Issues and Challenges in Preparing Multilingual Digital Lexicons

Dr. Atanu Saha
School of Languages and Linguistics
Jadavpur University
atanu.saha@jadavpuruniversity.in

Abstract

The paper probes into the issues and challenges of making a multi lingual lexicon especially in case of endangered and indigenous languages by using a combination of software/interfaces e.g. Sheetswiper, FLEX and Lexique Pro developed by SIL. The paper also takes note of the persisting problem of linguistic prescriptivism while preparing such dictionaries. Additionally, the need, for a change in the attitude as well as methodology, is addressed in this paper. The work concludes that there is a difficulty on the part of a researcher for this particular technique since it involves the manipulation of at least three interfaces but argues that the difficulty can control the quality of publication in several ways.

Keywords: Dictionary, open source, endangered languages, lexicography

1 Introduction

The present paper addresses the challenges of building multilingual dictionaries or lexicons by an open source software based technology developed by Summer Institute of Linguistics. Apart from that, the Paper also focuses on the methodology for preparing such dictionaries in the context of endangered and lesser-known languages. There are several possible ways by which a digital lexicon can be built such as electronic dictionaries for diverse NLP applications (See [Wilks,1995]), as well as Lexical Databases [Miller,1995], World Knowledge Bases [Lenat,1990], ontologies [MikroKosmos], cited in (Sáenz & Vaquero, 2014) programming based (java class) (Saha, 2010) and web-based (Genee & Junker, 2018). In this paper, instead of a single software interface or program oriented interface, I have argued in favor of a combination of software and interfaces i.e. Sheetswiper, Flex and Lexiquepro developed by SIL (Products , 2018) . The lexical data have been used from Kurmali and Koda two endangered and lesser known languages spoken in parts of West Bengal and Jharkhand in India. I have tried to show that an effective way of preparing a dictionary requires the lexical entries to be put in multiple languages/scripts along with the glosses in multiple languages/scripts. This, in turn, makes the dictionary user-friendly and can reach out to a wider population.

The paper is organized into four sections. In Section 2, I have surveyed the current literature pertaining to the issues of endangered languages and the role of dictionaries to preserve the lexicon of such languages. The methodology has been articulated in section 3 while in section **Error! Reference source not found.**, I have noted the challenges within the existing technology.

2 Endangered languages and dictionaries

(Kroskrity, 2015) shows that the dictionaries are the most useful resource for an endangered language community. It is also the linguistic genre which makes a direct impact on the community interests. Dictionaries not only help the users to retrieve the lost words but it also assists the learners to thrive in their vocabulary. This certainly requires user-friendliness of a dictionary as pointed out by (Ivanishcheva, 2016). She shows that for a modern lexicographic resource of an endangered language, one should also take into account an anthropocentric and cognitive pragmatic perspective. While the anthropocentric information helps the users to understand the scope of the dictionary, the pragmatic aspect is going to define the purpose and the usage of the dictionary and cognitive aspect is going to tell us what skills one requires to read the dictionary in long run. One particular approach that (Ivanishcheva, 2016) upholds by citing (Landau, 2001, p. 9; Gao, 2013; Nkomo, 2015; Schryver – Prinsloo, 2011) that the dictionary of an endangered language should take into account is the active-passive language approach in the sense that the encoding and decoding of the lexical items should consider including the dominant language and a foreign language for giving it a better outreach. For example, A dictionary for Russians with explanations in the Russian language and for foreigners in their native languages. In this work, I have tried to give explanations such as encyclopedic information in Bangla and Hindi (dominant languages of West Bengal and Jharkhand respectively) and in English.

