार मम्मी का नाम Y[8] है> <पापा का नाम है Z[9]> <हमारे पड़ोस में दादी दुकान करती हैं। बाबू भी दुकान में दुकान में रहतें हैं। और लक्ष्मी दीदी पुलिस में नौकरी करती हैं > **hin**>

<**hin** < həmar nam X hɛ> <həmar məmmi ka nam Y hɛ> <papa ka nam hɛ Z> <həmare pəɽos me d̪ad̪I dʊkan kərt̪I hɛ̃> <babu bʰi dʊkan me rəht̪e hɛ̃> <ɔr ləkʃmi d̪id̪I pʊlɪs me nɔkəri kərt̪I hɛ̃>**hin**>

<I am X> <My mother's name is Y> < Father's name is Z> < grandmother runs a shop in neighborhood> <grandfather also contributes in it> <sister Laxmi has a job in police>

[7] Personal information of language consultants such as their first name, family name, names of the place they work or study at has been kept confidential.

[8] Personal information of language consultants such as their family members, relatives, etc. has been kept confidential.

[9] Personal information of language consultants such as their family members, relatives, etc. has been kept confidential.

(Part II, storytelling)

< **bjj** < एक भाई, एक बहन राहें> <तो बहन कि पहले से सादी होइ गे राहे> <भाई छोटें राहें> <तो उई एक बार काहत हैं कि अम्मा बहिनी से हम मिलि आवन> <तो अम्मा कहेन कि बउआ अबै रुकौ कहे कि तुम्हरी बहिनी तो बहुत दूर राहती हैं > **bjj** >

<bjj<ek bʰaɪ, ek bəhen rahẽ>. <t̪o bəhen kɪ sadi pəhle se sadi hoɪ ge rahe>. <bʰai ʧʰoʈe rahẽ>. <t̪o ʊɪ ek bar kahət̪ hɛ̃ kɪ əmma, apni bəhɪni se həm mɪlɪ awən> <t̪o əmma kəhen kɪ bəʊa abɛ rʊkɔ kəhe kɪ t̪ʊmʰri bəhɪni t̪o bəhʊt̪ d̪ur rahti hɛ̃>**bjj**>

<There was a brother and a sister> <Sister was already married> <And brother was younger one> <So, once he asked his mother if he can go and meet his sister> <Mother said "Dear Lad, not now as you sister lives far from this place">

In 'Part-I, introduction' the girl fluently speak in Hindi; while in 'Part II, storytelling' she goes on speaking in Kanauji without a hitch. Girl, however, code-mixes some Hindi words and phrases such as 'ek bʰaɪ, ek bəhen', 'se', 't̪o' and 'd̪ur' like modern generation of Kanauji.

7. Recording specifications

Speech data was collected using two channels simultaneously: first, by Olympus LS-100 96kHz/24 PCM linear recorder which is outfitted with two built-in 90° stereo condenser microphones, second using Sony Digital Flash Voice Recorder (ICD-PX312). The major specifications of the recorder are listed below in Table 3.

Channel	**Frequency response**	**Sensitivity**	**SPL**	**Pick-up pattern**
Ch-1	20-20000Hz	high	140dB	Cardioids
Ch-2	75-20000Hz	mid	--	Mono

Table 3. Recording specifications

8. Speech data recording

The language consultants were asked to maintain an approximate distance of 10-14 inches from the microphones and keep their speaking style consistent and natural. The speech was digitized at a sample rate of 44.1 KHz/24 bit. It was all stored in wave format. Since speech data collection is far more sensitive than text collection process, continuous monitoring is required for any unaccepted deviation in pronunciation, styles, rate of speech, etc. However, in this case, data collector himself was native speaker of Kanauji and therefore it was easy to keep track of the mistakes. An onsite cursory review of the recordings was performed however. As mentioned earlier, collecting maximum amount of speech data should be the primary goal while working for an endangered language. It has two benefits: first, it provides us with the potential possibility of the phonetically rich data (covering the frequency count of the rarely occurring phones), second, more data usually cover the more diverse instances of linguistic events. Corpus has both types of speech data, i.e., read and spontaneous. Read speech data comprises of grammatical sentences, whereas spontaneous data is full of grammatical mistakes and speech disfluencies. Therefore, the corpus can further be customized for different goal oriented speech applications.

