ELKL – 4

4th International Endangered and Lesser-known Languages Conference

Abstracts

February 25 – 27, 2016

Department of Linguistics
K.M. Institute of Hindi and Linguistics
Dr. Bhim Rao Ambedkar University, Agra

Cover Designed by: *R. Karthick Narayanan*

Edited, Proofread and Typeset by: *Mayank Jain*

With contribution from: *Falak Kashyap & Atul Kumar Ojha*

Contents

Keynote Lecture

Plenary Talks

Panel Discussion
on
'Language Endangerment and Revitalisation: Social Prestige, Linguistic Rights and Technological Intervention'

Interactive Session with Nikolaus Himmelmann

Tutorials

Research Papers

Keynote

Endangerment of lesser-known tribal languages: With focus on Central and Southern India

B. Ramakrishna Reddy
Telugu University, Hyderabad
brkrin5@gmail.com

The Census of India 2001 lists the total number of languages in India as 122 which consists of 22 scheduled and 100 non-scheduled languages, eliminating all those having less than 10,000 native speakers. This understatement ignores another 180 or so lesser known languages of the 5 language families. The family wise division of total languages is: Indo-Aryan 20, Dravidian 30, Austroasiatic 20, Tibeto-Burman 110 and the Andamanese 4.

The present talk is confined to the endangerment of tribal languages of Central and Southern India, mainly of the Dravidian and Munda groups. Tribal bilingualism, the status of tribal language vis-à-vis major languages, endangerment and maintenance of tribal speeches, need for their protection and preservation and measures to revitalize the lesser known languages are some of the topics discussed in detail with copious examples.

Plenary Talks

Introducing the Google Endangered Languages Project

(via Google Hangout)
Anna Belew, University of Hawai'i at Mānoa, belew@hawaii.edu
Bryn Hauk, University of Hawai'i at Mānoa, bryn.g.hauk@gmail.com
Kevin Lowe, First Languages Australia

Languages are entities that are alive and in constant flux, and their extinction is not new; however, the pace at which languages are disappearing today has no precedent and is alarming. Over 40 percent of the world's approximately 7,000 languages are at risk of disappearing.

The Endangered Languages Project (www.endangeredlanguages.com) puts technology at the service of the organizations and individuals working to confront language endangerment with resources to document, preserve and teach them. Through this website, users can not only access the most up-to-date and comprehensive information and resources on endangered languages, but also play an active role in putting their languages online by submitting information in the form of text, audio or video files.

This presentation will explore the status of endangered languages on a global scale and look at ways of using and contributing to the Endangered Languages Project website.

Language technologies: What endangered languages can teach us

Dafydd Gibbon
Universität Bielefeld, Germany
gibbon@uni-bielefeld.de

The most common angle on language and speech technologies is to ask how existing technologies can be applied to improve documentation of endangered and other minority languages, and to modernise communication in these languages in order to enable participation in regional, national and global economies. But this approach begs an important question: the technologies concerned have been developed for a very small number of languages of affluent societies - far below 1% of the 7000 languages of the world. While it is often claimed that the technologies are language-independent, this is only half the story, and basically means simply that the technologies are interchangeably applicable to each of the 'affluent languages' to which they have already been applied.

While the other 99.9% of languages of the world do share many properties of the 0.1%, very many of them also have typological properties which require models which do not enter into the standard technologies, and which therefore require attention /before/ language and speech technologies can be applied to them. For this reason the focus of the talk is on what we can learn from endangered languages.

There are many general things that we can learn. Rather than taking our own stable languages for granted, we can learn to appreciate what it means for a language to disappear: not only the loss of cultural and environmental information contained in its lexicon, which is the first thing that generally comes to mind, but also deep insights into the range of human cognitive powers which are provided by the phonological, prosodic and grammatical structures of a language. It is these which enter into technological applications.

Building on this initial claim I will discuss some of the typologically different and interesting features which are not covered by present models, and which need attention both for documentation and communication. The focus will be mainly on the structures of speech sounds, from syllabic phonology and tonal prosody to speech timing, which are a primary requirement for successful deployment of language and speech technologies in comprehensible and natural speech synthesis and in automatic speech recognition.

Data integration for language documentation: Developing technologies for long-term, team-based investigation

Jeff Good
State University of New York at Buffalo, USA
jcgood@buffalo.edu

One of the great successes of the first decade or so of work on language documentation was the development of recommendations for the digital encoding of language resources. An emphasis on the adoption of open standards and lossless recording formats led to advisory statements which were relatively straightforward to follow and ensured that the products of documentary work could be used by diverse audiences far into the future.

However, the resources created by documentary projects too often remain relatively inaccessible due to a lack of discoverability. This is often because they do not have the rich metadata that is required for successful exploration of a documentary corpus, whether by members of the team which created the corpus in the first place or by outside users. One cause of this problem is the absence of widely adopted solutions for encoding metadata which allow links between various datasets to be explicitly encoded in a general way. Another cause is that many crucial aspects of the documentary workflow rely on the human documenter to consistently perform relatively complex tasks under difficult field conditions.

This talk will consider these issues from the perspective of the general computational domain of data integration. Solutions involving the use of Semantic Web technologies and the implementation of specific workflows into data collection tools will be sketched out; suggesting that, with appropriate effort, problems of discoverability can be effectively addressed using existing technologies. As part of the discussion, basic concepts of the Semantic Web and workflow modelling will be introduced along with illustrations of how they are being applied in the context of a team-based language documentation project examining multilingual language practices in rural Cameroon.

Taking spoken language seriously

Nikolaus Himmelmann
University of Cologne, Germany
n.himmelmann@uni-koeln.de

Despite proclamations to the contrary, the concept of language dominating modern linguistics is essentially written language. This has many repercussions in theory and practice. As for documentation, a major issue in this regard is the question of how best to represent spoken language, as (good) documentations mostly consist of audio or audio-video recordings of spoken (or gestured) interactions. The talk is particularly concerned with issues pertaining to the proper segmentation of spoken language when transcribing it.

Panel Discussion

'Language endangerment and revitalisation: Social prestige, linguistic rights and technological intervention'

Panelists
1. Girish Nath Jha teaches Computational Linguistics at Special Centre for Sanskrit Studies in Jawaharlal Nehru University, New Delhi. He has worked extensively on developing resources and technologies for several lesser-known (as well as well-known) languages of India, especially Sanskrit and is currently the leading researcher in the field.

2. David Nathan is currently working at Batchelor Institute of Indigenous Tertiary Education and the EL Training Group, Australia. Previously, he has directed The Endangered Languages Archive at SOAS, London and has been responsible for establishing it as one of the leading language archives for storing and processing the documentation data.

3. Monojit Chaudhary is a Researcher with Microsoft Research Labs India and works extensively in Natural Language Processing and Computational Linguistics. A large part of his current research is focussed on code-switching, code-mixing and multilingual processing of language data.

Moderator:
The panel will be moderated by Dafydd Gibbon who has worked extensively on documentation of languages (especially African languages) as well as development of language resources and technologies.

Motivation (Ritesh Kumar):
The idea behind this panel is simple.

1. I would like it to discuss the possibility of using documentation data (produced primarily for archives like ELAR and stored there) for development of language technologies for these languages. As we know, documentation data is generally quite varied, structured and rich in the sense that it generally has language data from a large number of contexts and situations and more importantly, they are carefully glossed (annotated) and aligned at word-level (high-quality parallel corpora). However, most of the times, these data are stored in formats which becomes difficult to process for developing NLP tools and applications. Another issue is, of course, related to the access to this data for research and development. So the first part would be to understand the format in which data is stored in the archives and are there some parsers/tools to process the data. If not, is it possible to develop some - how difficult / easy is that (depends largely on the structure of the data).

2. The second question is more theoretical and related to the utility of such an exercise - parsing the data and then development of language technologies. Could this in any way be a tool for language revitalisation or at least for altering the social prestige of the language? Could this - any piece of technology ranging from keyboards to machine translation systems - prove to be useful for the community.

My idea behind organising this panel is to find solutions to some of the practical problems related to this enterprise and also, if possible, forge some kind of collaboration (not necessarily involving finances) to start developing LTs for lot of endangered, lesser-known and less-resourced languages of India. But, of course, I don't want to limit it to that. I would like you to bring up different questions (and answers, may be?) and fresh perspective to these issues during the discussion.

Panel schedule
The total duration of the panel will be 90 minutes, including time for discussion among the panellists as well as with the audience after the presentations.The schedule will be somewhat flexible, but roughly as follows:

After a brief introduction by the moderator, each panelist will have up to 15 minutes for presentation with 5 minutes for questions and initial discussion (totalling 60 minutes). This will be followed by discussion with the audience and questions to the panelists (20 minutes), followed by final practically oriented statements by the panelists (10 minutes)

Brief outlines of panelist positions
Girish Nath Jha
I will start with the data/standards requirements for developing language technology for less resourced languages and in that context discuss whether language documentation data as being collected can be useful. I will then present some of our own platforms and tools for enabling less resourced languages.

David Nathan
I will make some suggestions from the wider perspective of language documentation, and its relation with archives. There are several things that could be done to increase - or at least improve - the usability of archive collections for the development of language technologies. I discuss under 3 headings: the aims of language documentation; curricula and skills; and division of labour.

Particular topics: "intermediate" technologies that can conceivably come out from "mobilising" materials in archives that are distinct from the kind of language and NLP technologies, e.g., pedagogical multimedia etc; computationally tractable data and the need for skills in data management.

Monojit Choudhury
I will talk about what are the different stages of language technology and NLP building blocks that one would need to claim "we have digitally and computationally preserved a language", and to do so, how much and what kind of documentation data is needed. My position is going to be: "we need to collect data a little differently than the traditional methods of documentation, especially if we want to build say a translation or speech recognition system for endangered languages (EL).

Also: how can we use technology to collect more EL data in less time and effort? I see some initiatives in this direction based on mobile apps (we have one such project in MSR: CGNetSwara, and of course there is Aikuma). Use of social media for spotting and scraping EL data sounds quite fascinating to me, and I know a couple of projects along that line too.

And how can we, as a community, come together to achieve this feat?

Dafydd Gibbon
1. Language documentation, for whatever purpose, needs to be accessed efficiently – either via metadata and systematic multi-tier annotation of text and speech, or via techniques for unstructured search of text, pictures and videos. In other words, methods for searching resources are a prerequisite to any application. Some of the search techniques will involve standard database models, others will require natural language processing for parsing and other analysis (such as dictionary induction from data), signal processing (automatic and semi-automatic speech annotation) and artificial intelligence methods.

2. What is the impact of language documentation on small communities with endangered languages? And what is the impact of speech and language technology applications on these communities? For example, do they lead to localisation for the endangered language, or to swamping the endangered language with a dominant language?

Collaborations and curricula: Keys to improving the mutual contributions of documenters and language technologists

David Nathan
Batchelor Institute of Indigenous Tertiary Education and the EL Training Group, Batchelor
dn2@soas.ac.uk

I will make some suggestions from the wider perspective of language documentation, and its relation with archives.

There are several things that could be done to increase - or at least improve - the usability of archive collections for the development of language technologies. I discuss under 3 headings: the aims of language documentation; curricula and skills; and division of labour.

Language documentation (aka documentary linguistics) is not only the only declared and systematic response to dealing with language endangerment and loss, but it also has its own distinct goals and methods. As is common, I take Nikolaus Himmelmann's definitions as a kind of "constitution" (Himmelmann 2002). To summarise: documentation defines methodology and outcomes for a *comprehensive record* of the *range of speech practices* of a *community*, to serve a range of *audiences* (including those speech communities themselves), *disciplines*, and *purposes*. This implies a variety of specific goals, methods, genres, outputs and outcomes only some of which relate to the creation of computationally tractable data (or even formal language description), and that *documentation* archives are not necessarily *data* archives.

Documenters who do intend to create and use tractable data can be better informed about how their work dovetails (or doesn't) with others working to document languages. Often, those coming from the NLP side are less aware of the breadth of documentation goals (and sadly we all too frequently see rash motivations for NLP "solutions" with little or no research, evidence or evaluation). On the other hand, those coming from the more humanistic side often have underdeveloped "data science" skills. Most of the woes of archives are not caused by depositors' lack of archiving skills but by lack of depositors' basic data management skills. It would be good to see improvement of under- and post-graduate curricula so that graduates (and teachers!) have a proper grasp of the principles and practices of data management, as well as more training workshops and suitable recognised training curricula. Related to this would be a better understanding of metadata (i.e. expressivity, not merely following "standard" schemas) and what Peter Austin (Austin 2013) calls "metadocumentation" - reflective descriptions of the contexts, motivations, methods, relationships, histories, attitudes and responses to the documentation activities. As a simple example, documenters could much more explicitly describe the characteristics of their collections, indicating which items are machine readable, the conventions and structures used, and the audiences and purposes of those items.

The third category, "division of labour" refers to the often covert rhetoric of those who focus on processing tractable data. Language documentation does not consist solely, or even necessarily, of "data". Audio and video recordings (and many other manifestations of documentation such as contextualising and enriching text material) should be understood as resources, not (yet) data. Turning, though, to one kind of data - interlinearisation or glossing - it is well known that to take an audio recording and annotate the morphology of its linguistic content can take up to 100 hours for each hour of recording. Because such work has been so frequently treated as a *sine qua non* of documentation, it has caused many documenters to spend vast amounts of time doing such work even though it may not be very relevant to achieving their project goals, or has even set up inexperienced documenters to fail. Those needing tractable material cannot expect the average language documenter to serve up such materials; nor should responsible funders allow such lapses in methodology. Those requiring tractable data need to be responsible for its creation and quality by contributing to creating it, or, ideally, collaborate with other documenters to realise the humanistic as well as technical goals,

rather than treating language documenters as data-producing slaves. Thus, “division of labour” is a matter of methodology, ethics, and financial imbalance or even exploitation.
ELAR (SOAS University of London) was addressing some of these issues. For example, we were the first and only language archive to be built based on a “social networking” (aka “Web 2.0”) model to facilitate communication about a deposit between the depositor and potential users, thus enabling, for example, an NLP practitioner to directly obtain further information about the details of file and data structures from the documenter who created them. Ideally, collaboration would begin earlier in the process of creating documentation, but this innovation was a step in the right direction.

To sum up, we might say that the combination of the original broad definitions of documentation together with the NLP community’s hunger for data is a dangerous mix. In the discussion, I will sum up the history and evolution of language documentation over the last 20 years in an attempt to suggest why this situation has come about, and suggest some ways forward.

References:

Himmelmann, Nikolaus. 2002. Documentary and descriptive linguistics (full version). In Sakiyama, Osamu & Endo, Fubito (eds.), *Lectures on Endangered Languages:* 5 (Endangered Languages of the Pacific Rim, Kyoto, 2002) Online at http://www.hrelp.org/events/workshops/eldp2005/reading/himmelmann.pdf

Austin, Peter K. 2013. Language documentation and meta-documentation. In Ogilvie, Sarah & Jones, Mari (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*. Cambridge: Cambridge University Press. Online at http://www.hrelp.org/aboutus/staff/peter_austin/AustinMetadocumentation.pdf

Interactive Session with Nikolaus Himmelmann

University of Cologne, Germany
n.himmelmann@uni-koeln.de

Documentation of Buksa language

Anshikha Adhikari
University of Hyderabad, Hyderabad
adhikarianshikha29@gmail.com

Language documentation as defined by Himmelmann (1998:166) is to "provide a comprehensive record of the linguistic practices characteristic of a given speech community". It deals with the compilation of data in recorded and digital form so that the information can be disseminated to its users. This recorded data can also be further used to create a corpus of the given language, which in turn can be helpful in developing dictionaries. Documentary linguistics in the present context is of great need. This is so because several languages are on the verge of endangerment or extinction.

Buksa is an Indo-Aryan language spoken by the Buksa community in Uttarakhand and few districts of Uttar Pradesh. As labelled by Ethnologue, Buksa is a separate language with Dangaura, Kathoriya and Rana as its sister languages. According to the data of 1997, it has around 43,000 speakers. There is no known documentation of Buksa language. This is the first attempt to create a lexical database for the language. Already the speakers of the language have started abandoning it thereby, favouring the major language (Hindi) for wider communication. The language does not face an immediate danger of extinction but this not means that it must left undocumented. It is so because the future is unknown and the language documentation is of an immediate necessity. There is always an impending threat to the languages of the world thus, the cultural and the intellectual knowledge must be recorded and documented for its users. The presentation explores the language documentation of Buksa language and also the difficulties which were encountered during the fieldwork. The main problem was the absence of the secondary data which if present could have provided an ample help to know and comprehend the language and the culture of the Buksa community.