The other issue that has been highlighted in (Kroskrity, 2015) is the ‘confronting ideologies’ between the endangered language researchers, linguists and the linguistic community. There are often tendencies of putting linguistic prescriptivism rather than descriptivism while producing dictionaries and in the process, a researcher may land up putting wrong or inappropriate analysis in the form of ‘research’ in a dictionary. For example, (Kroskrity, 2015) has noted that there is a considerable amount of evidence which suggests that ‘dictionaries and other language resources have been designed by colonizers, missionaries, and agents of nation-state supported assimilation

campaigns'. As a consequence, the way the dictionaries are designed is 'to control and subvert than to enhance indigenous resources'. Dictionaries, like other products of an outsider's 'research, have often been distrusted by the indigenous communities. Therefore, the emphasis needs to be given to the inclusion of community members. Three key issues also need to be kept in mind. Firstly, explaining to the community about the usage and necessity of such works, reducing the tension between academic and non-academic functions of the work, secondly representing the verb forms with the possible argument structures and thirdly careful representation of the culture-specific and sensitive nouns. In the current work, the verbs are included in their non-finite forms but often given with an example sentence. The culture-specific nouns such as the name of the clans are listed both alphabetically in the dictionary as well as separately as a list. Let me discuss the details of this work in the methodology section.

3 Methodology

To address all the issues raised by (Ivanishcheva, 2016) and (Kroskrity, 2015) above and for our ongoing research project on Endangered and lesser-known languages[1] of India sponsored by the University Grants Commission of India, we have worked on five endangered and minor languages e.g. Koda, Toto, Kurmali, Mahali and Lodha-Shabar spoken mainly in the states of West Bengal and tried to use a combined open source technology to prepare multilingual digital lexicons. In this paper, I have elicited examples of Koda and Kurmali.

3.1 Data collection method

The languages do not have a script at the moment. We have collected the data through a questionnaire at the first phase of the dictionary.

3.2 Community participation

After entering the data, we have invited three Koda and Kurmali language experts to cross-validate our data. In the process, they have given additional words which eventually helped the dictionary to become rich and more useful. Since Kurmali is spoken in the neighbouring state of

[1] The work is an outcome of the ongoing project on to study and research on the indigenous and endangered languages of India during the XII plan period funded by the university grants commission, India. The data on Kurmali language was collected from Purulia District of West Bengal, India by the three research associates Dr. Bornini Lahiri, Dr. Arup Majumder and Dr. Dripta Piplai. The fieldwork was done between 28th February-5th March, 2017.

Jharkhand (Hindi as the official language), we have decided to build a trilingual dictionary (Bangla-Hindi-English) with the following schema:

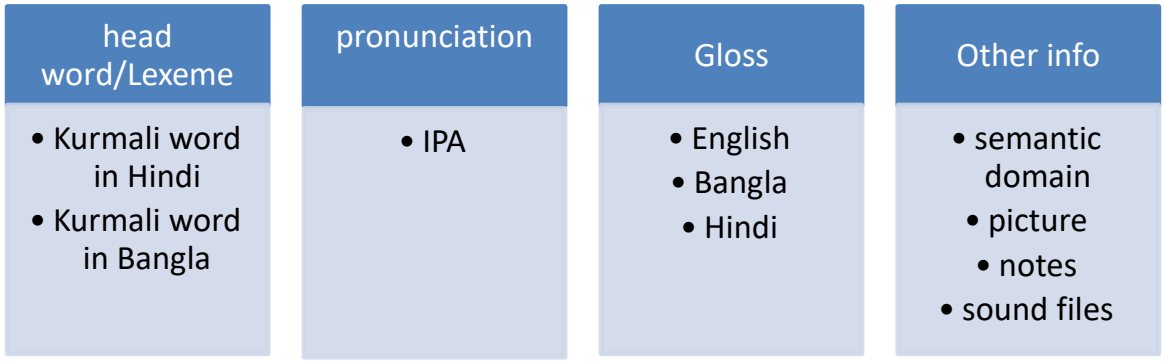


Figure 1 Schema of Kurmali dictionary

However for Koda, since it is mainly spoken in West Bengal, we have built a bilingual dictionary for this language.