9. Establishing phoneme inventory

Establishing phonemic inventory is one of the most important things for a speech corpus. Depending on the requirements the researcher may choose biphones, triphones, syllables, words or sentences as basic speech unit. However, Indian languages are syllable centered, where pronunciations are mainly based on syllables and therefore it can be the best unit for Indian Language Speech Synthesis (Kiruthiga & Krishnamoorthy 2012).

Vowels in Kanauji are also represented by glyphs and consonants to get combined to form ligatures. Its phonemic inventory consists of 53 phones: 33 consonants and 20 vowels. Hence, the ideal number of possible biphones and triphones should be 2809 and

148877, respectively. However, due to phonotactic constraints this number will be considerably lower. A comparison is given below in Table 4:

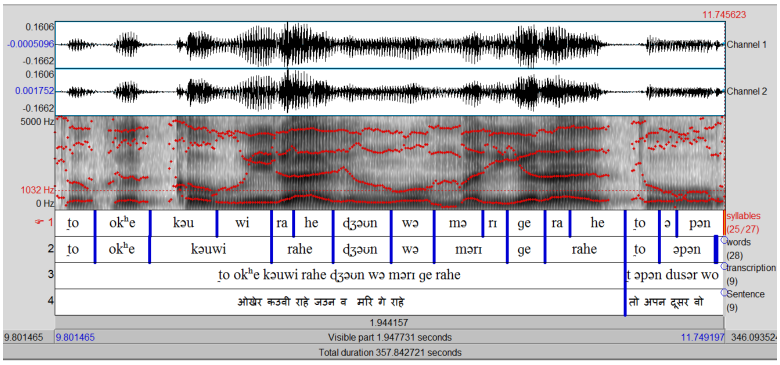
Phones	53		
Biphones	53x53	2809	< 2809
Triphones	53x53x53	148877	< 148877
Syllables	Not counted	Not counted	
Words	Not counted	Not counted	

Table 4. Analysis of phones in Kanauji

10. Phonetic transcription and annotation

The phonetic characteristics of a speech corpus play a key role in the robustness of the future speech corpus (Stănescu *et al.* 2012), and therefore phonetically transcribed text is needed to provide the complete phonetic coverage (Murtoza *et al.* 2011). Transcription is often achieved with the help of some phonetic dictionary, but, in case of unavailability of resources, one needs to design the appropriate automatic grapheme-to-phoneme converter. Task of phonetically balanced sentences cannot be achieved unless every single phrase is phonetically transcribed.

The unedited speech data Kanauji lasts for more than 16 hours. Speech data contains instances of repetitions, code mixing and switching, and another non-speech sounds such as laughter, coughing, background noise, etc. It requires a manual cleaning and editing. As it is presupposed, data tends to reduce by several hours. Annotation is supposed to be of two types: basic and extensive. The basic annotation refers to the annotation of regular pronunciation such as boundaries of phones, syllables, words and sentences; whereas extensive annotation refers to the annotation of the actual speech. Since, work on transcription and annotation is still to be done, no conclusive remarks can be made on it. Speech files will be annotated using PRAAT in text grid files. Depending on the purpose, EAGLES, ELRA, ICAME, LDC etc. may be used as guidelines for transcription and annotation.



1 t̪o | okʰe | kəu | wi | ra | he | dʒəun | wə | mə | rɪ | ge | ra | he | t̪o | ə | pən — syllables (25/27)

2 t̪o | okʰe | kəuwi | rahe | dʒəun | wə | mərɪ | ge | rahe | t̪o | əpən — words (28)

3 t̪o okʰe kəuwi rahe dʒəun wə mərɪ ge rahe | t̪ əpən dusər wo — transcription (9)

4 ओखेर कउवी राहे जउन व मरि गे राहे | तो अपन दूसर वो — Sentence (9)

Figure 2. Annotation snapshot of Kanauji speech database

11. Spoken text markup

Usually text corpora are marked up using standard markup languages like XML and SGML. However, depending on the purpose behind the collection of speech corpus and availability of economic and manual resources, level and type of marking up (STML, JSML, SSML, etc.) could differ from one speech corpus to another one. While for a corpora to be used for the purpose of language teaching, a general purpose and content mark up is done; corpora to be used in speech applications require more extensive and intensive level of content markup that can cover minute features such as intonation, prosody, types of noise, pause or long pause, repetition and so on. This mark up usually has to go thorough further post editing and normalization process. Starting the general purpose markup, to be categorized as being standard, a certain level of markup is a must.