Documentation of an endangered language: Beda

Yogesh Dawer
Dr Bhim Rao Ambedkar University, Agra
yogeshdawer@gmail.com

Beda is an endangered language spoken in Leh-Ladakh region of Jammu & Kashmir. This language belongs to Tibeto-Burman language family. The speakers are less than 1000. The Members of the community are mostly uneducated. The community has very limited resources, therefore lives on humble means. Their main profession is music and dance. The male members work as a musician and the women perform as a dancer. They usually perform in hotels. It is one of the sources of income for the community. Some members work as taxi drivers. One community member claimed that his grand-father was a musician with the King. He informed that the community was also known as 'MON'.

The community members are multilingual in Hindi, Bodhi/ Ladakhi and Beda. Bodhi/Ladakhi is the major regional language. The younger generation is bilingual in Bodhi/ Ladakhi and Hindi. But, they don't use Beda language. I visited the community in Leh for the linguistic field work in September

2015 for a fortnight. The field work was conducted for a language documentation project of Beda under SPPEL.

Challenges: Some of the challenges faced during the field work are as follows:

1. Availability of the community members: It was difficult to get time from the community members as they are mostly occupied with their work. When I first visited the place, I went to Chuchot Yokma village to meet some musician families of the community. But, most of the people were not present in the village. Thereafter I visited the village in the early morning and could talk to some people and get some data. Their villages have limited facilities; therefore they live on humble means.

2. Geographical Accessibility: Their villages are located at the remote location in the Himalayan hills. For almost six months during the winter season, the community lives in isolation because of heavy winter situation.

3. Attitudes towards their language: As the community mainly works as musicians, which is not seen as a good profession, they are reluctant to speak Beda with others so that they are not identified as Bedas, the musicians. It stops them to use Beda publically and instill negative attitude towards their language.

Documentation of an endangered language: Malayan

D. Jenifer
Madurai Kamaraj University, Madurai
jenijeniroy@gmail.com

An endangered language is a language that is at risk of falling out of use as its speakers die out or shift to speaking another language. Government of India has started a project called SPPEL. SPPEL stands for Scheme for Protection and Preservation of Endangered Languages. The main objective of the project is to document endangered languages of India. The SPPEL project is about collecting data from the languages which are spoken by less than 10,000 speakers. This paper is on documentation of Malayan, a language which is spoken in southern region of India and belongs to Dravidian family.

Malayan is one of the tribal communities in Kerala. Malayan tribes are mainly live in the four districts of Kerala (Trissur, Ernakulam, Idukki and Kottayam). I went to the field in September 2014 with my team. I visited Malayan community of Trissur district. First we collected data from the people who are living in Vachamaram colony located in Vazhachal forest division in Athirappally area. The community has 8 families with around 29 people. Then we visited Thavalakuzhipaara which is located in Erunakkulam district (45 minutes short cut route via taxi journey from Vachamaram) in the centre of the forest. The community, there consists 50 families with a population of around 250 people. I tried to document Malayan tribe's language and ethno-linguistic information from the Malayan community of Trissur district. The data include words, sentences, stories, narrations and folk songs and some indigenous knowledge shared by the tribe. It aims to provide a comprehensive record of the linguistic practices characteristic of the given speech community.

One of the major reasons for this tribe to shift to the major language of the area is that the government of Kerala offers free education to tribal people with hostel facility. The tribal people send their children to hostel when they reach the age of five. The children above the age of five can get education, hostel facility and food free of cost form the government. That is very attractive for the poor Malayan tribe. As the present generation is staying and studying in hostel for years, most of the time, they are not staying with their parents. So they speak Malayalam which is their medium of instructions. Middle ages people also speak Malayalam. Malayalam is the major and official language of that area (Kerela).

However, only elderly people speak Malayan to converse with the people of their age group. So during data collection we have to mainly depend on the older people of the community. But when these old people see outsiders they speak in Malayalam. So when we go for data collection, they do not use Malayan language in front of us this makes the data collection and documenting the language a difficult task. Moreover, we also faced some practical problems like all the elderly people are eating betel leaves with tobacco and betel nut so it is very difficult for us to record their speech. They are not even ready to spit as they think that will be wastage of the thing. They are not even ready to stop eating betel for some time as they are highly addicted to it. There were also some more challenges which I and my team faced when we were in the field. I would like to discuss these in details in the presentation.

Documenting Dhimal

Bornini Lahiri
Central Institute of Indian Languages, Mysore
lahiri.bornini@gmail.com

Dhimal is a Tibeto-Burman language spoken in Morang and Jhapa districts of Nepal and Darjeeling district of West Bengal, India. The population of Dhimal speakers in Nepal is 19,300 (according to Ethnologue). But in India the number of speakers is much less. The number of Dhimal speakers in India is estimated to be around 900.

In the presentation, I would discuss my field experience in documenting Dhimal and the problems I faced in collecting and analyzing the present data. I did a pilot survey for documenting Dhimal. I mainly used questionnaire method. A questionnaire of around 300 pages which had basic wordlist and sentence list was made which I carried to the field. The whole questionnaire was not completed but some sections like Kinship Terms, Colour Terms, Body-part terms, were completed. The questionnaire was mainly used as a guideline and I did not totally stick to it. Other than questionnaire I also collected data by pointing out at objects of surroundings like different types of plants & trees, birds and insects, household items etc. Group discussions with the informants also gave us a lot of information, mainly about Ethno-linguistics.

I also collected some narrations, folksongs, folkdances etc.

Due to continuous contact with Bengali, Rajbanshi and Hindi, the Indian variety of Dhimal has borrowed lots of words from their neighbouring speech communities. Dhimal community has adopted lots of cultural activities from these speech communities. The cultural changes have also made changes in the language. The effect of Indo-Aryan languages in Dhimal can be witnessed at the level of lexicon, phonology, morphology and ethno-linguistics. However it is interesting to note that some domains have borrowed more words than the other domains. Domains like body parts have borrowed words for basic body part terms like finger (*aŋuli*), palm (*ʈ̤ala*) , beard (*ɖ̤arih*), which have replaced the Dhimal words. Whereas, in kinship terms, no borrowings were witnessed. The speakers easily gave terms for both proximal and distal kinship terms like, wife (*be*), daughter (*camɖ̤i*), grandfather (*adʒu*), husband's elder brother (*puɟu*), husband's sister (*hulme*) etc. It would be interesting to explore why borrowing is more in one domain like body part terms but not so in another domain, kinship terms.

Due to close cultural affinity with the Bengali community, Dhimal community has adopted many cultural aspects from this community. This 'strong cultural pressure' (Thomason & Kaufman 1988) has resulted in borrowing of concepts and culture. Dhimals have borrowed the concept of religion, festivals etc. Due to immense contact and hence borrowing, it is difficult to understand whether certain terms are inherent part of Dhimal, which have been replaced by the borrowed words or whether the whole concept is borrowed. It is a known fact that borrowing is a process of learning and

acculturation (Mesthrie et.al 2008). This has made the distinguishing line between the inherent Dhimal culture and the adopted culture, bleak. Now for some part of the culture and the language, it is difficult to judge whether these are inherent parts of Dhimal or an adopted version. For example, Dhimal has adverbs out of reduplicating verbs, e.g.

(1) *ka* *jimtiŋ-jimtiŋ* *maigʰa*
I sitting-sitting tired
'I was tired of sitting.'

However, in some sentences Dhimals use adverbs out of reduplicating Bengali adverbs, e.g.

(2) *ua* *gʰənəŋ-gʰənəŋ* *lokʰe*
He/she frequently come
'He/she comes frequently.'

Now, it is a challenging task to unveil whether Dhimal has replaced the reduplicated words (frequently) with Bengali words or whether it has borrowed the whole structure, wherein Dhimal does not have structures where adverbs are reduplicated. In the presentation I would like to focus on some more problems related to this.

References:

Mesthrie, R. & Swana, J. & Deumert, A & Leap, W. 2008. *Introducing sociolinguistics*. Edinburg: Edinburg University Press.

Thomason, Sarah Grey & Kaufman, Terrence. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Challenges and prospects of (linguistic) fieldworker as a facilitator in language documentation projects: Lessons and experience from the Blue Mountains (Nilgiri)

R. Karthick Narayanan
Jawaharlal Nehru University, New Delhi
karthick1988.nara@gmail.com

"The number of trained investigators is very small, and the number of American languages that are mutually unintelligible [is] exceedingly large." (Boas, 1911. as quoated in Mihas, 2012.). This statement made almost a century before is unfortunately true in today's context. The most optimistic claim (Campbell & Et al.) projects language loss at about one language every three months. This alarming rate of language loss is not just a catalyst for large scale language documentation effort but also a call to redesign our methods and approach to language documentation.

Language Documentation unlike description seeks to produce a "lasting, multipurpose record of a language" (Himmelmann 2006). Creating this comprehensive record of a language is massive task. Such task is not feasible within the traditional 'lone-ranger' methodology. This necessity has kindled passions among many linguist to include native speakers in their documentation projects as active partners in research. This non traditional approach to language documentation include "cooperative", "participatory", "collaborative", "empowerment", "community-based" and "sub-contracting" research models (Chaykowska-Higgins 2009). Even though the benefits of such methods have been well discussed in literature and perfected in many countries, these research models have not been extensively applied by Indian scholars. This presentation will provide a description of probably the first native speaker centric documentation effort in India and discuss in detail the challenges that were encountered by the fieldworker during the documentation.

The study: 'Toda Tales and Songs' on which this presentation is based on is part of the Oral literature archive being prepared under the aegis of Prof Anvita Abbi, Center for Oral and Tribal Literature, Sahitya Akadami. The aims of the project was to create audio-visual documentation of Toda folktales and songs, transcribe and annotate the collected materials, translate them in to major Indian languages and publish them. Accomplishing these tasks within the traditional research models was clearly not feasible and this was identified at the commencement of the project. Hence the native speaker who was hired as the consultant was trained in the basics of linguistics fieldwork and was annexed as a co-producer of knowledge in this project. This practically meant a fieldworker becoming a facilitator. Challenges faced during this transformation were numerous. However for this discussion I would restrict to discuss the challenges in transforming myself from a fieldworker to facilitator and my consultant/informant to a co-producer of knowledge. These challenges include: the negotiation of (linguistic) ideological differences between the fieldworker and consultant, and developing co-working strategies. Challenges discussed during this presentation are mostly personal but these challenges I believe are directly connected to the way in which linguistic field work is taught in Indian and in general at other universities around the world.

References:

Boas, Franz. 1911/1966. *Introduction to the handbook of American Indian Languages*. Bulletin 40, Part 1.1-83. Washington, DC: Smithsonian Institution, Bureau of American Ethnology, Government Printing Office.

Chelliah, Shobhana L. & de Reuse, Willem J. 2011. *Handbook of descriptive linguistic fieldwork*. London: Springer.

Crowley, Terry. 2007. *Field linguistics: A beginner's guide*. Oxford: Oxford University Press.

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working with Canadian indigenous communities. *Language Documentation & Conservation* 3(1).

Dixon, R.M.W. 2010. *Basic linguistic theory: Methodology*, vol.1. Oxford: Oxford University Press.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork. In Gippert, Jost & Himmelmann, Nikolaus P. & Mosel, Ulrike (eds.), *Essentials of language documentation*, 31-66. Berlin: Mouton de Gruyter.

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4(1-4). 123-155.

Mihas, E. I. 2012. Subcontracting native speakers in linguistic fieldwork: A case study of the Ashéninka Perené (Arawak) research community from the Peruvian Amazon.

Gippert, J., Himmelmann, N., & Mosel, U. 2006. *Essentials of Language Documentation* 178. Walter de gruyter.

Tutorials

TypeCraft - an online annotation tool

Dorothee Beermann & Lars Hellam
Norwegian Institute of Science and Technology, Norway
dorothee.beermann@hf.ntnu.no & lars.hellan@ntnu.no

TypeCraft (TC) is an Open Infrastructure that allows the creation and retrieval of Interlinear Glossed Texts (IGT) - the standard data format in linguistics. It is a user-driven database which offers functionalities needed for the management of textual data. TypeCraft's main function is to enable the sharing of linguistic data, such as transcribed and annotated oral narrations, annotated small texts, and linguistic collections exposing phenomena of special interest to linguists, such as multi-verb constructions, valence frames, tense-aspect systems, infinitival and other hypotactic construction types (to just name some). At present TC hosts 2137 texts from 146 languages, corresponding to 5 million words.

In our first session we talk about general potential of user-driven databases and answer the question: What can you do with TypeCraft?. We use a tiny Bangla linguistic collection, annotated by Gautam Sengupta and Lars Hellan for illustration. There are other tasks that you can use TypeCraft for. In session 2 (tutorial. Get a LOGIN[1]), we will run an introductory tutorial introducing the basic steps needed to start a successful annotation project in TypeCraft.

Language archiving: Making documentation archivable

David Nathan
Batchelor Institute of Indigenous Tertiary Education and the EL Training Group, Batchelor
dn2@soas.ac.uk

The digital era offers excellent new ways to support the production and archiving of language documentation. Especially with the increased use of mobile devices, digital resources are potentially:

- Adaptable and repurposable
- Seamless across distance, languages, devices, and time
- Accessible to language communities as well as to privileged people (e.g. academics)

To actually realise these benefits, linguists' workflow must include these activities:

1. *Documentation* - Collect the best possible data and resources using good equipment and field methods.
2. *Data management* - Design data models and representations that reflect both the collection domain and best-practice methods, standards, formats etc.

1 *Login* instructions. From the TypeCraft Homepage at http://typecraft.org click on Request account, located on the grey bar at the very top right of the page. Please get a login latest one day before the event.

3. *Mobilisation* - Deliver useful materials to the language community honouring their goals and contributions. Documentation outcomes should be of value to linguists, community members, and others.
4. *Preservation* - Documentation records must be preserved for scientific, cultural and moral reasons.

These four activities are mutually dependent and reinforcing if done well. Mobilisation, preservation, and well-organised research all depend on data being well-designed, clear, explicit, technically standard and "agnostic". Therefore, data management is at the heart of language documentation workflow.

This tutorial will focus on data management and data preservation. Good data management fulfils many of the requirements of enabling and preparing for archiving. Until recently, this was widely misunderstood, because data management principles were typically only encountered for the first time when facing archiving.

Topics to be covered in the tutorial will be drawn from the following:

- Data domain modelling and representation
- File naming, file management, tips and tricks
- Inventories and metadata
- Choosing software
- Choosing an archive; local/personal archiving
- Access and distribution

In the hands-on component of the tutorial, groups of participants will work on a sample data set, design a data representation for a hypothetical domain, and use software to prepare an inventory and metadata.

A workbench for linguistic annotation and documentation of data

Anil Singh & Manish Kumar Singh
Banaras Hindu University, Varanasi
aksingh.cse@ iitbhu.ac.in & maneeshhsingh100@gmail.com

Introduction:
Linguistic annotation and documentation of human language data can be useful not only for the purposes of Computational Linguistics (CL) or Natural Language Processing (NLP), but also for linguistic analysis and even documentation of a language.

Linguistic annotation can be of various kinds and it is difficult to create separate tools for different kinds of annotation. If one uses only some text editor to create annotated data, it is almost impossible to ensure that the annotated data will be properly machine readable. A well designed annotation interface takes care of such problems and also makes it easier and faster to create annotated data. Inter-operability of different kinds of data is also a very important issue and an annotation interface should manage it seamlessly.

Sanchay (http://sanchay.co.in) is a collection of tools and APIs for working on natural language data. The most relevant tools in it for the purposes of this tutorial are the annotation interfaces. There are several annotation interfaces in Sanchay, but the most used out of these is the syntactic annotation interface. This interface allows syntactic annotation at various levels such as part-of-speech tagging, chunking, morphological analysis, phrase structure annotation, dependency annotation, i.e., the complete treebanking process. This interface is an easy to use interface that has evolved over the years based on the feedback provided by the users. For easy access of the annotated data, the interface has

search and statistics facilities. Sanchay also has its own concise and expressive query language that allows searching of complex linguistic patterns in the data and even allows modifying the data based on the conditions provided. The highlight of this interface is the easy drag-and-drop interface for dependency annotation. This interface (though called syntactic annotation interface) can also be used for some other purposes such as named entity annotation. For the IJCNLP 2008 workshop on named entity recognition for South and South East Asian languages, the corpus for some languages was prepared using Sanchay.