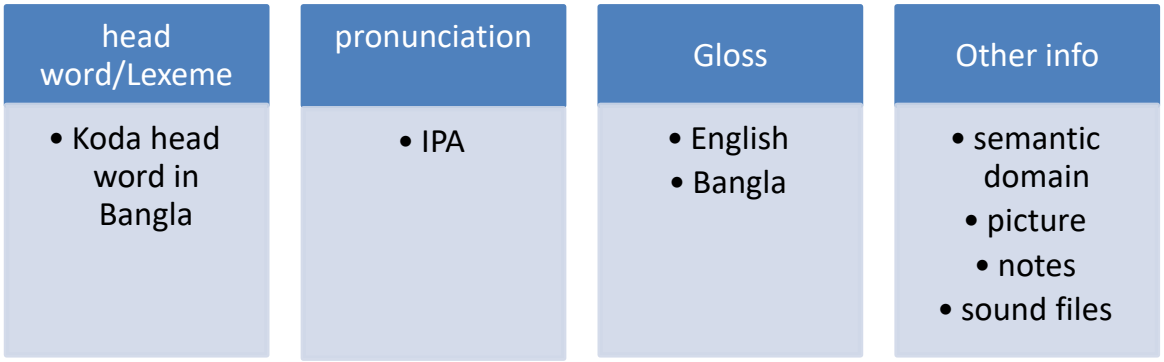


Figure 2 Koda

3.3 Use of technology

The dictionary is built up by mixing a number of software and to get the output files in several formats as would be useful for language documentation and build an archivable lexicon. In order, to accomplish this task we have used an excel/spreadsheet to enter the data.

The schema for excel sheets are shown as under

\lx	\lc	\ph	\ge	\gr	\gn	\sd	\nt
পৃথিবী	পৃথ্বী	pit̪ʰibi	world	পৃথিবী	পৃথ্বী,	Physical environment	

					धरती	
মাটি	माटी	maʈi	soil	মাটি	मिट्टी	Physical environment
ধুলা	धुला	dʱula	dust	ধুলা, ধুলো	धूल	Physical environment
কাদা	कादा	kaɖa	mud	কাদা	कीचड़	Physical environment
বালি	बाली	bali	sand	বালি	रेत	Physical environment
পাহাড়	पाहार	pahaɽ	hill, mountain	পাহাড়	पहाड़	Physical environment
মাঠ	माठ	maʈʰ	field	মাঠ	मैदान	Physical environment
ঢিপু	डीपू	dʱipu	island	দ্বীপ	द्वीप	Physical environment
পানি	पानी	pani	water	পানি, জল	पानी	Physical environment
সমুদ্দর	समुददर	səmuɖɖər	sea	সমুদ্দুর, সমুদ্র	समुद्र, सागर	Physical environment
বান	बान	ban	lake	ঝিল, হ্রদ	झील	Physical environment

Figure 3 Kurmali Lexicon

The koda lexicon is shown below.

\lx	\ph	\ge	\gr	\ps
রেমবিল	rembil	world	পৃথিবী	Noun
হাসা	hasa	soil	মাটি	Noun
ধুরি	dʱuri	dust	ধুলা, ধুলো	Noun
ল?সত	lɔʔsɔʈ	mud	কাদা	Noun
বালি	bali	sand	বালি	Noun
বুরু	buru	hill, mountain	পাহাড়	Noun
পাহার	pahaɾ	hill, mountain	পাহাড়	Noun
বাদ	baɖ	field	মাঠ	Noun
বাইক	baik	field	মাঠ	Noun
মাঠ	matʰ	field	মাঠ	Noun
ধিপা	dʱipa	upland village	পাহাড়ী গ্রাম/ উঁচু জমি	Noun
ডারশা	darʃa	island	দ্বীপ	Noun
দা?য়া?	dʱaʔaʔ	water	পানি, জল	Noun
শমুদ্র	ʃɔmudʱo	sea	সমুদ্দুর, সমুদ্র	Noun
গাডিয়া	gadija	lake	ঝিল, হ্রদ	Noun
কাসাই	kasai	river	নদী	Noun
ধুলাট	dʱulat	whirlpool	ঘূর্ণি	Noun
কুঁয়া	kũa	well	কুয়ো, কুয়া	Noun

দিড়ি কুঁয়া	ɖi̯ɾi kũa	well made of rocks	পাথরের তৈরি কুয়ো	Noun
ঝরনা দা?য়া?	ɟʰərna ɖa̯?a?	waterfall	ঝরণা	Noun

Figure 4 Koda Dictionary

Then the data was imported into a scripting software called Sheetswiper (SheetSwiper, 2018).