Listening to our speech database of Kanauji, we have noted down quite a few transcripts. Though, this corpus is still in its initial stage and not ready to be used in some application process, we are still exploring on what would be an appropriate level and type of markup for it. Endangerment of a language has little to do with its application markup and therefore we are not particularly engaged to provide details on it.

12. Advantages of spoken corpora over text corpora

Speech is primary in nature than writing. It is ever evolving and more contemporary in nature. Speech contains far more linguistic information than that of text. Speech applications can be very useful even for the people who cannot read and write. Therefore, it becomes even more useful in case of people speaking endangered language, as they are often those who belong to backward section of the society. A good audio recording, accompanied by video, can provide articulatory and acoustic information. Speech data can be collected with minimum second party interference. Spoken corpora become more important when the language in focus is in endangered state because one may not have time to properly document it at once. The next attempt may lead to certain amount of setback as few elderly speakers may pass away. Endangered languages often don't have the written scripts and lack the written texts; speech corpora are particularly useful in these kinds of cases.

In the language teaching and revitalization process speech corpora can become a useful tool for learners' integration into the concerned speech community. Speech corpora present the people with near-to-immediate natural linguistic environment and almost authentic form of language. It can be utilized as an authentic effective tool or aid serving as a richer language resource than that of text materials. Brown & Yule (1983) claim that it is essential to give learner the opportunity to acquire some of the idiosyncrasies of the native natural speech in order to fit in to the culture of the target language.

Spoken corpus, which is robust and balanced in size, can be a very useful resource for those working on dialect study. It provides the evidence from the phonetic perspective and helps understand the first and second language acquisition theory. Evidence, as spoken form of the language provides finer nuances of a language. It also helps in the tasks such as modeling of articulation and prosody.

13. Challenges in constructing Kanauji spoken corpus

The spoken corpus consists of two kinds of speech samples: near-to-spontaneous and text read speech. Though we had wanted to create a phonetically balanced corpus yet things came out be different as project moved ahead. Earlier goal of the project was to record maximum amount of natural, unscripted and spontaneous speech, i.e., everyday conversations of the people. But as it happens, members of the communities were uneasy to have their personal conversations recorded in the database. Therefore, it was decided to resort to individual and group interview type of sessions and exercises. However, later on some youths agreed to let their speech recorded during their personal conversations. Some enthusiastic early age teens also had let their speech recorded. Another problem that came up was of 'NOISE'. In fieldwork, however, this is a common problem. Due to Kanauji being a variety of Hindi and unavailability of written records, transliteration of the speech was done in Devanagari script. During the process of the phonetic transcription, it appeared that there are noticeable prosodic differences in some speech samples. And these differences perhaps are due to intra-dialectal differences. However, our transcription is limited to broad level and no analysis has been done for now. As it usually happens with data collection process (especially spontaneous speech), language consultants often switched to Hindi and therefore recordings were full of the instances of code-mixing and code-switching. Sample 5 will throw a bit light on it.

Sample 5 (Recording Entry No. 12121402)[10]

This sample is from a 34 years man who works as a primary teacher in a school. The medium of instruction in his school is Hindi. Notational conventions are same as in Sample 1. In this piece of conversation primary teacher starts with requesting to data collector if he (data collector) can help him (primary teacher) to get a particular literary work from the city and then this teacher goes on informing the data collector about the

[10] Sample 5 has previously been used as a reference example in one of our other research reports and in a seminar presentation.

training he (primary teacher) is currently undergoing in his school. A look into the conversation (given below) tells how the teacher switches from Kanauji to Hindi and then vice versa. Recording shows us that the teacher does not even realize this shift and goes on talking in the same manner for about ten minutes.