To ensure inter-operability of annotated data, Sanchay uses Shakti Standard Format (SSF) as the underlying representation scheme. SSF is human readable as well as machine readable and makes it possible to encode various levels of linguistic annotation at the same place. The query language in Sanchay works on this kind of data. Sanchay has an extensive API for handling this data, which can be used for building other NLP tools that are supposed to work on similar kinds of data. Another annotation interface in Sanchay is available for creating parallel corpus by aligning sentences as well as the words within the sentences. It is possible to perform some annotation even on the parallel data.

Methodology:
The tutorial will be a combination of lecture and demonstration. The participants will be encouraged to try using Sanchay hands-on.

Intended Outcome of the tutorial:
It is hoped that the tutorial will prepare the participants to be able to use Sanchay for the language of their interest and create linguistically annotated data for it. The participants can find out how to get Sanchay and use it for their projects.

The material taught could be useful for creating various kinds of linguistically annotated data such as building treebanks and other purposes such as named entity recognition. Sanchay has many other tools besides annotation interface and the participants can learn to use these tools to work on language data. After the tutorial, interested people who know programming can also find out (with help from the resource persons) how to use the Sanchay API for building some kinds of NLP tools.

As annotation interface, Sanchay is currently being used by several research teams around India to develop various kinds of linguistically annotated data. It is now almost ten years since its use started. The kinds of annotation include part-of-speech tagging, chunking, named entity recognition, dependency annotation and even complete complete treebanking. Sanchay is probably the most used tool for annotation of data for Indian languages.

As API, Sanchay is being used in the government sponsored consortia project for Inidan language to Indian language machine translation (ILMT). The resulting machine translation system from this project, called Sampark (http://sampark.iiit.ac.in/) is ready for six language pairs and many others are on the way.

Some researchers are also using parts of Sanchay for NLP research.

References:
Agarwal, Rahul & Ambati, Bharat Ram & Singh, Anil Kumar. 2012. A GUI to detect and correct errors in Hindi dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC).* Instanbul.

Bharati, Akshar & Sangal, Rajeev & Sharma, Dipti & Singh, Anil Kumar. 2014. SSF: A common representation scheme for language analysis for language technology infrastructure development. In *Proceedings of the COLING Workshop on Open Infrastructures and Analysis Frameworks for HLT*, 66–76. Dublin, Ireland.

Singh, Anil Kumar. 2014. A set of annotation interfaces for alignment of parallel corpora. *The Prague Bulletin of Mathematical Linguistics* 102. 57–68. (doi: 10.2478/pralin-2014- 0014.)

Singh, Anil Kumar. 2012. A concise query language with search and transform operations for corpora with multiple levels of annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Instanbul.

Singh, Anil Kumar. 2011. Part-of-speech annotation with Sanchay. In *Proceedings of the National Seminar On POS Annotation for Indian Languages: Issues & Perspectives.* Mysore, India.

Singh, Anil Kumar & Ambati, Bharat. 2010. An tntegrated digital tool for accessing language resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC).* Malta.

Singh, Anil Kumar. 2008. A mechanism to provide language-encoding support and an NLP friendly editor. In *Proceedings of the Third International Joint Conference on Natural Language Processing.* Hyderabad, India.

Singh, Anil Kumar. 2008. Named Entity Recognition for South and South East Asian languages: Taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.* Hyderabad, India.

Research Papers

Ethical and practical issues in documentation of Khortha language in Jharkhand

Atul Aman & Niladri Sekhar Dash
Linguistic Research Unit, Indian Statistical Institute, Kolkata
atul.aman1@gmail.com & ns_dash@yahoo.com

The present paper seeks to address some of the major ethical issues that have recently become factors of paramount importance in the task of language/dialect documentation. These issues involves various sensitive events and attitudes (e.g., like and dislike, opinions and views, preference and objection, etc.) of the informants which may have direct impact in the overall scheme of language documentation as well as in the act of inference deduction from the data and information obtained from the people involved in the survey. Since involvement of informants as one of the active agents in a documentation project has been accepted for ages, and since in recent times, it has raised various questions regarding the nature of involvement of informants as well as the nature of treatment of informants by the investigators, the issues of ethics and willful consent of the informants are now placed on a global plate for approval by the ethical committee formed for this purpose to safeguard the right of the people involved as informants as well as to protect the scheme of the works of investigators. Against this background it is absolutely necessary to look into the ethical factors and issues that are linked with documentation of Indian languages/dialects that are thriving at the verge of extinction. Our present paper is perhaps, the first of its kind in India that deals with some of the ethical and practical issues while the investigators are involved in collection of language data from Khortha - one of the least resourced languages spoken in several districts of the state of Jharkhand, India.

The paper first investigates the importance of ethics in linguistic field surveys as well scrutinizes the actual reactions (i.e., approval and disapproval) of the informants in the task. Also, it explores the possibility of allotting compensation to the informants, and devising methods for protecting appropriate rights of the informants as a part of a notable concern for protecting human rights of the informants. In principle, ethics in the area of linguistic surveys, is primarily concerned with what we perform on the field, i.e. the applied perspective. Creating maximum interest among the informants through developing amicable rapport with them may keep us ethically alert to carry out our linguistic surveys. In case of less resourced and less explored languages like Khortha, there is a necessity to focus on ethical factors in order to understand how the informants actually respond or react to any linguistic property or cultural element that is being questioned and recorded. Although it is true that some ethical factors are community-specific and culture-bound, in most cases, these are controlled and triggered based on the nature of study or purpose of a researcher.

We have experienced a notable change in the behavior and attitude of the native Khortha informants to certain questions and this has guided us to prepare an ethical framework on the basis of which it is safe to interact with the informants. The major challenge was to convince them for their precious time, and make them understand the purpose and importance of the work. Since, we covered remote villages along with major towns and cities, we had to deal with varied issues in comprehending the attitudes of the informants as well as interact accordingly. What was most striking is that ethical factors, even within the same speech community, varies based on certain demographic factors and

cultural backgrounds, such as, education, social awareness, sociopolitical background, economic status, gender, depth in cultural knowledge, ethnic orientation, etc

After analyzing the drawbacks of our initial survey we (including our institutional authorities) came up with an ethical framework by keeping all the loopholes of our survey in mind. Accordingly, we gave the documentation rights, moral rights, access and usage rights, and copyrights a special importance with giving emphasis to both the consent and their disagreement through a consent from the native informants. So there was a noticeable improvement in collecting complete rather than unfinished speech data. Apart from this we gave a special concern to compensate informants for their effort, time and contribution with adequate incentives. This also resulted as a profitable step in keeping the informants at their interest and accumulating a quality speech data without any discontinuation and pause.

It is true that the ethical code of conduct may vary among the different organization but the importance should be always given to the requirement of the research work and the area of the field survey. Since there is a big difference among the linguistic communities even at the nearby villages, towns and cities in their speech varieties, tolerance level, and nature of work, staying to the basic sense of the ethics we can give importance to the moral and universal values such as the need of informants, their rights, concern for health and safety. Although it can be said that drafting the points of ethics in any theoretical framework is an easy task, keeping these points alive in real practice is a tougher ball game.

The last part of the paper focuses on the methods and strategy adopted in collecting empirical speech data from the native Khortha informants. There are elaborate descriptions of the plan of work and the methodology adapted in conducting field surveys. It also highlights the attitude of the native speakers in order to describe their collective community desires for their native language. The theoretical relevance and practical importance of this paper may be attested in a language documentation frame where approval of the target speech community as well as of the competent informants is mandatory before conducting linguistic survey of any kind either for language documentation or for language description.

Khortha is considered as one of the least resourced languages in the IA family. Its marginal identity is observed in speech of the local people living in several districts of the state of Jharkhand, India. Although it is officially recognized as the second most populated language variety with 47,25,927 speakers in the state after Hindi (Census Report 2001), some scholars are willing to identify to as a dialect of Maithili (Prasad and Shastri 1958:9), while others are interested to refer it as a sub-variety of Hindi (Census Report 2001). The name ‘Khortha’ itself refers to a corrupt or impure form of a language (Prasad and Shastri 1958: 9). The language is widely spoken in the fifteen districts of the Jharkhand state. They are: Gadhwa, Palamu, Latehar, Chatra, Hazaribagh, Ramgarh, Bokaro, Dhanbad, Koderma, Giridih, Deoghar, Jamtara, Dumka, Godda, Pakud and Sahebganj (Ohdar 2012).

Keywords: Khortha, language documentation, informants, Ethics and Issues

References:

Alshenqeeti, Hamza. 2014. Interviewing as a data collection method: a critical review. *English Linguistics Research.* 3(1): 39-45.

Aman, Atul & Dash, Niladri Sekhar & Chakraborty, Jayashree. 2015. Investigating the patterns of syllable structures in Khortha as observed in spoken Khortha text. Presented at the *43rd All India Conference of Dravidian Linguists (AICDL-43)*, 18-20 June 2015, Annamalai University, Tamil Nadu, India.

Austin, Peter K. 2010a. Communities, ethics and rights in language documentation. *Language Documentation and Description*. 7(1): 34-54.

Austin, Peter K. 2010b. Current issues in language documentation. *Language Documentation and Description*. 7(1): 12-33.

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide.* New York: Palgrave Macmillan.
Crowley, Terry. 2007. *Field Linguistics: a Beginner's Guide*. Oxford: Oxford University Press.
Crystal, David. 2000. *Language death.* Cambridge: Cambridge University Press.
Dangi, Aanand Kishor. 2012. *Khortha Bhasha: Ek Parichay (Khortha Language: An Identity).* Jamshedpur, Jharkhand: Spardha Publications.
Dash, Niladri Sekhar. 2014. Language attitude of Khortha speakers in Giridih: a survey report. Presented in the *National Conference on Inter-disciplinary Researches in Social Sciences in Eastern India with special reference to Jharkhand (NCIRSSEI-2014)*, 27-28 February 2014, Sociological Research Units, Indian Statistical Institute, Kolkata and Giridih, Jharkhand, India.
Dash, Niladri Sekhar & Aman, Atul. 2013. An attempt towards documentation and preservation of the Khortha language of Jharkhand. Presented in the *2nd Seminar on Endangered and Lesser Known Languages (ELKL-II)*, 23-24th October 2013, Dept. of Linguistics, Lucknow University, Lucknow, India.
Dash, Niladri Sekhar & Aman, Atul. 2015. Generation of a dialect corpus in Khortha used in Jharkhand India: Some empirical Obsevations and theoretical postulations. *Journal of Advanced Linguistic Studies (JALS)*, Vol.4, No.1-2, Jan-Dec 2015, pp.151-16, ISSN: 2231-4075.
Labov, William. 1972. Some principles of linguistic methodology. *Language and Society.* 1(1): 97-120.
Ohdar, Bhawnath. 2007. *Khortha Bhasha Evam Sahitya: Udbhav EvamVikash (Khortha Language and Literature: Origin and Growth).* Ramgarh, Jharkhand: Khortha BhashaSahitya Academy.
Prasad, Bishwanath & Shastri, Sudhakar, 1958. *Linguistic survey of the Sadar sub-division of Manbhum and Dalbhum (Singhbhum)*. Patna: Bihar Rastrabhasha Parishad.
Samarin, William J. 1967. *Field Linguistics: A guide to Linguistic Field Work*. New York: Holt, Reinhart and Winston.

Issues and challenges in corpus collection and annotation of Sambalpuri: The case of a lesser-known language

Pitambar Behera
Jawaharlal Nehru University, Delhi
pitambarbehera2@gmail.com

Lesser-known languages are the languages that are less known in spite of huge number of population speaking because of their lack of advancement and empowerment in terms of technology. As rightly pointed out by Ostler (1993) languages that lack active participation in the electronic media are doomed to be endangered from the state of lesser-known. The situations of languages in South Asia in general and in Indic languages, in particular, are 'relatively bleak' (McEnry et al., 2000). Although India is a land of approximately 1500 languages, only 22 are scheduled and the rest are non-scheduled. Low-density languages have fewer resources in terms of the availability of voluminous corpus (McEnry et al., 2000) for NLP applications.

Two issues such as corpus collection and annotation have been taken up here. The unavailability of a corpus for a low-density language proves to bear adverse impacts on its future NLP development. If the data is available, it is either in the image form or non-Unicode encoding or not in machine readable format. Since these languages are not technologically empowered, there are not enough number of tools to convert the texts into Unicode. Owing to the facts that Sambalpuri being a lesser-known language and guidelines developed for Indian languages have been devised for only the scheduled languages, there are significant issues and challenges one faces with regard to annotation of a voluminous corpus. These underlying issues mostly pertain to the non-incorporation of unnoticed unique linguistic features of these languages into the uniform guideline devised for the scheduled languages.

The present paper brings out the issues and challenges underlying the collection and annotation of a voluminous corpus collected for Sambalpuri, a lesser-known language spoken in the eastern region of India (Kushal, 2015). It is an Indo-Aryan (IA) language otherwise known as Dom, Kosali, Koshal, Koshali, Western Odia[2]. It is spoken in the ten districts of western and south-western Odisha. The total corpus collected for it amounts to 121k tokens covering five domains: literature, sports, tourism, entertainment and miscellaneous.

Methodology:
The whole corpus has been annotated using the ILCIANN App. 2.0 (Kumar, et al., 2012) version following the BIS-ILCI tagset devised for Odia language since there is no tagset available for it. In addition, Sambalpuri was earlier considered to be one of the dialects of Odia and both are closely related morpho-syntactically but not syntactically. The BIS tagset (Baskaran et al., 2006) is a hierarchical set designed by the POS Standardization Committee appointed by the Department of Information and Technology, Government of India. The categories of reciprocal pronoun and foreign word have not occurred in the whole corpus during annotation job. The issues and challenges pertaining to corpus collection and annotation for Sambalpuri have been outlined below and the solutions have been proposed.

Issues and challenges:

1. Corpus collection (Behera et al., 2015)

- Unavailability of unicode text

The data in pdf format for Sambalpuri corpus has been collected[3] and converted into Unicode text[4] as the available text was not in a machine readable format.

- Non-standard usage

Sambalpuri is not a scheduled Indian language and is written and spoken with varying standards in different regions of the western and south-western Odisha. For example: Sambalpuri, Bargadia (spoken in Bargarh), Bolangiri/a (spoken in Bolangir district), Sundargadi/ia (spoken in Sundargarh), Deogarhia (spoken in Deogarh region) etc. The table (see Table 1) demonstrates dialectal variations of Sambalpuri with reference to negative morpheme ‘no’, adverbs ‘now’ and ‘this way’. Lexical similarity within the varieties of Sambalpuri is considerably high which ranges from 90 to 95 percent (Mathai & Kelsall, 2013).

Table 1: Dialectal Variations in Sambalpuri (Adapted from Patel)

Variety of Sambalpuri	Negative Morpheme [*naĩ*] ‘no’	Adverb [*ɪhaɖ̪e*] ‘now’	Adverb [*ɪaɖe*] ‘this way’
Bargarh	*nʊhe/nɪhe*	*ɪhaɖ̪e/ɛcʰɛn*	*ɪaɖe/ɪpʰale*
Bolangir	*nĩ*	*ɛkʰɛn*	
Deogarh			
Kalahandi	*nĩ*	*ɛkʰɛn*	*ɪbaʈe*
Sambalpur	*nɪhe/nʊhe*		
Sundargarh		*ɪgədɪ*	

- Different orthographic conventions

A large number of words in Sambalpuri has different orthographic conventions; especially the ligatures. In Sambalpuri, there are several writing conventions used for a given word form because of

2 http://www.ethnologue.com/language/spv
3https://koslisahitya.wordpress.com/
4https://22bc339da9ca3e2462414546a715752e4c2c5e0d.googledrive.com/host/0B5rBGd680WZFemVLa3RxY0preE0/AkrutiUnicode

the non-uniform usage of language. For instance, in the following examples two forms are used for one word with two of them having different POS labels with the change of form.