Figure 5 Interface of Sheetswiper developed by SIL

3.3.1 Sheetswiper

Sheetswiper recognizes the fields such as *lexeme, headword, pronunciation, semantic domain* etc. by the same convention used in the fieldwork explorer software (Flex), we needed to use the same field markers in the title of each column in the excel sheet. Noteworthy, Sheetswiper looks up only the first sheet of the excel file so all the data must appear in that sheet and not in some other place and secondly, it works only for the compatible version .xls and not .xlsx. Hence, we needed to use the Flex recognized field markers.

3.3.2 Language code and field markers

For each language, an ISO code is assigned which is needed for preparing a dictionary in Lexique Pro (Lexique Pro, 2018) especially. For Kurmali, the code is KYW. Apart from that, a uniform convention has to be followed in excel, Sheetswiper and subsequently in all the interfaces. For Kurmali-Bengali-Hindi-English we have used the following field markers.

\lx lexeme

Data were entered by using \lx for Kurmali in Bangla and then \lc was used for the Kurmali headword in Hindi.

\ph pronunciation

For pronunciation, International phonetic alphabet or IPA has been used.

\ge gloss in English, \gn gloss in a nationalized language & \gr gloss in a regional language

Hindi is used as \gn and Bangla as \gr.

\sd semantic domain

The dictionary entries are divided into semantic domains. We know that FLEx comes with a predefined.

Once the excel sheet is imported to sheetswiper, it converts the file into a .sfm file also known as standard format data and saved at the desired location.

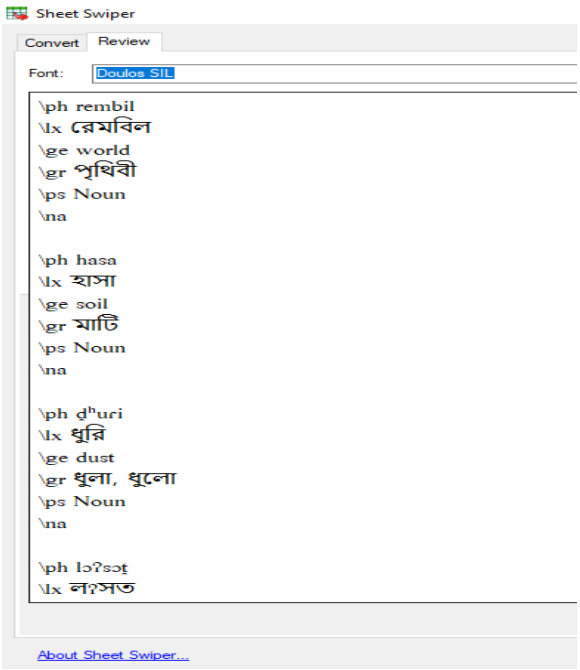


Figure 6 Data in sheetswiper

3.3.3 .sfm files to FLEX or language explorer

After the field work explorer or Flex has been opened, it provides one of the options as import standard format data into a new project.

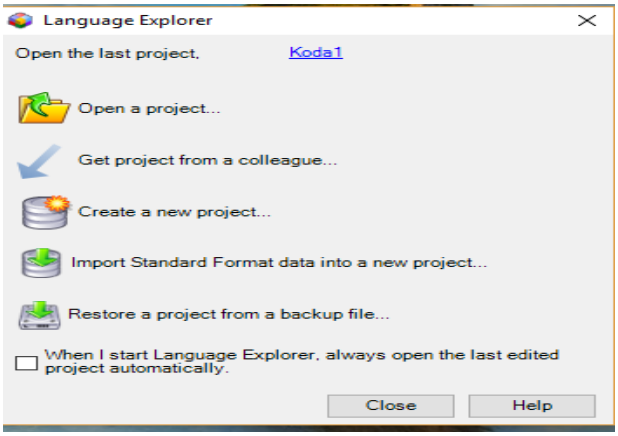


Figure 7 Import options in Flex/ language explorer

Once we click it and relevant information such as keyboard language and font information is mapped the flex interface shows up. If there are any typographic errors, formatting issues or some missing entries then those can be sorted out here. If things go alright, Flex generates a report showing the details of the entries.