<bjj < एत्ता काम करे जैव > <एत्ती चिन्हारी अपन देहे जैव **> …<** सात साल होइगे पढ़ावत > **bjj**> < hin <अभी जो **ट्रेनिग** चल रही है उसमें जब ये चीज आई **तो** दिमाग **हमारा** चकरा गया > **hin**> <**bjj** <हेई सासा का पढ़ै वाला लरिका आये। > <नीतू का जानत है > **bjj**> < hin <जब उसने चालू किया तो **हमनें** कहीं पढ़ा था ये > < दिमाग में था लेकिन **होई कुछ >** <लेकिन वो **बेसिक** से ही आधारित है तो उसका ज्ञान आवश्यक है> <**ट्रेनिंग** एक दिन के लिये थोड़ी न होती है> < **आल लाइफ** के लिये होती है> **hin** >

<**bjj <** et̪t̪a kam kəre dʒɛw >. <et̪t̪i tʃɪnhari əpən d̪ehe dʒɛw >**…**<sat̪ sal hoɪge pəɽʰawət̪ > **bjj**> <**hin** < əbʰi dʒo ʈreniŋ tʃəl rəhi hɛ ʊsme dʒəb je tʃidʒ aɪ *t̪o* d̪ɪmag həmara tʃəkra gəja. > **hin**> <**bjj** < hẽɪ sasa ka pəɽɛ wala lərɪka aje> <nit̪u ka dʒanət̪ hɛ. > **bjj**> <**hin** < dʒəb ʊsne tʃalu kija t̪o həmne kahĩ pəɽʰa t̪ʰa je. <d̪ɪmag me t̪ʰa lekɪn hoɪ kutʃʰ **>… <**lekɪn wo besɪk se hi ad̪ʰarit̪ hɛ. t̪o ʊska gjan awəsyək hɛ> ʈreniŋ ek d̪ɪn ke lɪje t̪ʰoɽi nə hot̪ɪ hɛ. > <al laɪpʰ kɛ lɪje hot̪ɪ hɛ. >**hin**>

< please do this much favour to me>. < It will remind me of you >…< I have been teaching for seven years> < When I came across this thing in the current training, my mind went blank> < He (the trainer) comes from Sasa and knows the Neetu > < As soon as he started, **I** knew it as **I** read about it somewhere>. < I was in my mind >. But then, may be something**>…** < But, it starts from basic concepts. So, it's necessary to know it >. < Training doesn't help only for a day (Meaning) > < It's for whole life>

This sample provides us with a fair idea that Kanauji speakers frequently code-mix (underlined words and phrases) and code-switch from Kanauji to Hindi and vice versa. Words 'training' and 'basic' and phrase 'all life' from English language is also used in different statements.

Embedded Foreign Languages: When the target language is under-resourced or low resourced it is likely to be influenced by one or more embedded foreign languages. The development of speech applications requires a corpus which is adequately developed (having good amount of transcribed speech data) and which in turn need good amount of text data for the construction of prompts, language models, pronunciation dictionaries, etc. Gathering all these components of an under-resourced language is quite a difficult task.

14. Conclusion

There is a dire need for multilingual and multipurpose speech corpora for Indian Languages. Despite many efforts, the number and varieties of such corpora are not yet enough to facilitate the academic research beyond their developers. It is fact that the majors' efforts were made for the larger languages of India. However, for corpus-based studies on smaller or endangered languages, even theoretical in nature, it is still the responsibility of the concerned researcher to develop the corpus of that language. This is a challenging, but important task ahead of any search studies.

These corpora would serve two perspectives: to provide language documentation and to make the benefit of information technology available to all sections of the society. Speech corpora can wonderfully work in language revitalization process by assisting in the process of product development for endangered languages. Considering the fact that India is one of the most linguistically richest countries in the world but hundreds of its languages are on the verge of being lost, a need for preservation and maintenance of linguistic heritage becomes all the more important and thereby role of language corpora also become very significant.