କାଞ୍ଜେଁ N_NN, କାଁଜେ DM_DMQ
କନ୍ଡକ୍ଟର N_NN, କଣ୍ଡକ୍ଟର୍ N_NN
କାନ୍ଚନ N_NN କାଞ୍ଚନ N_NNP

- Hindi-like constructions

In the examples instantiated below /*bɑʊəɟʊɖ̚*/ and /*ke*/ are postpositions as used in Hindi while the Hindi-like indefinite and reflexive pronouns are also used. For instance,

/bɑʊəɟʊɖ̚/ PSP
/*hərek*/ PR_PRI /*ke*/ PSP
/*əpna*/ PR_PRF /*əpnar*/ PR_PRF

2. Corpus annotation

2.1. Issues in framing Tagsets:

As has been discussed by Chandra et al. (2014), there are many significant issues in designing a standard tagset for any parts of speech annotation. There are linguistic issues as bulleted in the following:

- Finiteness vs coarseness

This issue deals with tagset designing where one faces challenges whether to create flat one or hierarchical one. For example, verb having one tag (VB) is an example of flat tagset whereas verbs having labels for their types (finite, non-finite, main, auxiliary, infinite etc.) is considered as a hierarchical tagset. BIS tagset strikes a balance between the flat and hierarchical nature; so as tagset devised for Sambalpuri.

- Morphological vs syntactic

It deals with the annotation issue and the agreement whether to go for the morphological or syntactic approach. In Indian languages there are a large number of adverbial constructions where one faces conflicts and disagreement. Annotating in either of the ways seems to be compromising with the linguistic information about the given structure. In this study, morphological-contextual information has been considered.

Morphological approach

If one goes for this approach, one needs to take into consideration the morphological information. Thus, in the following example, /*bʰɔlbʰɑbe*/ is an adverbial modifier which modifies the finite verb and should have RB tag. Contrastingly, they have the tags of adjective followed by postposition. For instance:

bʰɔl\JJ *bʰɑbe*\PSP *kɔrbɔ*\V_VM_VF ‘Do it properly.’

Syntactic approach

The same phrase is labelled as adverbial which results in misinformation too.

bʰɔl\RB *bʰɑbe*\RB *kɔrbɔ*\V_VM_VF ‘Do it properly.’

- New tags vs existing tags from a tagger

While designing a new tagset one considers to create an entirely one or to modify an existing tagset to make it suitable for their own. In the present study, an already existing tagset devised for Odia, which is closely related to Sambalpuri, has been used.

2.2. Issues in linguistic annotation (Behera et al., 2015)

- Reduplication

Generally, in Indian languages the reduplicated expressions follow the meaningful word (Abbi, 1992). Contrastingly, in Sambalpuri many of the reduplicated parts precede the meaningful words (see section 3.3). For instance, in the conjunct verb (adjective + finite verb), /*cʰɪcʰɪ*/ is the meaningless reduplicated part which is preceding the meaningful part /*bɪcʰɪ*/ ‘scattered’. For example,

cʰɪcʰɪ\RD_ECH *bɪcʰɪ*\JJ *heɪcʰɔn* ‘have got scattered’

Similarly, in the following verbal reduplication, the meaningless part is preceding the verbal part.
For example,

kʊʈ̪\RD_ECH *kʊʈ̪eɪ*\V_VM *dela*\V_VM_VF ‘has tickled’

These kinds of constructions pose significant linguistic challenges for the human annotators as to how to label them and so is for the statistical tagger.

- Agglutination of classifiers with postpositions

Agglutination is one of the common features in Odia (Behera, 2015 & Jena et al., 2011) and Sambalpuri along with some IA languages like Bengali and Marathi (Baskaran et al., 2008). In Sambalpuri, one of the peculiar constructions with agglutination is that the classifiers and postpositions agglutinate with each other which is also rarely found in the most agglutinating Dravidian languages. Here, to annotate these constructions as classifiers (RP_CL) or postpositions (PSP) is quite difficult. For instance,

/*bagɪr-ʈa*/ ‘as-CL’
/*lekʰeʈa*/ ‘like-CL’

- Spatial-temporal nouns and adverbs

/*kɛbʰe kɛbʰe*/ ‘sometimes’ is an expression which is an adverb of frequency in Sambalpuri. If we adhere to the BIS guideline we observe that they are to be annotated as spatial-temporal nouns N_NST. Thus, to annotate these elements properly becomes a tricky issue for the annotators.

- Serial verbs

In the following instance, *kʰʊa*\V_VM is the meaningless, verbal, partial and reduplicated part of the meaningful preceding verb. The issue is how to annotate /*kʰʊa*/ as annotating it an echo or main verb or auxiliary verb proves it to be a wrong judgement.

sekʰaɪ\V_VM *kʰʊa*\V_VM *kɔrɪ*\V_VM_VNF *asla*\V_VM_VF
‘He came after eating.’

- Multi-words (reduplication, onomatopoeic words, echo-words, idioms)

Onomatopoeic words are the imitation of a sound associated phonetically with its describing referent. These following expressions are parts of the multi words because individually these words do not have meaning, but when combined they are manner adverbs. As per the ILCI guideline, if we annotate

the first sound as noun and the following words as echo-words (RD_ECH), we are missing relevant linguistic information. For instance,

bʰɛ̃ bʰɛ̃ 'loudly'
ɟʰɒ ɟʰɒ 'heavily'
d̪ʰɔ̃ pɔ̃ 'gasping'
b̪ʰɒ bʰɒ 'bark'

- Punctuations

Punctuations have several functions other than to punctuate (Vaz, 2012). For example, colons are used as conjunctions, section headers and may be others. Similarly, the punctuation mark ':-' in Sambalpuri also functions as conjunctions and section headers other than its usual function.

- Ambiguities

Ambiguities at the POS level are owing to the homographic nature of words having different function labels. The verbal word form /kərɪ/ has more than three labels in the corpus and which is rightly so. It can be used as main, auxiliary, finite and non-finite verbs as instantiated in the following examples.
kərɪ (V_VAUX or V_VM or V_VM_VF or V_VM_VNF)
For instance,

kʰaɪ\V_VM *kərɪ*\V_VAUX *asle*\V_VM_VF
kam\N_NN *kərɪ*\V_VM *asle*\V_VM_VF
kərɪʈ̪ʰɪla\V_VM_VF
kʰaɪkərɪ\V_VM_VNF *asle*\V_VM_VF

Solutions proposed:

- Tagsets proposed

The tag sets proposed is the BIS tagset devised under the ILCI Project which a combination of both flat and hierarchy.

- Dealing with spatial-temporal nouns and adverbs

The nouns referring to spatial and temporal sense are annotated as spatial-temporal nouns N_NST whereas adverbs of frequency and manner adverbs have been annotated as RB.

- Tackling morphophonemics

The sandhi in the Indian phenomenon has been decided based on the right-headedness feature of the language. As Sambalpuri is a right-headed language, the label has been decided on the basis of the tag of the right placed word as in the following example.
sɔʈ̪\JJ + *pɔʈ̪ʰɔ*\N_NN = *sɔd̪pɔʈ̪ʰɔ*\N_NN

- Handling prefixes and suffixes

The typical prefixes and suffixes for different categories have been marked manually for each of the categories to annotate on the basis of the list. This has been done keeping in view the automatic annotation by the tagger in future.

- Multi-words

Multi-words have been annotated morphologically and contextually as there are several sub-categories.

- Incorporating labels for demonstratives and pronouns

One of the suggestions is to incorporate the labels for possessive demonstratives and pronouns which are missing in the BIS.

- Compound proper nouns with hyphenation

These types of nouns are generally foreign words transliterated into Sambalpuri that cause challenges during manual annotation. Thus, to address this issue one can create a special tag N_NNPC.

John-in-the-Wilderness\N_NNP
Kɪṭab-e-hɪnḍ\N_NNP

- Linguistic word sense disambiguation

For manual disambiguation, the morphological information has been taken into consideration. For automatic disambiguation, one can take recourse to different statistical models.

Reference:

Abbi, A. 1992. *Reduplication in South Asian languages: An areal, typological, and historical study*. Allied Publishers Pvt. Ltd, India.

Baskaran, S & Bali, K. & Bhattacharya, T & Bhattacharyya, P. & Jha, G. N. 2008. A common parts-of-speech tagset framework for indian languages. In *In Proc. of LREC.*

Behera, P. 2015. *Odia parts of speech tagging corpora: suitability of statistical models*. New Delhi: Jawaharlal Nehru Univerity. (M.Phil. Thesis.)

Behera, P. & Ojha, A. K. & Jha, G. N. 2015. Issues and challenges in developing statistical POS taggers for Sambalpuri. In *Proc. LTC-2015*, Poland: Springer (accepted).

Chandra, N. & Kumawat, S. & Srivastava, V. 2014. Various tagsets for Indian languages and their performance in part of speech tagging. In *Proceedings of 5th IRF International Conference*. Chennai.

Jena, I. & Chaudhury, S. & Chaudhry, H. & Sharma, D. M. 2011. Developing Oriya morphological analyzer using Lt-toolbox. In *Information Systems for Indian Languages*,124-129. Berlin: Springer.

Kumar, R. & Kaushik, S. & Nainwani, P.& Banerjee, E. & Hadke, S. & JHA, G. N. 2012. Using the ILCI annotation tool for POS annotation: A case of Hindi. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012), New Delhi, India*.

Kushal, G. (2015). *Case and agreement in Sambalpuri*, New Delhi: Jawaharlal Nehru Univerity. (M.Phil. Thesis.)

Mathai, E. K. & Kelsall, J. 2013. *Sambalpuri of Orissa, India: A brief sociolinguistic survey*. SIL International.

McEnery, T. & Baker, P. & Burnard, L. 2000. Corpus resources and minority language engineering. In *LREC*.

Ostler, N. 1999. Language technology and the Smaller Language. *Elra Newsletter*, *4*(2).

Patel, Kunjabana. (undated). *A Sambalpuri Phonetic Reader*. Sambalpur: Menaka Prakashani.

Vaz, E. & Walawalikar, S. V. & Pawar, J. & Sardesai, M. 2012. BIS annotation standards with reference to Konkani language. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)*, 145-152. COLING. Mumbai

Developing an automated SVM POS tagger for Sambalpuri: The case of a lesser-known language

Pitambar Behera & Atul Kumar Ojha
Jawaharlal Nehru University, New Delhi
pitambarbehera2@gmail.com & shashwatup9k@gmail.com

Language Technologies (LT) play a vital role in making a language available for the perusal of a larger public which ultimately empowers its status at several levels. In order to develop LT for a lesser-known, low-density, poorly-described, less-resourced, minority or less-computerized language,

a large annotated and supervised corpus is indispensably desired. Considering the situations in non-scheduled (lesser-known) Indian languages it is quite unfavorable in comparison to the scheduled languages since some of the Indian institutions have either worked on or are presently developing language resources and technologies for these languages. Because of the indifference of the government towards the lesser-known languages and financial patronage to only the scheduled languages, the former are getting disempowered gradually. As outlined in (McEnery et al., 2000), Ostler (1999:3) says "languages which do not take a full part in the electronic media are doomed to stagnate, if not atrophy". Therefore, the need of the hour is the large-scale development of the language resources and technologies for these languages.
There are several issues ranging from heterogeneous language standards to developing tagset and annotating at the linguistic levels. First, since India has approximately 1500 languages and out of them only 22 are scheduled and the rest among them are non-scheduled, it is obvious that only the former are financially patronized in all matters while the latter are not. Thus, the lesser-known languages are spoken and written with a non-standardized form which ultimately poses challenges in a machine-readable corpus by affecting the accuracy rates of the NLP applications. Second, the tagsets developed for Indian languages are meant for only scheduled languages, for instance, ILMT, IL-POSTS, LDC-IL and BIS. Since these languages are less-described, these former languages may have interesting unnoticed linguistic structures and may not be incorporable in these tagsets. Finally, owing to the fact that lesser-described languages have several interesting linguistic structures, there are significant challenges pertaining to their linguistic nature.

The envisaged research presents an SVM parts of speech tagger for Sambalpuri. It is an Indo-Aryan (IA) language otherwise known as Dom, Kosali, Koshal, Koshali, and Western Odia [5]. It is spoken in the ten districts of western and south-western Odisha. The paper presents the issues in corpus collection for a less-resourced language and salient features of the collected corpus. Furthermore, the training, testing and development data sets, the experimental set-up, the accuracy rates (overall, per POS categories, ambiguity sets) and word sense disambiguation of the conflicting sets have been vividly dealt with.

The data demonstrated below represents the overall accuracy, accuracy per POS categories and ambiguity sets. The corpus collected for the Sambalpuri language amounts to 121k tokens in totality extended to almost all the domains. The data has been collected from the online webpage [6] in pdf version and converted into UTF-8 on the Unicode Converter[7] for Odia as the scripts for both the languages are the same. The whole corpus has been annotated with the BIS [8] (Baskaran et al., 2006) standard annotation scheme devised for Odia under the ILCI Project on the ILCI ANN App 2.0 version (Kumar, et al., 2012) semi-automated tool. The tagger has been trained and tested with 80k and 13k tokens respectively with a development set of 30k tokens. The system has been trained with SVM algorithm (Giménez & Màrquez, 2006) and provides around 83% accuracy.

Salient linguistic features:
These features are taken from Behera et al., 2015.

3.1. Agglutination:
In the examples instantiated below, all the case endings or markers /*ke*/, /*ɔr*/, /*nʊ*/ agglutinate with their head categories verb, noun and pronoun. /*ke*/ which is equivalent to the English infinitive and preposition is alternating here with both verb and spatial noun /*kʰaɛbar*/ and /*bahar*/ respectively. For instance, *kʰaɛbar-ke* 'to eat', *lʊk-ɔr* 'peoples', *bahar-ke* 'to outside', *mɒr-nʊ* 'from me'.

5[4] http://www.ethnologue.com/language/spv
6[5] https://koslisahitya.wordpress.com/
7https://22bc339da9ca3e2462414546a715752e4c2c5e0d.googledrive.com/host/0B5rBGd680WZFemVLa3RxY0preE0/AkrutiUnicode
8 BIS scheme is a standard annotation scheme for Indian languages prepared by the POS Standardization appointed by the DeiTY.

3.2. Classifiers:
It is a dominant linguistic feature in Sambalpuri as well. The classifiers mainly occur either as proper classifiers, attached to numerals or to the quantity word */keʈe/* 'how many; some', or as indefinite markers, in combination with the suffix */-e/* (Neukom, 2003) as */ʈe/, /ʈa/, /ʈe/, /ʈa/, /kʰɔ̃ɖe/, /ɟʰɔne/, /ʈɪ/* etc. in Sambalpuri. One of the rarely observed phenomena of Indian languages found in Sambalpuri is that classifiers also occur with post positions. For instance, */mɒrlekʰe-ʈa/* 'like me'.
/kʰɔɖe/, /ɟʰɔne/, /ʈɪ/ etc. in Sambalpuri. One of the rarely observed phenomena of Indian languages found in Sambalpuri is that classifiers also occur with post positions. For instance, /mɒr lekʰe-ʈɑ/ 'like me'.

3.3. Reduplication:
In the following instances, the first two are fully reduplicated while the rest of the following are partial. In the partial reduplication, the final syllables */nɑ/* of both the words are reduplicated like in the first example whereas the final example contains the reduplications of the initial syllables */hʊ-/*. For instance,

çɪk çɪk 'shining', *ɖʰɪre ɖʰɪre* 'slowly'
ɟɔnɑ sʊnɑ 'known', *hʊɖɑ hʊɖɪ* 'abusing'.

Serial verbs are reduplicative in nature in Sambalpuri. In the below-instantiated example, it is quite confusing as to how to annotate the verbal occurrences. Because, the initial verb */nɔrɖɪ/* is a non-finite verb followed by a verbal reduplication which is behaving like a manner adverb modifying the finite following verb */pɔleɪgɔlɑ/*. The issue here is how to annotate the verbal reduplication. For example,

se mɑrkɔrɪ nɔrɖɪ nɔrɖɪ pɔleɪgɔlɑ
he V-Nonfinite V-reduplication V-Finite
'He went away beating.'

3.4. Compounds:
In the following instance, the first word is an adjective and the second is a noun but when they get combined comprise a nominal category. Since Sambalpuri a head final language, the annotation label is decided on the basis of the category of the head. Here the head is a nominal element and hence the judgment goes in favor of the category of the label of the head word. For example,

sɔɖ \JJ + *pɔʈʰɔ*\N_NN = *sɔʈpɔʈʰɔ*\N_NN 'good path'

So, in the above example, the decision whether to annotate the word as an adjective or noun, goes for the right-headedness feature of Sambalpuri. This feature is typical to most of the IA languages and the word */sɔʈpɔʈʰɔ/* is labelled as a noun.