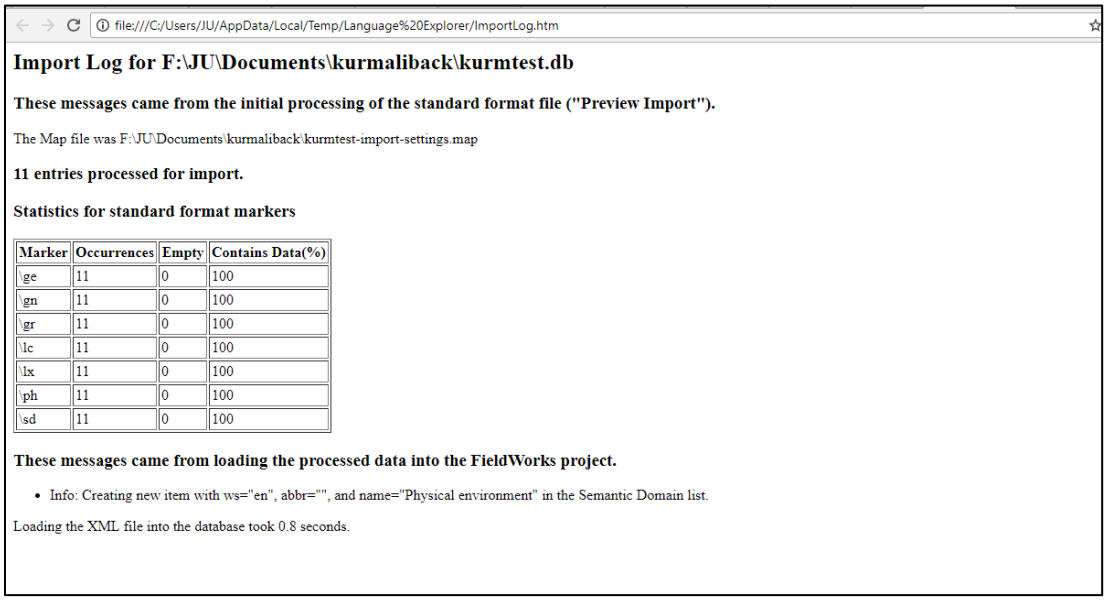


Figure 8 Log report

After the data is inserted, it can be noticed in the figure above that a Kurmali lexical item is entered in Bangla and Hindi preceded by the IPA transcription, Part of speech and subsequent glosses are written in English, Bangla and Hindi.

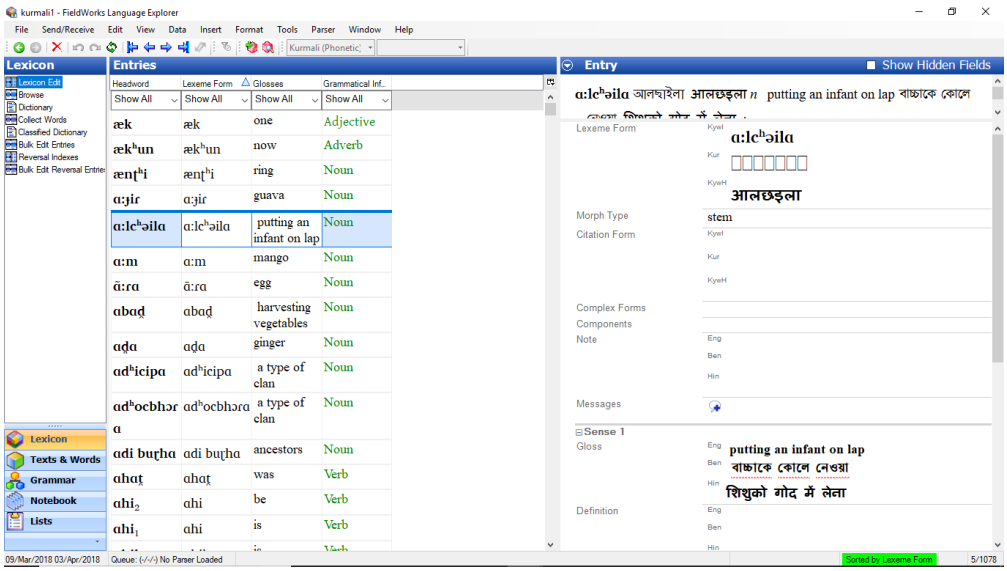


Figure 9 Flex lexicon edit view

After clicking on the dictionary tab, the interface of Flex is going to generate a listed lexicon such as the following:

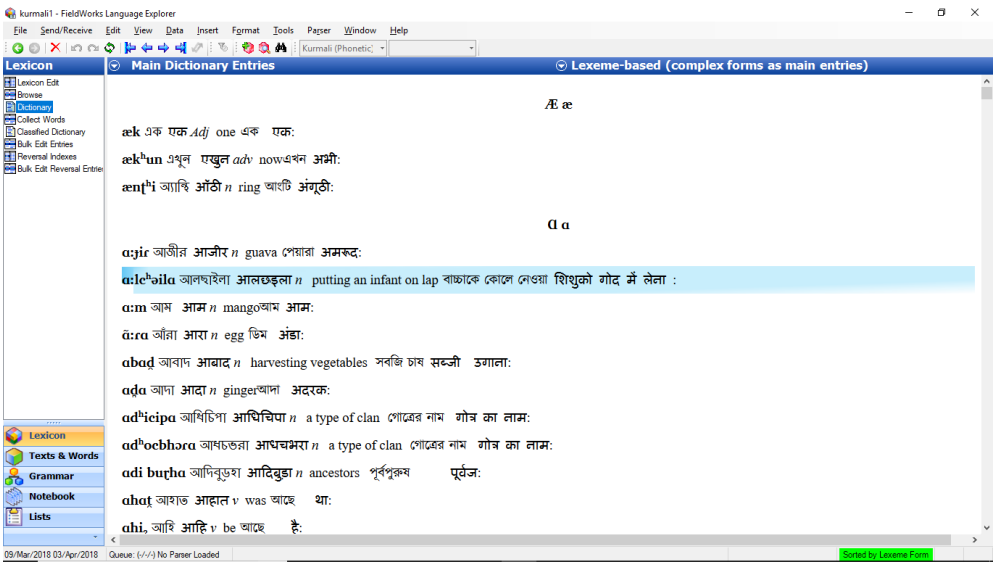


Figure 10 Dictionary publication view

Once the flex file is ready there are two requirements as such. One is to export the data into an archivable format and publish it in both digital and print forms. The imported dataset looks like the following at the FLEX interface.

The suitable formats are Full lexicon lift 0.13 XML for an archive and a full lexicon SFM file (extension is .db) which is imported to Lexique Pro for generating a .doc/.rtf file.

3.3.4 Lexique Pro

Lexique Pro is an interface that quite useful to prepare a digital dictionary and a printable dictionary. The .sfm file generated by Flex can be opened in LexiquePro. After providing a few inputs such as the name of the language (which will be the name of the dictionary) the languages used for glossing and lexeme forms, font, picture details etc., the interface generates the digital dictionary along with picture and sound.

