Building corpus for endangered language a case study of Malayan Tribe

Mrs.D.Jenifer
Research Scholar, Department of Linguistics, Madurai Kamaraj University, Madurai – 21

Abstract-The study of language referred to as *corpus linguistics* has largely become accepted as an important and useful mode of Linguistic inquiry. While corpora were first mainly used as aids to lexicography and pedagogy, they have more recently been deployed for a wider range of purposes. Here we are collecting the corpora from different languages for protection and preservation which are found endangered languages in India. 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.

Keywords : SPPEL, Programme, language , endangered, corpus, Malayan

INTRODUCTION

The study of language referred to as *corpus linguistics* has largely become accepted as an important and useful mode of Linguistic inquiry. While corpora were first mainly used as aids to lexicography and pedagogy, they have more recently been deployed for a wider range of purposes. Here we are collecting the corpora from different languages for protection and preservation which are found endangered languages in India. 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. It aims at documenting the endangered languages which are recognized by Government of India. This project is supervised by Central Institute of Indian Languages (CIIL) which is working on the protection, preservation and documentation of endangered languages. The main objective is to document the endangered languages. We are collecting the corpus from the languages which are spoken by less than 10,000 speakers. On this I selected the south region endangered language. This paper is mainly going to talk about how to build corpus for endangered languages in the case study of Malayan Tribe.

CORPUS LINGUISTICS

The program in corpus linguistics supports analysis of a large amount of language use data and compilation of corpora, which feed into linguistic informatics research and into descriptive and typological research. Some of the specific targets are building electronic corpora and developing analysis and processing tools in order to support of new analyzing language data and multipurpose of the multifunctional integrative corpora of language use for languages. It used development and utilization of tools for corpus creation, morphological analysis, electronic dictionary creation, text analysis.

DEFINING MALAYAN TRIBE

According to the census of 2001, the total number population of tribes in Kerala is 3.64 lacks. This is 1.14% of Kerala's population. In this state there are 36 types of tribal people live. Malayan tribe people are mainly living in the four districts of Kerala. (Trissur, Ernakulam, Idukki and Kottayam) for data collection we have chosen the people of Malayan tribal living in Trissur District. According to the census of 2001, 4,826 tribal people are living in Thrissur district. Athirappally, Kodassery, Varantharapally, Mattathur, Pancherry the south east hill areas of the district have most of the Malayan people.

LOCATION

We collected data form tow settlements in Athirappally area. One is Vachamaram colony located in vazhachal forest division; there are 8 families with a population 29 people. Another one settlement is Thavalakuzhipaara it is come under erunakkulam

forest division there is 50 families with a population of more than 250 people.

LANGUAGE OF MALAYAN

Basically Malayan is the south Indian Dravidian family. The people of Malayan speak Malayan language. Malayan is the name of their race and language. 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 age people also speak Malayalam. But only elder people speak Malayan to converse within the same age group. So according to the UNESCO report one language is only spoken by the oldest generations means it is "severely endangered''

CORPUS

A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. There are four important characteristics to note here.

1. Authentic
2. Electronic
3. Large
4. Specific criteria

1. Authentic

Language and consists of a genuine communication between people. Eg. Normal conversation with people

2. Electronic

A text in electronic from is one that can be processed by a computer, Eg ; Essay that you into a word processed, and article that you scanned from a magazine or found massage from internet

3. Large

Large corpus means a greater number of texts then you would be able to easily collect and read in printer form. Eg; books

4. Specific criteria

It is very different from other corpus. The corpus is selected according to explicit criteria in order to be used as a representation sample of a particular language or subset of that language.

MALAYAN CORPUS

The Malayan language corpora come under the fourth type, that specific criterion. Particularly we took the Malayan tribe's spoken language which is identify the endangered language in Dravidian family. The purpose of study the language and collect the corpus is protection and preservation of the endangered languages and make the electronic dictionary for the languages. So we are design the corpus in different way as we want.

COLLECTION OF THE CORPUS

The data were collected over 10 days, from December 3rd, 2014, to December 12th 2014. I went two main location in Trissur district. On competition of the audio recordings, I collected the personal details age birthplace, current area of residence, occupation and level of education and their socio cultural behaviours. Most of the data collected from old age people those who are known their won mother tongue Malayan language. The corpus collected from both males and females. That gives a total word count of approximately 800 words, 50 sentences, 5 stories, 15 folk songs and some narrations.