4. Methodology:
This section deals with (a) the total corpus collected in four major domains, (b) the BIS annotation guideline adapted for Sambalpuri, (c) size of the corpus for training, testing and development stages (d) features selection for SVM and CRF++ POS taggers.

4.1. Corpus size:
The tabulated data (see Table 1) demonstrates the total corpus size collected for the development of the Sambalpuri POS taggers. The whole corpus size comprises of five domains, viz. literature, sports, tourism, entertainment, and miscellaneous. The highest corpus size is registered in the domain of entertainment i.e., approximately 40k while the 'miscellaneous' section accounts for the lowest number of data.

Table 1: Total corpus size domain-wise

Domains	**Tokens**	**Data set**	**Tokens**
Literature	30, 344	Training	80, 288
Sports	21, 121	Development	30, 022
Tourism	26, 767	Testing	12, 791
Entertainment	40, 554		
Miscellaneous	2, 424		
Total	1,21,210		

4.2.Corpus Annotation:
The whole Sambalpuri corpus is annotated using the ILCIANN2 App. 2.0 version (Kumar et al., 2012)

4.3 Experimental Set-up:
The below figure explains the features selection for SVM tagger which takes into account the word, POS, ambiguity and may be's. Behera et al., (2015) discusses that learning phase contains medium verbose (-V 2) and the mode of learning and tagging is set to left-right-left (LRL). The rest of the features such as sliding window, feature set, feature filtering, model compression, C parameter tuning, Dictionary repairing and so on are set to default.

word features	$w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$
POS features	$p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}$
ambiguity classes	a_0, a_1, a_2, a_3
may_be's	m_0, m_1, m_2, m_3

Figure 1: Template for feature selection of SVM

The tabulated data demonstrates the accuracy rates of each of the POS categories by the SVM POS tagger with recall (71%), precision (83%) and F-measure (76%). It further suggests that categories such as relative pronoun, verbal noun, echo-words, gerundive verbs have registered less accuracy whereas POS categories like interrogative pronoun, punctuation and symbol have highest amount of accuracy. The lesser accuracy rates in categories like verbal noun and gerundive verbs can be accounted for the fact that the tagger is not able to distinguish between these types of words.

Table 2: Accuracy per POS category

POS	**Recall**	**Precision**	**F-measure**
CC_CCD	90.338165	83.48214	86.77494
CC_CCS	76.99115	70.731705	73.72881
DM_DMD	76.83398	95.67308	85.22484
DM_DMI	82.08955	98.21429	89.43089
DM_DMQ	82.10526	86.666664	84.32432
DM_DMR	83.63636	93.877556	88.46154
JJ	49.619484	76.16823	60.09217
N_NN	88.107346	77.877655	82.67727
N_NNP	56.19048	43.703705	49.16667
N_NNV	14.285715	57.14286	22.85714
N_NST	81.48148	93.333336	87.00565
PR_PRF	77.41935	96.0	85.71428
PR_PRI	73.46939	83.72093	78.26087
PR_PRL	44.444447	66.66667	53.33334
PR_PRP	89.0625	85.84337	87.42331
PR_PRC	Gold data not found	**	**

PR_PRQ	100.0	100.0	100
PSP	91.31313	91.68357	91.49798
QT_QTC	75.419	80.838326	78.03469
QT_QTF	69.03226	88.429756	77.53623
QT_QTO	71.42857	76.92308	74.07407
RB	66.46342	87.2	75.43253
RD_ECH	21.95122	75.0	33.96226
RD_PUNC	98.40357	98.84541	98.624
RD_SYM	97.72727	100.0	98.85057
RD_UNK	77.77778	29.166666	42.42424
RP_CL	82.02247	97.333336	89.02439
RP_INJ	50.0	92.85714	65
RP_INTF	28.57143	83.33333	42.55319
RP_NEG	99.41521	91.89189	95.50562
RP_RPD	95.97701	98.81657	97.37609
V_VAUX	35.0	100.0	51.85185
V_VM	72.935776	81.53846	76.99758
V_VM_VF	89.95046	83.78378	86.75768
V_VM_VINF	59.740257	81.415924	68.91385
V_VM_VNF	61.695904	77.85978	68.84176
V_VM_VNG	31.25	76.92308	44.44444
TOTAL	70.61526	83.415063581	76.48334

The data below states that the classes of ambiguity by the tagger have been categorized into three major classes. The two-class ambiguity is the category where the machine does not distinguish between coordinating and subordinating conjunctions and other such categories. Similarly, in the three-class, it confuses with negative particle, main and finite verb. In the more than three-class category, verbal classes are a matter of concern for the tagger.

Example of a four-class ambiguity,
One word form can be used as main, auxiliary, finite and non-finite verbs as instantiated in the following examples.

kərɪ (V_VAUX or V_VM or V_VM_VF or V_VM_VNF)
For instance,
kʰɑɪ\V_VM kərɪ\V_VAUX ɑsle\V_VM_VF
kɑm\N_NN kərɪ\V_VM ɑsle\V_VM_VF
kərɪʈʰɪlɑ\V_VM_VF
kʰɑɪkərɪ\V_VM_VNF
ɑsle\V_VM_VF

Table 3: Ambiguity Classes

Classes of Ambiguity	Label Sets
2 Sets:	CC_CCD_CC_CCS, CC_CCD_QT_QTF, DM_DMD_DM_DMQ
3 Sets:	JJ_N_NST_V_VM_VF, RD_ECH_V_VM_V_VM_VNF, RP_NEG_V_VM_V_VM_VF
More than 3 Sets	V_VAUX_V_VM_V_VM_VF_V_VM_VNF, RP_INJ_RP_NEG_V_VM_V_VM_VNF, RD_UNK_RP_INJ_RP_RPD_V_VM_V_VM_VN

	F

Results of the study show that Sambalpuri SVM tagger (83) does not perform well in comparison to Odia (94%). The significant difference can be accounted for the fact that unlike Odia, Sambalpuri is non-standard and has several orthographic and spoken varieties.

Labels for reduplication (RD_REDP), possessive pronouns (PR_POS) and demonstratives (DM_POS), interrogative adverbs (WRB) can be introduced. For handling agglutination, a stemmer or a lemmatizer could be used with statistical POS taggers. For punctuations, fine-grained labels should be incorporated based on their functions in a given context as they can be used as coordinators, section headers, list item markers and so on.

References:

Baskaran, S. & Bali, K. & Bhattacharya, T. & Bhattacharyya, P. & Jha, G. N. 2008. A common parts-of-speech tagset framework for Indian languages. In *Proc. of LREC*.

Behera, P. 2015. *Odia parts of speech tagging corpora: Suitability of statistical models*. New Delhi: Jawaharlal Nehru Univerity. (M.Phil. Thesis.)

Behera, P. & Ojha, A. K. & Jha, G. N. 2015. Issues and challenges in developing statistical POS taggers for Sambalpuri. In *Proc. LTC-2015*, Poland, Springer.

Giménez, J. & Màrquez, L. 2006. *Technical manual v1.3*. Universitat Politecnica de Catalunya, Barcelona.

Joachims, T. 1999. *Making large scale SVM learning practical*. Universität Dortmund.

Kumar, R. & Kaushik, S. & Nainwani, P. & Banerjee, E. & Hadke, S. & Jha, G. N. 2012. Using the ILCI annotation tool for pos annotation: A case of Hindi. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012)*, New Delhi, India.

McEnery, T. & Baker, P. & Burnard, L. 2000. Corpus resources and minority language engineering. In *LREC*.

Neukom, L. & Patnaik, M. 2003. *A grammar of oriya*. Seminar für Allgemeine Sprachwissenschaft der Univ. Zürich.

Ostler, N. 1999. Language technology and the smaller language. *ELRA Newsletter*, 4(2).

An account of personal pronouns and pronominal agreement in Vaiphei

Nandini Bhattacharya
University of Delhi, Delhi
nandini.poetry@gmail.com

Vaiphei is a language of the Kuki sub-group of the Tibeto-Burman sub-family of languages. This North-eastern language is spoken in the Churachandpur district of Manipur, and in pockets of Assam, Meghalaya, Mizoram, and Nagaland, as well as in the Kabaw Valley and Chin State of Myanmar. Vaiphei is one of the 100 non-scheduled languages in India and also an endangered North-eastern language (2001 census data). Vaiphei is the name of the tribe who speaks this language. The basic word order is SOV in Vaiphei. In this language, there are distinct forms of personal pronouns [i.e. /*kei*/ (1st), /*naŋ*/ (2nd) etc.] as well as there are distinct pronominal agreement markers [i.e. /*ka*/ (1st), /*la*/ (2nd), /*aʔ*/ (3rd)] that occurs within the verb phrase. The pronominal agreement markers occur as a prefix to the main verb.

However, all of the personal pronoun forms are optional in the subject position and the person feature is encoded by the pro-markers. This suggests that Vaiphei is a pro-drop language. For example,

(1.) *zapoaŋ pani hi la o̱oi hi*
clothes both PROX 2nd like AUX
'You like both of the clothes.'

(2.) *aʔ hoi puai*
3rd good NEG.SG
'It is bad.'

The third person singular pronoun is a composite DP, containing obligatory deictic marker and determiner, for example: /*zepahi*/ [3rd.sg.mas]. In contrast, the plural 3rd person pronoun has a distinct form, for example: /amao/ [3rd.plu.mas.]. The personal pronoun paradigm can be observed in the following examples:

(3.) *zanixan kei ka t̪ai hi*
yesterday I 1st run AUX
'I ran yesterday.'

(4.) *zanixan naŋ la t̪ai hi*
yesterday you 2nd run AUX
'You ran yesterday.'

(5.) *zanixan ze pa hi aʔ t̪ai hi*
yesterday DET MAS PROX 3rd run AUX
'He ran yesterday.'

Moreover, the occurrences of personal pronoun forms are optional, whereas, the pronominal agreement markers are obligatory. This phenomenon is illustrated in the following example:-

(6.) *zanixan t̪ui ka d̪on hi*
yesterday water 1st drink AUX
'I drank water yesterday.'

Another essential feature is clusivity that can be observed in Vaiphei. This distinction is drawn by the 1st person plural pronominal agreement marker. The pronominal agreement marker is /*ka*/ for exclusive "we" and is /*i*/ for inclusive "we". Moreover, the split-ergative- case marking feature can also observed in Vaiphei. The ergative case marker is /*n*/. The ergative marking can only be observed in 3rd person pronouns.

Therefore, this paper gives an adequate account of the personal pronoun paradigm of Vaiphei and provides insight into the pronominal agreement feature, as well. This paper thus demonstrates a descriptive morphological account of the personal pronouns in Vaiphei.

References:

Abbi, Anvita. 2001. *A manual of linguistic fieldwork and structures of Indian languages.* München, Germany: Lincom Europa.

Bauman, James. 1975. *Pronouns and pronominal morphology in Tibeto-Burman*. California: University of California. (PhD Dissertation.)

Bauman, J. 1974. Pronominal verb morphology in Tibeto-Burman. In *Linguistics of the Tibeto-Burman Area* 1. 108-155.

DeLancey, S. 2010. Towards a history of verb agreement in Tibeto-Burman. In *Himalayan Linguistics* 9.1.

Lust, Barbara. & Wali, K. & Gair, J. & Subbarao, K.V. (eds). 2000. *Lexical anaphors and pronouns in selected South Asian languages : A principled typology.* Berlin: Mouton de Gruyter.

LaPolla, Randy J. The inclusive-exclusive distinction in Tibeto-Burman languages. In Filimonova, E. et al. (eds). 2005. *Clusivity: Typology and case studies of the inclusive-exclusive distinction.* Amsterdam: John Benjamins.

Payne, Thomas E. 1997. *Describing Morphosyntax.* Cambridge: Cambridge University Press.
Shophen, Timothy. et al. (eds). 2007. *Language typology and syntactic description.* Cambridge: Cambridge University Press.
Siewierska, Anna. 2003. Person agreement and the determination of alignment. In *Transactions of the philological society* 101.2: 339-370.
Thurgood, G. & Lapolla, Randy J. (eds). 2003. *Sino-Tibetan Languages*. London & New York: Routhledge Taylor & Francis group.
http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement8.aspx

Historical relationship among Great Andamanese languages

Mayank Jain
Jawaharlal Nehru University, New Delhi
jnu.mayank@gmail.com

Great Andamanese tribes have been living in the Andaman Islands since time immemorial. Some genetic and linguistic studies claim it to be present in the Islands for 70,000 thousand years. Until very recently till the mid of 19th century, before the British occupation of the islands, Great Andamanese tribes were constitute of 10 different tribes speaking 10 different languages inhabiting whole of the Great Andaman islands. These 10 languages (Sare, Kora, Bo, Jeru, Kede, Kol, Juwoi, Pujjikar, Bale, Bea) constituted Great Andamanese language family which spread across these islands. Each individual community used to live in seclusion from each other with limited interaction among the neighbouring communities. They were hunter and gatherer communities who relied on forest and sea for their living. The linguistic scenario changed completely since the outsiders occupied these Islands. The British penal settlement in the Andaman Islands leads to their brutal interaction with the tribes which resulting in the dangerous dwindling in the tribal population, which reached at the alarming level during the first half of the twentieth century. It further reached its lowest count in the early 1970s which forced the Indian government to put them in Strait Island near Port Blair with government aids which helped in improving their numbers.

Present Great Andamanese language (PGA) is a moribund language spoken by a few (4-5) of the elderly members of a small Great Andamanese Negrito community, around 50 (Abbi, 2006). Among those who can speak PGA, none can speak it fluently and use it as a secret language to give them privacy from outsiders. They cannot frame sentences in PGA but rather code mix with the state language Hindi. The community has been completely shifted to the state language 'Hindi'. An Andamanese variety of Hindi is used for day to day communication within and outside the community. Youngster and children are completely shifted to Hindi as medium of education and instruction is Hindi (Abbi, 2006). The present Great Andamanese community consists of representation from the 10 different languages once spoken in the Island. The reduction in their population had prompted marriages among those who speak different Great Andamanese language varieties. This has created a situation of interaction among different languages. Due to such interactions, there has been found much variation among the speech of the speakers of PGA, which is manifested at each level of linguistic analysis particularly at the phonetic level. Many speech sounds are found in free variation. It is a linguistically stimulating question to find the relationship of 10 previously spoken Great Andamanese languages with PGA. The present paper tries to discover such relationship between now extinct Great Andamanese languages and PGA. The study employs the comparison of the cognates of PGA and four (Sare, Kede, Pujjikar and Bea) of the ten previously spoken Great Andamanese languages. Out of these four languages, Sare is a north Andaman language, Pujjikar and Kede are central whereas Bea is the south-most Great Andaman languages. The study shares some interesting results explaining the relationship of these languages with PGA. For instance, it has been found out that Bea has least affinity with PGA. This may be attributed to the fact that Bea speaking Great Andamanese community, as they were living near Port Blair was the first who came in unfriendly contact with the British. This community has to bear most of the atrocities of the British

which resulted in wiping out of their population. In contrast, PGA shares closer affinity with Great Andamanese languages spoken in the northern and central Andaman Island. Sare, a north Andamanese language and Kede, a central Andamanese language have been closer to PGA.

This study also discusses the sound change in the PGA lexicon. It has been observed that PGA lexicon has some systematic sound distinction with the lexicon of the two languages used in the study (Kede and Sare). It has been found by comparing cognates of these languages and PGA, for instance in some cognates word final [b] in Sare and Kede is devoiced to [p] in PGA. There are several such instances which show similar sound change among these languages.

References:

Abbi, Anvita. 2006. *Endangered Languages of the Andaman Islands*. Lincom-Europa: Munich.

Abbi, Anvita. 2008. Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? *Language Sciences*. (doi:10.1016/j.langsci.2008.02.002)

Abbi, Anvita. 2012. *Dictionary of the Great Andamanese Language. English-Great Andamanese-Hindi.* Ratna Sagar: Delhi.

Mayank. 2009. *Comparative Lexicon of Great Andamanese Languages*. New Delhi: Jawaharlal Nehru University. (M.Phil. Dissertation.)

Portman. M.V (1898) 1992 (reprint). *Manual of the Andamanese Languages*. Delhi: Manas Publications.

Radcliff-Brown, A.R. 1922. 1948 (3rd print). *The Andaman Islanders*. Glencoe, Illinois: Free Press.