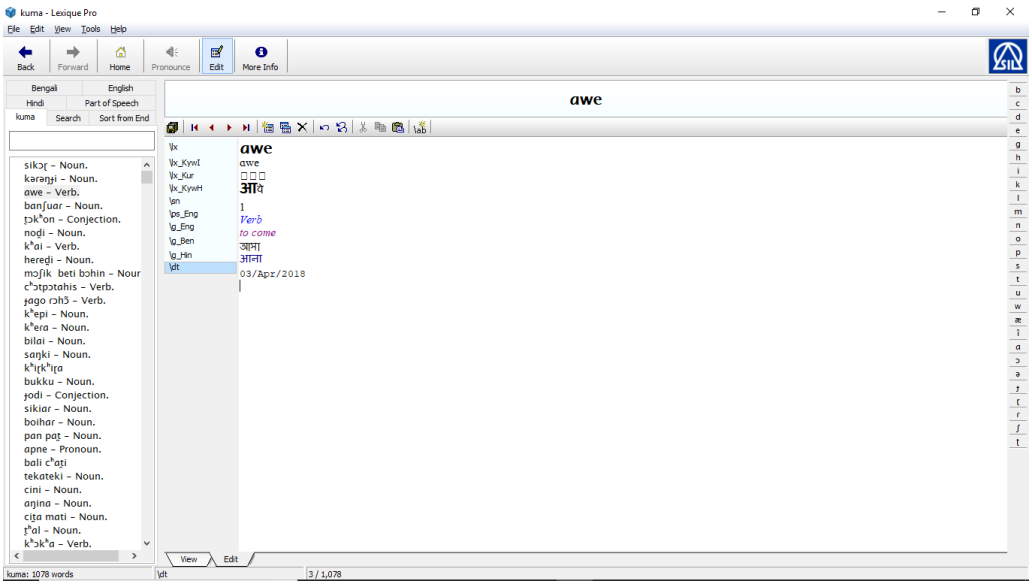


Figure 11 interface of Lexique Pro

The software also gives the output in the form of a .rtf file which is needed for the printing of the dictionary.

3.3.5 Insertion of the lexical entries

In this section, I am going to discuss the details of the lexical items produced by the Lexique Pro in a .doc/.rtf file. A couple of examples of Kurmali noun is cited below. The proper noun 'ant' comes with the following information. The headword shows up in the dominant languages Bangla and Hindi in the form of headwords followed by the pronunciation.

1) **চিমটি चिमटीcimti** *Noun.* পিঁপড়ে; *Ant;* चींटी.

The entry is glossed in three languages Bangla, Hindi, and English. Example of a Kurmali clan which pertains to the community sensitive concept has been represented as under:

2) **চিরুআর चिरुआरciruar** *Noun.* গোত্রের নাম ; *a type of clan ;* गोत्र का नाम.

The dictionary contains a number of pictures out of which I have demonstrated one example below.

3) **চুক্কা चुक्काcukka** *Noun.* ছোট পাত্র; *small pot;* छोटे मिट्टी का घड़ा.

Figure 12 small pot

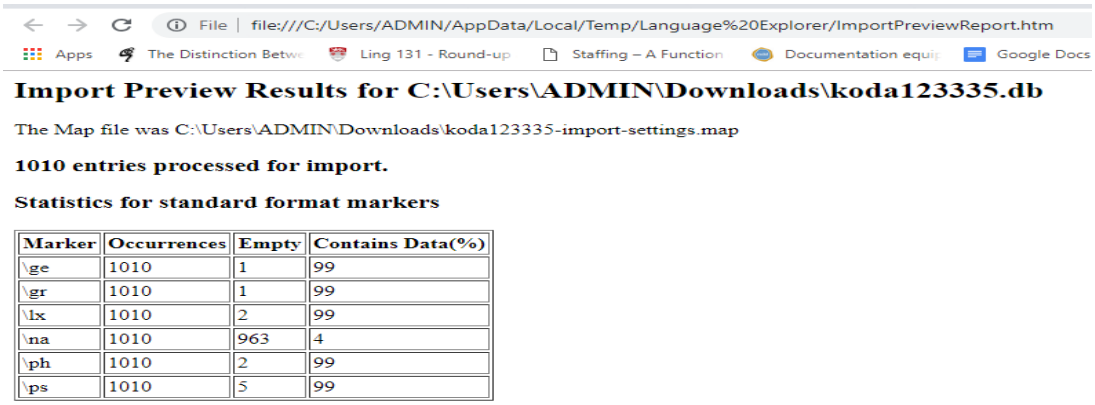
For a digital version, segmented audio files have been incorporated against a number of entries. The dictionary contains information about semantic domains such as physical environment, flora, and fauna which is useful for the speakers to recollect the words belonging to a particular domain. In the concluding section, let me summarize the entire work.

4 Conclusion and challenges

As desirable, the dictionaries are required to be built both in the digital platform as well in a printable form. Flex helps us not only build a digital dictionary and upload it via webonary but it also helps the researchers to generate XML files which are a prerequisite for the archives and a .sfm file. It further helps the researchers to build a digital dictionary using the lexique pro interface without any extra effort.