References

ALLWOOD, J. (2006) "Language survival kits", in A. Saxena & L. Borin (eds.), *Lesser-Known languages of South Asia: Status and policies, case studies and applications of information technology. Trends in linguistics studies and monographs*, vol. 175, Berlin: Mouton De Gruyter, 279-292.

BIRD, S. & G. SIMONS (2003) "Seven dimensions of portability for language documentation and description", *Language*, 557-582.

BORIN, L. (2006) "Supporting lesser-known languages: The promise of language technology", in A. Saxena & L. Borin (eds.), *Lesser-Known languages of South Asia: Status and policies, case studies and applications of information technology*. *Trends in linguistics studies and monographs*, vol. 175, Berlin: Mouton De Gruyter, 317-337.

BROWN, G. & G. YULE (1983) *Discourse analysis*, Cambridge: Cambridge University Press.

COX, C. (2009) “Corpus linguistics and language documentation: Challenges for collaboration”, in John Newman, R. Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, Amsterdam: Rodopi, 239–264. <Retrieved on October 23, 2013 from http://www.ualberta.ca/~aacl2009/PDFs/Cox2009AACL.pdf>

CRYSTAL, D. (2000) *Language death*, Chicago: Ernst Klett Sprachen.

SOMAN, S. (2013) *India lost 220 languages in last 50 years*, (August 9, 2013), Times of India reports. <Source: http://timesofindia.indiatimes.com/india/India-lost-220-languages-in-last-50-years-survey-finds/articleshow/21720601.cms>.

JANSEN, J. B. & S. B. SØRENSEN (2005) “Voices of the world: The extinction of language and linguistic diversity [motion pictures]”, USA: Final Cut Productions. <Retrieved from http://digital.films.com/play/2VTLH2 on October 17, 2013>

JOHNSON, H. (2004) “Language documentation and archiving, or how to build a better corpus”, *Language documentation and description*, 2, 140-153.

KEAY, F. E. (1920) *A history of Hindi literature*, Mysore City: Wesleyan Press.

KIRUTHIGA, S. & K. KRISHNAMOORTHY (2012) “Design issues in developing speech corpus for Indian languages—A survey”, in *2012 International Conference on* Computer *Communication and Informatics,* 1-4.

MCENERY, T., & N. OSTLER (2000) “A new agenda for corpus linguistics-working with all of the world’s languages”, *Literary and linguistic computing*, 15(4), 403-420.

MURTOZA, S. M., F. ALAM, R. SULTANA, S. A. CHOWDHUR & M. KHAN (2011) “Phonetically balanced Bangla speech corpus”, in *Proc. Conference on Human Language Technology for Development 2011,* Alexandria, Egypt, 87-93.

PAUL, L. M., G. F. SIMONS & C. D. FENNIG (eds.) (2013) E*thnologue: Languages of the world*, Seventeenth edition, Dallas, Texas: SIL International [Online version: <http://www.ethnologue.com> accessed: August 4th, 2015].

RUSSELL, R. V. (1916) *The tribes and castes of the central provinces of India,* vol. 3, London: Macmillan and Company, limited.

STĂNESCU, M., H. CUCU, A. BUZO & C. BURILEANU (2012) “ASR for low resourced languages: Building a phonetically balanced Romanian speech corpus”, *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Bucharest, Romania, IEEE.

STEPHENSON, P. (2013) *Living Language*, Faculty Q&A With Linguist John McWhorter (February 20th, 2013): <http://news.columbia.edu/content/living-language-faculty-qa-linguist-john-mcwhorter> (accessed on August 20, 2015).

TOGNINI-BONELLI, E. (2010) "Theoretical overview of the evolution of corpus linguistics", *The Routledge Handbook of Corpus Linguistics.* Abingdon: Routledge, 14-27.

TROSTERUD, T. (2006) "Grammatically based language technology for minority languages", in A: Saxena & L. Borin (eds.), *Lesser-Known languages of South Asia: Status and policies, case studies and applications of information technology*, vol. 175, Berlin: Mouton De Gruyter, 293-315.

UPADHYAYA, A.S. (1934) *The origin and growth of Hindi language and its literature*, Patna: Patna University.