The relevance of dialect planning in Indian context: Situation of languages of Bihar

Chandan Kumar
Jawaharlal Nehru University, New Delhi
chandajnudelhi@gmail.com

The paper is going to raise and discuss some of the crucial questions concerning language loss and maintenance like the question of language loyalty (Joshua Fishman 1966), identity, language choices and language need, question of language right and scope of its use. The paper is also an effort to see the implication and limitation of the article, written by Bernini (2014) titled ‘Preserving languages beyond the political dimension: Some proposals for a dialect planning’, in the Indian context.

India being the home of multilingualism lacks the official profiling of many languages/ dialects. It is matter of debate whether all dialects should be politically/ officially recognized, given the political scenario and societal attitude of the country. The article written by Bernini talks about the situation of dialects that are categorized as sub-languages whose functional and formal domain in the society is limited or absence. It has suggested some ideas to safeguard endangered languages that claim to be more effective and fruitful. Though, the suggestions provided by Bernini sound good lacks the practical aspects of sociolinguistic. This might be because every sociolinguistic milieu serves totally different perspective and challenges when it comes to language preservation and its political identity. By willingly calling the politically ignorant languages a ‘Dialect,’ she has certainly tried to invoke the idea that ‘Dialect’ word should ignorant and asked for the measure to call every idioms a language. However, that will not change the political and social attitude towards a language. She herself used the term ‘dialect’ in the title as ‘dialect planning’. She has suggested five steps towards the dialect planning i.e. corpus planning, status planning, prestige planning, acquisition planning and family planning. Though, the paper claims to avoid the political recognition of language; it hardly makes sense, status planning can’t be even thought without government and/or political authorities’ involvement. Even the languages with political status are vulnerable to the standard variety of it e.g. the status of Maithili a politically recognized language lacks proper political attention and the language seems helpless before the need and demand of Hindi and English in the region (Kumar, 2001). The various factors that contribute in the discouragement of a language use have not been

discussed taking very practical aspects of it e.g. she has discussed under the status planning that everyone should be allowed to speak his/her preferred tongues in all spheres of public life, local, nation, international; at the first place this is not practical, one uses language to be communicated and that is the sole purpose of a language, second the country like India where citizens have legal right to speak, encourage and advertise their mother tongues, sees no such activities. Without the support of political community one can't introduce the language in schools, curriculum and then the question of language nests, immersion education, and school based revitalization under language acquisition approach can't be thought of without the government support. The paper, however, is not primarily concerned with the critique of Bernini paper but will add up the issues and concerns arising in Indian context.

The paper, however, dealing with the concept or idea of language endangerment or dialect planning also talks about the prevalent myths concerning language endangerment. One of the widespread beliefs is regarding the role of multilingualism in language survival. This has been talked in the literature how multilingualism is helpful or an important tool in maintaining of two or more languages in a speech community. India was always a home of multilingualism but still there is always a threat of losing a language with the time. It might be the case that multilingualism helps one language to sustain a little more than its natural death, but ultimately it has to die. Multilingualism can't be the ultimate solution for the language maintenance.

Another prevalent belief in the literature is mutually intelligible characteristics of the languages and the situation of diglossia (Ferguson, 1959) also supported by Bernini, which considered to be good for the two languages to sustain together because of its characteristics, however, what I see in the case of Magahi that it is also subject to time and level of intelligibility; in the diglossic situation one language enjoys the political and economic benefit (H) with respect to other (L), therefore feel superior. The mutual intelligible characteristic helps speaker of the language to master themselves in the politically strong language easily and separate themselves from the one which is no good for them. This part also highlights the speech community association with the language, their language ideology, and language identity. There are certain features of language which acts as identity denotation to the people, however, is not endorsed by the people taken very badly. People don't want to be associated with the language.

I have taken the situation of Magahi, Bhojpuri and Maithili, the three so-called dialects of Hindi. However, all are not in the same situation, so, I have categorized the issues and measurement to see and discuss the challenges. e.g. the situation of Magahi, major prevalent issues with Magahi is that it is mutually intelligible with Hindi (more than Bhojpuri and Maithili), no politically affiliation, no documented literature, known as oral language, less worked, considered less-sober language, etc. Bhojpuri is politically ignored language, a major language, people take pride in the language, known for its great literature, print and electronic media, Bhojpuri cinema is very famous and established, considered less-ober, sexist language, etc. Maithili a politically recognized language, but doesn't enjoy the same popularity as Bhojpuri, known for its literature and sweetness, enjoy the attention of print and electronic media, taught as subject in universities, etc. These three languages of Bihar and I am sure many others that are subordinated by a dominant languages, present such and very different scenario that is unique to them. In this paper, however, I shall limit myself to these three major languages of Bihar.

Reference:

Andrea, Bernini. 2014. Preserving languages beyond the political dimension: Some proposals for a dialect planning. Italy: *Sustainable Multilingualism, 4*. 14-24.

Kumar, R. 2001. Shift from Maithili to Hindi: A sociolinguistic study. *Studies in the Linguistic Sciences* 31:2, 127-142.

Ferguson, C.A. 1959. Diglossia. *Word*, 15.2, 325-340.

Sense of loss or triumph of adaptation: An enquiry into the history of a language policy, its politics and turn of events in the age of technology

Preeti Kumari
Jawaharlal Nehru University, New Delhi.
kumaripreeti.2293@gmail.com

Various methods of recording human language in the written form have transversed a huge distance from the ancient engravings on rocks to the early lithographs until finally reaching the present state of Unicode generated in modern computational devices. This evolution demanded several modifications in the body of a language. These demands ranged from standardization of the script out of multitude of prevailing options to adapting to a completely different script, of a distant or not so distant linguistic cousin. While the choice of Devanagari for the written form of Hindi in the 19th century is an example of the former, the latter is manifested in the example such as the preference to Devanagari for Bodo in place of Assamese. However, it would be naive to accept that such shifts are guided purely by the compulsions of adaptability to contemporary technological shifts like better suitability for printing, or linguistic factors like better representation for all the phonetic expressions in the language. Had it been so the shift would have been directed towards the most viable choice of script among most closely related languages but as we know from the examples of Bodo language in Assam and Maithili in Bihar the choices often spill over to the distant relatives and neighbours.

I focus on the case of Maithili which is an eastern Indo-Aryan language spoken by more than 35 million people in the state of Bihar in India and Tarai region of Nepal. It has a long and rich past of written and oral tradition which dates back to at least the end of 14th century. While the regions under the influence of Malla rule in Nepal saw Newari adopted as a script for Maithili side by side with Tirhuta also known as Mithilakshar, the Mithila region in India saw the dominance of Mithilakshar only. The political shifts in these territories also had consequences over the scripts. For example, the arrival of Gorkhas after ousting Mallas in 1768 saw Newari being replaced by Devanagari. Similarly, in Mithila region in Mughal India, in the revenue administration under Todarmal, Maithili was used for keeping records by the Kayasthas who used Kaithi script for doing so and as a result Kaithi became the dominantly used script for representing Maithili in daily life while Mithilakshar was limited only to the literary and cultural purposes. Again, after the educational dispatch of 1854 the government took steps for the propagation of modern education, the Mithila region was considered to be a part of Hindi speaking area (as it was as later as 1881 that the Maithili language was beginning to be considered as a separate language and not a dialect of Hindi) and Hindi in Devanagari script became the medium of instruction. Slowly and gradually printing facilities and a large number of books were available in Devanagari while the traditional scripts of Maithili were left behind. It was followed by adaptation of Devanagari for writing and printing Maithili by a certain literate section of Maithili community. In independent India, Maithili was accepted by Sahitya Akademi as a language category in 1965 which nurtured and promoted the Maithili literature and finally in 2002 it was recognized as a major Indian language by inclusion in the VIII schedule of Indian Constitution making it one among the 22 scheduled languages of India. Today, Devanagari is the officially recognized script of Maithili in primary and secondary education making it the official script of Maithili for all practical purposes while Tirhuta or Mithilakshar is recognized only optionally along with Devanagari in higher education. Was this shift fuelled by the concern for adaptability of Maithili language with contemporary technological development like lithographs or other printing techniques, or was it a Said-ian case of misplaced paternalistic oriental knowledge produced by colonial apparatus subsequently internalized by nationalists who saw the linguistic heterogeneity as a threat to the conception of united India? What did it mean to the immensely rich body of literature hitherto written in traditional scripts- was it dominated by the literary currents of the Hindi belt or it resulted in a healthy exchange of ideas prospering both languages? Was Devanagari an apt choice based upon the fact that it is linguistically considered to be distantly linked to Mithilakshar in comparison to scripts such as Bangla and Odiya?

In the light of the above observations and questions this paper proposes to examine the dynamics behind the choice of Devanagari in place of Mithilakshar or Kaithi script for Maithili. It will examine the hypotheses of the role of technology, influence of political factors or a complex interplay of several socio-political or technological factors. This examination is expected to yield an insight into the colonial and post-colonial language politics and policy in a vast multilingual nation like India. This analysis would be followed by a study of the consequences or effects of the shift over Maithili language shedding light on the advances made or the drawbacks suffered on account of the aptness or inaptness of the adopted script for Maithili. It would focus on the phonological issues involved in such shifts in a comparative manner such as the presence of symbols for efficient representation of the speech sounds of the language. In the same trajectory, the paper would finally look at the recent technological developments such as various projects for the development of Unicode for Mithilakshar and make it friendly for the user on digital platforms. This gaze on the technological development would also focus on the potential of the language technologies in preserving and possibly reviving the endangered and lesser known languages and scripts including the challenges to be overcome and opportunities staring in the face.

A strange case of endangerment and revitalization

Bornini Lahiri
Central Institute of Indian Languages, Mysore
lahiri.bornini@gmail.com

Dhimal is a Tibeto-Burman language, spoken in both Nepal and India. The population of Dhimal speakers in Nepal is 19,300 (according to Ethnologue). But in India the number of speakers is much less. In India it is spoken in Darjeeling district of West Bengal. The number of Dhimal speakers in India is around 1000. The Indian variety of Dhimal has become endangered as the Dhimal speakers are shifting towards Bengali, the major and the official language of the area. The paper deals with the Indian variety of Dhimal.

Dhimal too like any other endangered language has lost many of its lexical items like numerals, names of colours and names of objects of day to day use and has replaced them with borrowed words from neighbouring speech communities. Due to intense contact with Bengali, a so-called prestigious language of the area and inter-community marriages, Dhimal community has adopted lots of cultural activities of Bengali community, which is reflected in their linguistic behavior.

Government of India gives reservation to certain communities which it feels, need special promotion. Article 46, of the Indian Constitution marks that special attention will be given to the weaker section of the society of India and reservation in some fields will be given to these communities for their development and promotion. These communities are categorized in different groups, popularly known as reserved categories. Some of the reserved categories are Other Backward Class (OBC), Schedule Class (SC) and Schedule Tribe (ST). Each group gets some set of facilities and benefits from the government. However all the reserved groups do not enjoy equal benefits. The benefits depend upon social and economic status of the group like OBC gets certain percent of reservation in the government job while for ST the percent is different.

Dhimal was given the status of Other Backward Class by the Government of India. However, in recent times, Dhimal community realized that if they could establish themselves as a distinct tribe, then they could get the status of Schedule Tribe. The status of Schedule Tribe is more beneficial than the status of Schedule Caste as the Government provides more benefits to the ST than the Schedule Class.

The community is continuously writing letters and application to the Government, to give them the status of Scheduled Tribe. But along with that Dhimal community is trying to reestablish their distinct features. The community feels that if they can focus on their individuality then they can be recognized easily by the Government and then the Government can give them the status of ST.

As the language and the identity cannot be separated so they are taking help of language to establish their identity. The community is trying to relearn the language to show that their language is different from the neighbouring communities. To establish this fact they are promoting their language and culture. They are publishing books in Dhimal and teaching Dhimal to young people. Dhimal do not have a script of its own so they are using Bengali script for the purpose. They have formed "Dhimal existence preservation welfare society". The attitude towards the language is shifting from negative to positive. As Spolsky and Shohamy (2000) marks, that nation's language policy is born from the unique interplay of its political, cultural, religious, educational and economic ambitions and realities. In the case of Dhimal it can be clearly witnessed that political and economic ambition is shaping its language attitude.

Earlier the speakers thought that Dhimal language cannot be helpful in providing job but now they think that Dhimal language can make them get the special status of ST which can get them jobs easily along with other facilities. This is the main motivation to relearn Dhimal. Spolsky (2004) points out four motivations behind the language policies of modern independent nation. One among them is an increasing interest in the claims of identity and the other is an increasing interest in linguistic rights. Both can be seen effecting Dhimal speech community. In fact both the motivations are interrelated in the case of Dhimal.

But Indian scenario is different from the ones described by the scholars like Spolsky (2009), Albury (2015) as here the linguistic rights also involve politics of caste, tribe and religion. As Indian society is divided into various groups of caste, tribe and religion, the linguistic scenario becomes complex mixture of all these.

The processes of endangerment and revitalization were both internally driven in the sense that no external force has been applied to the community, although the political condition plays its role in designing the fate of the language. Attitude towards languages primarily depends on social and political conditions (Sasse 1992). Dhimal also challenges the idea, languages and language varieties usually become endangered because their speakers are in contact with a dominant language (O'Shannessy 2011). As the dominat language Bengali still stands as the neighbouring language yet Dhimal speakers are relearning Dhimal. The paper describes the process of endangerment and revitalization of Dhimal based on first hand collected data. It talks about the reasons leading to a change in the attitude of Dhimal speakers towards their language.

References:

Albury, J. N. 2015. National language policy theory: Exploring Spolsky's 4 models in the case of Iceland. *Language Policy*. 1-18.

O'Shannessy, Carmel. 2011. Language contact and change in endangered languages. In Austin, Peter K. & Sallabank, Julia (eds.) *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Sasse, Hans-Jurgen. 1992. Theory of language death. In Brenzinger & Matthias (eds.), *Language Death*. Berlin: Mouton de Gruyter.

Spolsky, B. 2009. *Language management*. Cambridge: Cambridge University Press.

Thomason, Sarah G. & Kaufman, Terrence. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.

Spolsky, B., & Shohamy, E. 2000. Language practice, language ideology and language policy. In R. D. Lambert & E. Shohamy (eds.), *Language policy and pedagogy essays in honor of A. Ronald Watson*.1–42. Amsterdam, Philadelphia: John Benjamins Publishing Company.

The Grandfatherly relation of God with the Mundas: An inquiry into the endearing grandparent-grandchild relation as seen through their kinship terms

Gunjal Ikir Munda & Deep Lakshmi
Central University of Jharkhand, Ranchi
gimunda@gmail.com & deeplakshmimunda@gmail.com

The Munda tribe, which is a part of Proto-Austroloid race, presently resides in the modern day Indian states of Jharkhand, West Bengal and Orissa. According to the 2011 census, their population is around 2 million. Mundari is the name given to their mother tongue, which belongs to the Austro-Asiatic language family. The sister languages of Mundari are Santhali, Ho, Kharia, Asuri, Birhori, with whom the Munda people share many cultural as well as linguistic similarities .

Not much research related to literacy has been done on the philosophy of the Indian tribes' way of life. The obvious reason being the absence of tribals from the literary field who could explain their way of life better than the others. Recently, with the spread of literacy among the tribals, they have started writing about themselves and slowly a 'Tribal Philosophy' is coming into light. Of all the philosophies among the tribals, they also have a distinct religious philosophy, which is very different from all the organized religions of the present day. The tribals have moulded the spiritual image of God in light of their grandfather (or grandmother) instead of the more prevalent 'father-like God'. This is very much due to the emotional connection which the grandparents have with their grandchild. It is said that of all the relations a tribal community, the relationship between the grandparents and grandchildren is the most endearing.

Presently, we wish to inquire into the grandparent-grandchild relation among the tribals, with specific reference to the Munda community. The proximity in grandparent-grandchild relationship is very much visible in the rituals and practices by the Munda community but that proximity is also very much implicit in their kinship terms. For instance, the Supreme Being of the Mundas, known as Singbonga (literally, the Supreme Spirit) is also known as *haram*, or the Old man, also the ancestral spirits being very much a part of the belief system of the Munda are also known as *haramko*, the Old people, and of course in the world of living, any grandfatherly person is referred as *haram*, the old person. Also, in the Munda epic *Sosobonga*, when the Supreme Spirit assumes human form, he chooses himself to serve an old couple as their grandson. The fact that the Munda people like to think of their gods as someone like their grandfather, makes the grandparent-grandchild relation all the more sacred. All the more, the present pressure which the Mundari language is feeling from the other dominant languages is also due to the lack of the grandparent-grandchild relation; the general observation being, the trend in the modern families to exclude grandparents due to various reasons is actually harming the younger generation in terms of language and culture transmission.