This combinatory technology has several advantages. At every stage, there can be two sets of misrepresentation of data. Firstly, any key information such as pronunciation or POS can be left out while entering the data or there can be major typographic errors. Given that these mistakes are inevitable; this combinatorial technology helps us in pointing out these errors. For example, once we import the .sfm files into Flex it maps every entry with its respective information. If there is any missing entry the import results directly reflect it into its output result.

The import in flex happens following eight steps i.e. overview & backup, file &setting, language mapping, content mapping, key markers, character mapping, readiness and ready to import. For example from our Koda sfm file, we have gathered the following report after the data was imported to FLEX.



File | file:///C:/Users/ADMIN/AppData/Local/Temp/Language%20Explorer/ImportPreviewReport.htm

Apps The Distinction Betw Ling 131 - Round-up Staffing – A Function Documentation equip Google Docs

Import Preview Results for C:\Users\ADMIN\Downloads\koda123335.db

The Map file was C:\Users\ADMIN\Downloads\koda123335-import-settings.map

1010 entries processed for import.

Statistics for standard format markers

Marker	Occurrences	Empty	Contains Data(%)
\ge	1010	1	99
\gr	1010	1	99
\lx	1010	2	99
\na	1010	963	4
\ph	1010	2	99
\ps	1010	5	99

Figure 13 Report highlighting the missing entries

It shows that out of 1010 entries, the anthropological note is missing for 963 entries and POS data was missing against 5 entries. Two empty categories were found in case of headword and 1 each was missing for vernacular and English glosses respectively. During the import of the data from Flex to Lexiquepro (generates an offline browser interface and a printable doc/rtf file), several typographic errors were observed and resent to the data entry person for correction.

The only challenge is that one has to learn all the software and the step by step combination of these interfaces. The glitch we have found her is, for example, the glossing in multiple languages is easily recognized by all the three software but the lexemes written in multiple scripts cannot be imported to Flex or LexiuePro as of now. If the lexeme or the headword needs to be typed in multiple scripts it has to be done indirectly. For example, Flex gives the option of keeping a headword and a lexeme form for the same entry separately. If we put the entry in one script against the headword and another script in the lexeme form and then export the data in Lexiquepro or XML, the entries show up in multiple scripts.

The current methodology helps the researchers and linguists in several ways such as the data entry is easier in excel sheet or an open office spreadsheet because of its user-friendly interface. Sheetswiper helps the excel data to be prepared in a format that can easily be imported to Flex. Since Flex currently does not allow an option to generate a .doc file or a .rtf file, Lexique Pro serves that purpose quite effectively.

5 Bibliography

Genee, I., & Junker, M.-O. (2018). The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language documentation and conservation*, 274-314.

Kroskrity, P. (2015). Designing a Dictionary for an Endangered Language Community: Lexicographical Deliberations, Language Ideological Clarifications. *Language documentation and conservation, 9*, 140-157.

Sáenz , F., & Vaquero, A. (2014). *Development_of_an_Electronic_Dictionary_based_on_Ontology*. Retrieved from www.researchgate.net: https://www.researchgate.net/publication/228815433

Abbi, A. (2001). *A Manual of Linguistic Field Work and Structures of Indian Languages.* Lincom Europa .

Flex. (2018). Retrieved from https://software.sil.org: https://software.sil.org/fieldworks/

Ivanishcheva, O. N. (2016). Dictionaries of Critically Endangered Languages: Focus. *Journal of Linguistics*, 73-86.

Lexique Pro. (2018). Retrieved from https://software.sil.org/lexiquepro/

Products . (2018). Retrieved from SIL language technology: http://software.sil.org/products/

Questionnaires. (2018). Retrieved from https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php.

Saha, A. (2010). Challenges in Building Multilingual multi-directional search. *Proceedings of Knowledge Sharing event 2 CIIL,*. Mysore : CIIL.

SheetSwiper. (2018). Retrieved from http://software.sil.org/sheetswiper/

Simons, G. F. (2018). Retrieved from Ethnologue: Languages of the World: Twenty-first edition. Dallas Texas: SIL International. http://www.ethnologue.com