For the present paper, we would also be discussing other kinship terms among the Mundas and would be trying to interpret them to bring out the peculiarities of their society. For instance, we would be looking into the reference term for wife, which is generally the name of the village from where she belongs. Another reference term for wife is *era*, which is also a generic term for women and is also used to refer to goddesses or spirits (*Jaher era*, the spirit of the sacred grove). Also, another term which we also would be looking into is the address term for elderly people- *saking*, meaning the person after whom he/she has been named, the term is used by the youngsters even though they might not have been named after that particular person. We would also be looking at the socially strict relation with one's wife's elder sister (*aji hanar*) and with one's husband's elder brother (*bau honjar*).

While interpreting the kinship terms, reference would be taken from folklore and interviews to substantiate it. Through this investigation, we intend to unravel the social structure of the Munda people, their beliefs and world view to some extent.

Selected Bibliography:
Bhattacharya, Sudhibhushan. 1970. Kinship Terms in the Munda Languages. *Anthropos*, *65*(3/4), 444–465.
Bodding, Paul O. 1934. *A Santal dictionary.* Oslo: Det Norske Videnskaps Akademi I.
Hoffman, J. 2009. *Encyclopedia Mundarica.* New Delhi: Gyan Publishing House.
Koboyashi, Murmu & Osada. 2003. Report on preliminary survey of the Dialects of Kherwarian languages. *Journal of Asian and African Studies* 66.
Munda, Ram Dayal. 2013. *Mundari Vyakaran.* Ranchi: Rumbul.
Munda, Ram Dayal. 2015. *Adi-Dharam: Religious beliefs of the adivasis of India.* Kolkatta: Adivaani.
Parkin, R.J. 1885. Munda Kinship Terminologies. *Man New Series*, 4(20).
Roy, S.C. 1995. *The Mundas and their country.* Ranchi: Catholic Press.
Van Exem, A. 1982. *The religious system of the Munda tribe.* Ranchi: Satya Bharti.

Evidentiality: Evidences from the Dura languages of Nepal

Kedar Bilash Nagila
Tribhuvan University, Kathmandu,Nepal
kedarnagila@yahoo.com

Evidentiality is a linguistic category whose primary meaning is the source of information. Marking one's information source indicates how one learnt something. Languages vary on how many types of information sources they have to express. Many just mark the information reported by some one. Others distinguish first hand and non first hand sources or witnessed and unwitnessed, visually hearing or smelling or through various kinds of inferences. The paper focuses on evidentiality system of the Dura, one of the endangered Tibeto-Burman language spoken by approximately and 5,169 (CBS 2001:170); 2,156 (CBS 2010) living in Lamjung in west Nepal. Data were collected at Pokharithok in Swami Bhanjyang VDC, recently declared to be Madhyabindu Municipality by GON, in the southern belt of Lamjung in Gandaki zone in 4 Number .Federal State (Constitution of Nepal 2015) in central Nepal. The number of the speakers in the villages are dispersed in Am Danda and Pokharithok. The first reported to be is lately settled and the second is the oldest villages comparing 37 houses in the past and now has been left by people in search of better life in Damauli, Pokhara, and Kathmandu. The methodology is descriptive in line with Pyane and Weber (2007:1) and Chellia (2011).

Case marking and alignment in Kinnauri

Harvinder Negi
Delhi University, New Delhi
negi.harvinder@gmail.com

This paper describes and analyzes the select Kinnauri varieties and tries to account for the phenomenon of ergativity with an aim to give a complete structural description of this alignment in different varieties of Kinnauri and thus show the kind of diversity that exists even within a language with very few speakers such as Kinnauri. Like most of the languages of Tibeto-Burman family, Kinnauri language varieties also exhibit split ergative pattern in which the subject of transitive verbs are case marked as ergative in perfective.
e.g. (All examples taken from standard Kinnauri)

(1.) *nu-s* *anu* *kamang* *lanA*
he-ERG his work do-PST
'He has done his work.'

(2.) *kina-s* *baaliga* *zog-zog*
you-ERG earings buy- PST
'You (all) bought earrings.'

(3.) *nuga-s* *seo* *zog-zog*
they-ERG apple buy-PST
'They brought apples.'

The ergative case appears optionally on di/transitive verb subject in 1/2P and obligatory on 3P. e.g.

(4.) *gi/-s* *ral* *jaak*
I- ERG rice eat.PST
'I ate rice.'

(5.) *ki/-s* *khau* *jaya*
you-ERG food eat.Q
'Did you eat food?'

(6.) *do-s* *khau* *ja-jaya*
He- ERG food eat-PST
'Did he eat food?'

In 4 and 5, ergative case appears optionally and in 6, ergative should come obligatorily.

About the language:
Kinnauri is a Tibeto Burman language spoken in the tribal region of Himachal Pradesh, India. UNESCO lists the language as a 'definitely endangered language.' It is a sporadic language. It is a SOV language and attests all features of a verb final language. Data to be discussed in this paper is from three Kinnauri languages- Standard Sinnauri, Sunnam Kinnauri and Bhoti Kinnauri. Except the Standard Kinnauri, there is no available work on other two varieties.

Developing a Machine Readable Multilingual Dictionary for Bhojpuri-Hindi-English

Atul Kumar Ojha
Jawaharlal Nehru University, New Delhi
shashwatup9k@gmail.com

This paper is an attempt to develop a machine readable dictionary (MRD) for Bhojpuri to Hindi and English. Bhojpuri, an Indo-Aryan Language is spoken in the western part of Bihar, the north-western part of Jharkhand, and the Purvanchal region of Uttar Pradesh. Not only in India, but Bhojpuri is also spoken outside of India in the countries such as Mauritius, Nepal, Guyana, Suriname, and Fiji. However, in India, socially and politically, Bhojpuri is considered merely a dialect of standard Hindi. Because of this socio-political view, Bhojpuri has been mainly ignored outside the linguistic studies despite the presence of quite a large population 37,800,000 (Census of India, 2001). Despite having a large number of speakers, there has been no language technology developed and almost no language resources available for the language. On the other hand, Hindi and English are very popular languages. Hindi belongs to the scheduled languages of India and it is one of the official languages of India too. English is spoken worldwide and now considered as an international language. Both languages have very large number of speakers. There are many language technology tools and language resources available for both languages.

Dictionary is a very valuable for lexical resources. Zgusta (1971) defines dictionary as "A systematically arranged list of socialized linguistic forms compiled from the speech-habits of a given speech community and commented on by the author in such a way that qualified reader understands the meaning". Thus, a dictionary is a primary source to understand the meaning of a language. Creating a MRD is a very difficult job for any language irrespective of the language belongs to the scheduled or non-scheduled or a regional or an endangered language. However, it is more difficult when a language had fewer resources. But, once the dictionary is prepared, it is a very useful tool and source of information that is used in natural language processing (NLP) task because it contains an enormous amount of lexical and semantic knowledge. The MRD can be used in the various areas of NLP such as Machine translation, speech recognition, morph analyzer and POS Tagger etc.

In the technology era, it is necessary that we create lexical resources and language technology tools for all the languages. In this paper, my aim is to develop a lexical database in the MRD form for Bhojpuri, Hindi, and English. My particular goal is to connect Bhojpuri, a lesser known/ non-scheduled language to the official languages Hindi and English. Hence, Bhojpuri is selected as a source language and other languages are used as the target languages in the dictionary.

There is vast number of literature available in Bhojpuri. However, very few grammar books and dictionaries are available in Bhojpuri. In my knowledge, there is only one traditional dictionary (edited by Arvind Kumar, 2009) and two online dictionaries (http://dictionary.anjoria.com/ and http://bhojpuria.com/v2/dictionary) are available. However, both online dictionaries have only approx. 750 words.

Methodology

1. Compiling the corpus

My first step was to compile corpus in Bhojpuri, Hindi, and English. For English and Hindi corpus, I used already available resources. Getting data for Bhojpuri corpus was a big challenge. The following development has been made.

1.1 Data source and the structure of the corpus

(a) I have developed 7K parallel sentences for Bhojpuri-English & Bhojpuri-Hindi corpus from general domain. To create parallel sentences, I have collected 7K monolingual data of Bhojpuri language from various sources such as Bhojpuri grammar book, magazine and web sources (www.anjoria.com) etc.

(b) After collecting sentences, I have extracted approximate one thousand five hundred words form the Bhojpuri-Hindi-English Lok Shabdkosh and some words extracted through on-line website.

Table 1: Statistics of collected words

From the parallel sentences	From lok-shabd kosh	From online websites	Total collected words
5,750	1,500	400	7650

2. Word alignment and removing duplicate words

(a) In the first step, collected parallel sentences were aligned by sentences alignment and then the words were aligned by word alignment module. The module used for this alignment was Giza++ module.

(b) In the second step, after alignment of these words, I have manually validated and annotated the parts of speech level.

(c) In the third step, I removed the duplicate words from the corpus.

3. Compilation through Java program

(a) When the corpus is annotated, I exported all the .xls format data's in RDBMS and compiled them through the Java program.

(b) In the next step, very soon, I will create a web-interface of this dictionary and will open it to access to everyone.

Result:
In the present form, this machine-readable dictionary consists of simple word-to-word mappings. The word from the source language i.e. Bhojpuri can be mapped into several optional words in the target languages i.e. Hindi and English. Moreover, the dictionary also provides the facility that you can type a word from any of these three languages and you get its equivalent in other two languages. For example, if you type a word in Hindi, you will get output in English and Bhojpuri. If you type a word in English, you get the output in Hindi and Bhojpuri. In addition to this, the dictionary also provides the parts of speech category of Bhojpuri words.

Issues and challenges in developing Bhojpuri-Hindi-English dictionary:
During the collecting of Bhojpuri head words, I faced many issues and challenges. The major issues and challenges that I had to deal with were regional variation, selection of headwords, linguistic variation and issues, script/ font issue, annotation issues, and standard vs non-standard issues.

Conclusion and future work:
To conclude, in this paper, I will be presenting two major points: one is the development of Machine Readable Multilingual Dictionary for Bhojpuri-Hindi-English and the second is the issues and challenges during developing a dictionary for lesser known languages. Currently, the dictionary is based on seven thousands six hundred fifty words, but in the future, I will create more parallel data and will add more words in the corpus.

References:
Bhojpuri English Hindi dictionary: http://dictionary.anjoria.com/. (Access date: August 10, 2015.)
Bhojpuri.com: http://bhojpuria.com/v2/dictionary. (Access date: July 22, 2015)
Kumar, Arvind. (ed.). 2009. *Bhojpuri-Hindi-English lok shabdkosh dictionary*. Agra: Kendriya Hindi Sansthan.
Mitkov, Ruslan. (ed.). 2010. *The Oxford handbook of computational linguistics*. London: OUP.

Saveski, M. & Trajkovski, I. 2011. Development of an English-Macedonian machine readable dictionary by using parallel corpora. In *ICT Innovations 2010*, 195-204. Berlin Heidelberg: Springer.
Zgusta, Ladislav. et al. 1971. *Manual of lexicography*. The Hague, Paris: Mouton and Company.

Establishing the phonemic inventory of a lesser-known language: Rathvi

Mona Parakh
The M.S.University of Baroda, Baroda
monaparakh@gmail.com

The Adivasis of Central India are divided into Mundas and Gonds. The Mundas are further divided into Bhils, Savars and Korkuns. Among the Bhils there is a sub division of the Bhiloris, the Mankars and the Rathwas (Rathwa, 1999). The Rathwas are divided into subgroups *(petha*) Bamania, Thebaria and Mahania (Singh, 1998). Rathvi is the language of the Rathwa tribe.

Rathvi (often also spelt as Rathwi) is a central Indo-Aryan language spoken in the Chhota Udepur, Jabugam and Nasvadi talukas of Baroda District in Gujarat. Rathvi is also spoken with added vocabulary of Hindi in the neighbouring state of Madhya Pradesh (Choksi, 2009), so that the varieties of Rathvi spoken in Gujarat and Madhya Pradesh show a clear difference in terms of the influence of Gujarati and Hindi vocabulary, respectively. It is a minority language which is on the records as a non-scheduled language. According to the 2011 census (Report no. 543 of NSS) the population of Rathvi speakers in the Chhota Udepur district of Gujarat was 6.42 lacs.

This paper is a descriptive study aimed at establishing the basic phonemic inventory of Rathvi as spoken in the Gujarat region. Even though considerable work has been done in the fields of literature, anthropology and folk tales, nothing significant has been done in terms of linguistic descriptions or analyses of the language. The current study, therefore, attempts to provide a preliminary description of the segmental phonology of Rathvi, in the hope that establishing the basic phonemic inventory of this lesser known language could provide a basic foundation for more advanced linguistic studies of an otherwise lesser-known and linguistically undocumented language.

The data provided in this paper has been mainly collected through field work carried out over a period of six months, on a weekly basis. The direct and indirect elicitation methods were employed for collecting the data from the primary informant, a native speaker of Rathvi. The data used for elicitation consisted of a word list of basic (core) vocabulary compiled using the CIIL list of basic vocabulary. Approximately one thousand words were transcribed using the IPA symbols. After each session of data collection, the data was cross-verified from other members of the community and the validated data was further organised into separate inventories for word-initial, word-medial and word-final occurrences of consonants. These sorted lists were then used to identify and list the minimal pairs.

Using the technique of finding minimal pairs 31 consonants and 15 vowels of Rathvi were established. To take an example: the trill [*r*] and lateral [*l*], belonging to the class of liquids are contrasted through minimal pairs, in order to establish them as phonemes.

[*leːʈiː*]	‘take’(Past, Feminine)
[*reːʈiː*]	‘sand’
[*həlkuː*]	‘light’
[*hərkuː*]	‘similar’
[*dʱoːl*]	‘drum’

[*dʰo:r*] ‘cattle’

The paper provides a list of around 33 sets of such minimal pairs on the basis of which the consonants of Rathvi were established. These consonant phonemes include, sixteen plosives consisting of bilabials, dentals, retroflexes and velars showing contrasts between the voiced- voiceless, aspirated-unaspirated phonemes; four palatal affricates; alveolar and glottal fricatives; bilabial, dental and retroflex nasals; an alveolar lateral and trill; a retroflex flap and two semi-vowels/approximants one, a bilabial and the other, a palatal. The paper also provides a description of the allophonic alternations and free variation found among these phonemes.

The paper provides around 10 sets of minimal pairs on the basis of which the 15 vowels of Rathvi were established. These include 8 oral vowels and 7 nasalized vowels. It is observed, that all vowels in Rathvi get relatively longer when they occur in stressed syllables. For lack of time, the stress pattern of Rathvi has not been dealt with in this work. However, minimal pairs for the nasalized vowels and their oral counterparts could not be obtained. It appears that these nasal vowels are surface level phenomena and that they arise through derivation from the underlying nasal consonant /*N*/. Vowels often become nasalized in the environment of nasal consonants. The typical scenario is for the nasalized vowels to become phonemic (contrastive) when later in time the nasal consonant is lost (Campbell, 1998). Over and above the brief description of nasalization of vowels, the paper also discusses free variation among vowels, and vowel sequences (diphthongs) in Rathvi.

In conclusion, this paper presents a preliminary study of the segmental phonology of Rathvi using the structural approach.

References:

Campbell, L. 1998. *Historical linguistics: An introduction*. Edinburgh: Edinburgh University Press.
Choksi, N. (ed.). 2009. *Tribal literature of Gujarat*. Mysore: Central Institute of Indian Languages.
Giegerich, Heinz J. 1992. *English phonology: An introduction*. Cambridge: CUP.
Hensen, Jette G. 2006. *Acquiring a non-native phonology*. London: Continuum.
Mackenzie, Ian. 1999–2013. *The linguistics of Spanish*.
http://www.staff.ncl.ac.uk/i.e.mackenzie/index.html
Rathwa, S. 1999. The adivasis of central India. *Dhol*. 3. Bhasha Research and Publication Centre.
Singh, K. S. 1998. India’s communities. *People of India, National Series*, 6. Anthropological Survey of India. New Delhi: OUP.

Social status of women in Bihar: How Bhojpuri and Magahi account for it

Sweta Sinha & Sandeep Kumar Sharma
Indian Institute of Technology, Patna.
sweta@iitp.ac.in & sandeep.phs16@iitp.ac.in

From time immemorial, language has played an important function as power determinant. In Indian society with magnified patriarchy, the role of language becomes manifold crucial especially in determining gender relation with power. As Lakoff (1975) puts it “If it is indeed true that our feelings about the world colour our expression of our thoughts, then we can use our linguistic behavior as a diagnostic of our hidden feelings about thoughts.”

Differences in male and female gender roles are related to the power differential between men and women. Structural and institutional power resides in the forms of access to educational, economic, and political resources and opportunities. In most societies, access to these structural forms of power is the aspect of male privilege. According to Kiesling (1997), “Along with the freedom brought by power…comes the expectation (or requirement) that a man will somehow embody this power in his identity”. The presentation tries to bring forth the status of women in Bihar as manifested in two Bihari languages: Bhojpuri and Magahi.

Women in a typical Bihari family are neither completely powerful nor completely powerless. It is the systematic shift from being powerless to powerful and at times vice versa that mark the various ways in which their identities have been popularly constructed and perceived. The presentation tries to draw such similarities and dissimilarities in the two languages. Some theorists believe that men's greater power and status in societies underlie the differences in gender roles. The powerful roles that men hold lead to the development of related traits, such as aggressiveness and assertiveness. Likewise, women who have less access to powerful roles develop traits consistent with their subordinate roles, such as submissiveness and cooperativeness. In sum, the power differential in favor of men may explain why stereotypical male traits are more valued than stereotypical feminine traits. Language portrays such power equations with versatility.

There are certain social issues emphasized by language which adversely affect the overall well being of women. In Bihar, the moment a woman gets married she receives the eternal blessing of "*dudho nahao aur puto phalo"* (being capable of having milk bath and giving birth to several sons); her function being restricted to her reproductive prowess only. A man with effeminate characteristics is sarcastically referred to as "*mauga"* (woman like) which is considered to be highly derogatory and demeaning. Language subtly expresses these power equations which in a patriarchal society like Bihar turns the set up against women. Expression like "*joru ka kamaai khana" (*maintaining livelihood on wife's salary) or "*beti ka kamaai khana"* (maintaining livelihood on daughter's salary) are considered highly derogatory which results in high rate of female unemployment even in educated families.

On the contrary the concept of "*ghar ke lachhmi"* (the goddess of domestic prosperity) suddenly catapults the social status of women, however, only within the confines of the household. Similarly, if a woman has male children, she is referred to as "*laikan ke mae"* (mother of male children) in an accented tone as if it were a trophy. Kulick and Cameron (2003: 95) say "The linguistic reflex of this (identity politics) is an impulse to claim for the community 'a language of our own'- a distinct way of speaking and or writing which serves as an authentic expression of group identity.

The presentation highlights several such issues as reflected in Bhojpuri and Magahi. The methodology for this research is simplistic comprising of data collection, analysis, inference and conclusion. The data have been collected through primary as well as secondary sources. The informants were provided with spontaneous as well as semi spontaneous settings in which they came up with gender based expressions. The study has been restricted to conversations in social gatherings like child birth ceremony, *upanayan* (coming of age) ceremony, and marriage and death ceremonies. Conversation in normal day to day life have not been recorded and analyzed. However, the inference that has been drawn is quite holistic and comprehensive.

The observations are then categorized so that the shift of power can be clearly witnessed from a girl's infancy through adulthood to her death. The observations are somewhat similar for both the languages but a comparative study makes the findings comprehensive and inclusive.

Keywords: Women in Magahi and Bhojpuri, Language and gender, power relation in Bihari languages, Gender and power in Bihari languages

References:

Keisling, Scott. 1997. Power and language of man. Sally, Jhonson & Meinhof, Ulrike hanna (eds.) *Language and Masculinity.* Blackwell Publishers Ltd. 65- 85.
Kulick, Don & Cameron, Deborah. 2003. *Language and Sexuality:* CUP.
Lakoff, Robin (1975). *Language and Women's Place.* Harper and Row, New York.

A Study of Person, Number and Gender of Halbi

Ajay Kumar Singh

Lucknow University, Lucknow
ajay.linguistics@gmail.com

Halbi is spoken mainly in the Bastar region of Chhattisgarh. It belongs to Indo-Aryan language family. It is transitional between Oriya and Marathi. According to George A. Grierson, Halbi is an eastern dialect of Marathi with much intimate connection with Chhattisgarhi dialect of Eastern Hindi *(The Linguistic Survey of India, Vol.-1, part-1, page 141)*, but according to Rai Bahadur Hiralal, Halbi is a mixture of Hindi, Oriya and Marathi. Halbi is also called Bastari, Halba, Halvas, Halabi, Halvi, Mahari, Mehari etc. Halbi or Halba is a tribal community of India. It is mostly found in the states of Chhattisgarh and some in Maharashtra, Andhra and Odisha.

Halbi is the second largest spoken language in Bastar region after Gondi. Although it is said the language of Halba tribe but it is spoken by many others and by the people who are native of this region. The use of Halbi is mainly found at Kondagaon, Jagadalpur, Bijapur, Sukama and Narayanpur districts. In other districts there is the mixture of other dialects. In Halbi we find the words of Awadhi, Bagheli, Chattisgarhi, Bhojpuri, Bhatri, Gondi, Marathi and Oriya etc. Halbi is a lesser known language. The present paper is a preliminary study of some of its grammatical categories, number, person and gender. In every language there are some unique features of grammatical categories. The purpose of this study is to look how do these categories function in this language.

Person in Halbi:
Person is a grammatical category that distinguishes speakers and addressees from each other and from other individuals. Grammatical person shows the relationship between the speaker and other participants in an event. It is a reference to a participant in an event, such as the speaker, the addressee or others.

Halbi has three persons: First, Second and Third. As like in Hindi, it has two types for second person.

Table 1: Second person in Halbi

2nd Person	You(singular)	*tuɪ/tuj*	*tuke*	*tuco*
	You(Plural)	*tumi/ tumən/ tumənəmən*	*tumək e*	*tuməco/ tumənəmən co*

Number in Halbi:
Number is a grammatical distinction which determines whether nouns, verbs, adjectives etc in a language are singular or plural and also it expresses count distinctions (such as "one", "two", or "three or more").

There are two numbers in Halbi - Singular and Plural.
Such as *leka* 'boy', *hun* 'he/she', *kɪtab* 'book', *pɪla* 'Child' are singular and *lekamən* 'Boys', *hunəmən* 'they', *kɪtabəmən* 'books', *pɪlamən* 'Children' are plural. For making plural the suffixes: *mən, ʈʰən, ɟʰən, kᵘbe, səpa/səpaj* are used. For example -

(1.) *ʈʰən* - It is used only for non-animate things.
duɪ ʈʰən pustək 'two books'

Agreement:
Halbi have number agreement with verb. Verb conjugations are different for singular and plural numbers. For example-

(2.) hun haʈ ɟajəse
'He/she is going to market.'

(3.) *hunənəmən haʈ ɟasot*
'They are going to market.'

Gender in Halbi:
Gender is a grammatical category dividing nouns into classes. Every noun must belong to one of the classes and there should be very few that belong to several classes at once. A language distinguishes between gender- masculine, feminine, or in some instances neuter, then each noun will belong to one of those genders.

Halbi has two genders - masculine and feminine or non masculine. Non feminine gender is used more than Masculine. For example, for sun, moon, stars, hills, rivers, trees non masculine is used. There are some affixes that are used for making non masculine from masculine, such as: *-i, -ni, -ɪn*. For example-

Masculine	Non masculine
leka 'boy'	*leki* 'girl'
kukəʈa 'cock'	*kukəʈi* 'hen'

Halbi have natural gender as well as grammatical gender, but in Halbi grammatical gender agree with verbs only in third person and singular number. For example:

verb - *ɟato* 'to go'

3rd Person	Singular	*hun gelo* 'He went'	*hun geli* 'She went'
	Plural	*hunəmən gela* 'They went'	*hunəmən gela* 'They went'

References:
Vaishnav, Harihar. 2013. *Bastar Ki lok kathae*. National Book Trust: India.
Puran Singh, Thakur. 1937. An introduction to the Halbi language. State Printing Press: Jagadalpur.

The need to prepare a Lambani lexicon

Zeenat Tabassum
Central University of Karnataka
zeenattabassum@yahoo.com

This paper calls attention to the need to document the lexicon of Lambani language as spoken by the community in Karnataka. I begin with introducing the topic of this paper following a short historical background of the community. I then discuss the emergence of new domains and the words related to these domains in the linguistic capacity of Lambani. Here I also mention the previous works done on the language which were scarce and not very satisfying thus proving the need for the present study. Finally, I bring forth the critical issue of language endangerment that lesser known languages like Lambani is currently facing in the hands of globalization thus again establishing the need to document it.

Studying a language essentially begins with the study of the lexicon of the language. Territorial acquisition by invasion, relocation, as well as linguistic groups sharing geographical boundaries contribute to sharing language features beyond genetic affiliation. Internally, language over time and in the hands of succeeding generations is interceded by gradual changes occurring very commonly within the vocabulary, although structural modifications also occur but relatively at a lesser degree. The vocabulary of a language tends to expand, modify or reduce for the major reason – variation in the degree of usage (increased, decreased or ceased to exist) considering their requirement by speakers in respective domains. From the standpoint that study of lexicon results in discovering historical contact, nature and direction of borrowing and the phonological changes involved validates the necessity to document lexicons of studied as well as unstudied languages of India; a linguistic area.

In the pages of historical records, anthropological surveys, and in the statistical data of the population, Lambani has been documented as a nomadic tribe that migrated from Rajasthan and spread across north-west and south-east of India in the early 17th century. Although the tribe is addressed by different names as Banjara, Banjari, Vanjara, Vanachara, Gormati, Lambadi, Lambada, Lamani, Lambani, Laman Banjara, Sugali etc, in different parts of the country where they settled, popularly they are known as Banjara. According to Cumberlege (1870), the first migrants evidently came to the Deccan valley with A'saf Khan, the Vazir of Shah Jehan in the year 1630. At present they are hugely populated in Andhra Pradesh, Telengana, Karnataka, Maharashtra, Madhya Pradesh, and Gujarat. Although Banjara spoken in Andhra Pradesh, Telengana, and Maharastra has been researched extensively little linguistic work is available on the Lambanis of Karnataka.

Initially Lambanis might have retained much of native speech as they settled in isolated pockets and came in contact with other ethnic groups only for trading. However, gradual social assimilation and modernization brought changes in the vocabulary of the language and has taken under its fold new domains accessible to the speech community. The introduction of crop cultivation to Lambanis is evidently a later practice in the community and also to the domains in which the language is used. This is because the community was engaged in rearing cattle and transporting food items from place to place since a very long time. This has been attested in historical records but the fact that various other occupations for sustaining livelihood as were practices by Lambani people is observed in their folk songs.

There are songs that narrate the service the child when grown up would offer to the community. Through the songs the depiction of Lambani men as shepherd, raveler, trader migrating for business and women as artisan, entertainer, and family raiser evidently speaks that land cultivation was not seen as a source for earning livelihood.

Grierson collected samples on vocabulary from all the Indian provinces where huge number of Banjaras settled in the early 20th century. However, Greirson did not record the vocabulary of Lambani settled in Karnataka stating that the dialect spoken here is similar to that spoken in Berar. Trail's lexion on Lambani (1970) on the other hand, provides a satisfying word list of farming and farming equipments though he does not include a cautionary note saying that the lexicon also contains non-native words borrowed from Kannada, Marathi, Urdu, and others as identified so far.

Edmondson (2004) states that in the case of borrowing the nature of contact leaves a cause and effect relationship between the recipient and the host language. If there is casual contact and little bilingualism noticed the borrowing is limited to only non-basic vocabulary. On the other hand, if the contact is intensive with much bilingualism we witness much lexical borrowing and moderate structural borrowing. And finally, overwhelming contact for a long time results in massive grammatical replacement.

In the case of Lambani we come across heavy lexical borrowing which clearly is the effect of continuous contact with Indo Aryan and Dravidian languages for over two hundred years. Words

originating in Urdu, Marathi, Kannada and English can be found occurring in almost every domain that language uses.

Few examples of borrowing are listed in the tables below.

Kannada > Lambani		
English meaning	**Kannada**	**Lambani**
coriander leaf	*kottambari*	*kotambi*
groundnut	*senga*	*senga*
roasted Bengal gram	*puthani*	*puthani*
coconut	*khopra*	*kopra*
sprout	*natik/ molake*	*matki*
school	*shaley*	*Sali*
bathroom	*baccalmane*	*baccal*
read/ study	*voduadu*	*vodero*
rake (tool)	*raigol*	*ragol*
saw (tool)	*garagasa/ karagas*	*kargas*
plow	*neligi*	*kurgi*

Urdu > Lambani		
English meaning	**Urdu**	**Lambani**
duration	*muddat*	*mudat*
love	*muhabbat*	*mobat*
hard times, trouble	*takleef*	*taklup*
luck, fortune	*taqdeer*	*tagdir*
virtue, favour	*neki*	*neki*
fate, fortune	*naseeb*	*nasib*

English > Lambani	
English meaning	**Lambani**
brush	*burus*
tiffin	*tipin*
pencil	*pensal*

Marathi > Lambani		
English meaning	**Marathi**	**Lambani**
bath	*anghol*	*angodi*
to pour	*ghalne*	*ghalnu*
to sow	*pedane*	*pedanu*

The above data only shows the lexical items which tend to be phonologically similar with words from neighbouring languages. It doesn't establish at this point the nature of borrowing as the primary concern is to document the lexicon of the language. The data presented above is collected from native speakers of the respective languages residing in Gulbarga, Karnataka. A few hundred words have been elicited so far, all relating to domains such as food, agriculture, construction work, and education which apparently were later additions in the everyday life of the Lambanis.

It is important to mention that while eliciting data it was observed that certain lexical items as recorded by Trail in his work in 1970 reduced to exist in the vocabulary of only older generation of

Lamani speakers. Words like *mohbat* 'love', *mudat* 'time/duration', *nekidar* 'virtuous', *vichitar padnu* 'to be astonished' etc are being replaced by simpler or commonly available forms as spoken by the majority society. So we have *prem/ priti* for 'love', *tem* for 'time/duration', *accho* for 'virtuous' as well as 'good', and *chamak janu* for 'astonished/surprised' used regularly by young Lambani speakers. If the young Lambani speakers are narrowing down the usage to single lexeme for conveying more than one meaning this might be an instance of gradually reduced vocabulary although it would require more detailed study to confirm the assumption.

"Language endangerment is caused primarily by external forces such as military, economic, religious, cultural, or educational subjugation" (Brenzinger & Graaf, 2007). In the case of Lambani, socio-economic subjugation till the recent past had been a common practice by the caste-ridden society. Although government in recent time has introduced literacy initiatives among marginalized communities as the Lambanis, the effort pushes the linguistically minority speakers to abandon their native language for accessing education. Given that, younger generation of Lambani is constantly moving away from their native language the need to document and preserve the language before it is completely lost is utterly crucial.

Reference:

Brenzinger & Graaf. 2007. *Documenting endangered languages and maintaining language diversity.* Berlin: Mouton de Gruyter**.**

Cumberledge. 1970. Sketch of the Banjaras of Berar. In *Gazetteer for the Haidarábád assigned districts commonly called Berár* (ed.), Lyall, A. O., Commissioner of west Berar, Bombay

Edmundson, Jerold A. 2004. Typology and language contact phenomena in Southeast Asia. In S. Burusphat (ed.), *Papers from the eleventh annual meeting of the Southeast Asian Linguistics Society*, Tempe, Arizona, 233-261. Arizona State University, Program for Southeast Asian Studies.

Grierson, George A. 1967. *Linguistic Survey of India* IX(3). Delhi: Motilal Banarsidas Publication.

Naik, D. B. 2000. *The art and literature of Banjara Lambanis*,' India: Abhinav Publication. Trail, Ronald L. 1970.

Trail, Ronald L. 1970. *Lambani – phonology, grammar and lexicon.* Pune: Pune University. (Doctoral Thesis) http://www.language-archives.org/item/oai:sil.org:9360,'