CORPUS CONTRACTION PURPOSE AND METHODS

Purpose:

These corpus collection main purposes are;

- Documenting the endangered languages in India collect the language data from the people those how are having the less then 10,000 speakers in their language.
- Documenting the language data in socio cultural and ethno – linguistics aspects.
- Preparation of a sketch grammar of the language.
- Tri-lingual electronic dictionary from the corpus. (Main endangered language- reginal language- English)

Methods

Word corpus

1. Sentence corpus
2. Stories
3. Narrations
4. Folk songs
5. Ethno linguistics Aspects

Word corpus

The word corpus Human body parts ,stages of life , kinship terms Address and Reference Terms, housing and related, Artifacts and items of daily use, adirnments and costums, food and related, Drinks and Beverages, Health ailments and remedies, Religious and ritual terms, Festivals and related, music and its instruments, Sports games and entertainments, Occupation and related hunting fishing and its tools, Transport and related, Health ailments and remedies, celestial bodies and related, earth and related, water and related, Air and related, fire and related, Domestic animals and related, reptiles rodents and related, wild animals and related trees,Action verb, stative or position verbs, utterance verbs, body funtion verbs etc.

Example:

Artifacts and Items of Daily Use
/kəikələʈuni/ 'waste cloth'
/pot kalam/ ' pot'
/ku:jaili/ 'long spoon'
/ kəiɭə/ 'spoon'
/vəʈʈəcelvəm/ 'round vessel'
/vəɭəkku/ 'lamp'
/mu:ɖi pa:ʈraɱ/ 'pot lid'
/kuɲɲiʈi kalam/ 'Small pot'
/ʈaʈʈə/ 'plate'
/ʈi: kumbam/ 'barrel'

Sentence Corpus

Simple and complex sentences, word order, cases, pronouns and demonstratives, reflexive and reciprocalism interrogative, gerundial constructions tense aspect mood, complex predicates, causatives, Passives, Reduplication, comparatives and superlative, Sentences with Adverbs, Quantifiers and intensifiers, Ad positions, Adjectival clause, Intercausal relations, Exclamatory, etc.,

Example

1. The sun rises in the east.
   su:rijaʃ keɻakku uɖikkum
2. The bird is flying
   pətci pərəkkɯnɯ
3. The girl is givig sweet
   ponbulla: muʈʈa:j koɖɯkkɯnɯ
4. That house is big
   a: u:ɖɯ vəljə u:ɖɯ
5. I used to drink, but now i stopped
   pənɖɯ ɲa: kuɖitcəɖa:nɯ ipo ɲa: kuɖikaʈʈɯ

STORIES AND NARRATION

Stories and Narration for folklore, about their God, any good or bad spirit, black magic, any historical happens, their origin, religious, any epics like Ramayana and Mahabaratha. Folk songs

Example: Narration of Malayan about their village Name

The name of this village has an interesting story Narrated by an old woman. 'While the road was laid between Vaalparai to Athirapilli, an English man who stood in this place and inspecting the work was chased by an Elephant. As he climbed up a tree in a hurry to save himself, his watch fell into a burrow in the tree. He tried as many times as he can to take the watch out, but he never succeeded. So he named the tree as 'vatchmaram' as it had watch in it. The word 'Maram' in Tamil literally translates as 'tree' in English. And after that incident this place was named as vatchmaram.

ETHNO LINGUISTICS

For Ethno linguistics part we have to collect the data form this topics: folk songs, Songs with dance and without dance (*to be videographed*), Lullaby, Seasonal songs (*holi*, *chaita*), Professional/non-professional songs, Ceremonial songs, festivals and related, worship and related, settlement and housing system, Life cycle and related, food and drinks, Measurement, occupation profession, Ethno medicines (powder paste and decoration) Games and entertainment, Attire adornments and related judiciary, communicative behaviour of the community.

Example for Marriage

- In ancient times, there was no the concept of marriage, but the man lived with the woman whom he liked, said the elders.
- Nowadays, the marriage happens within Malayan tribes only. They never marry the people who belong to other tribes such as Kaadar.
- The marriage held in bride's home.
- The bridegroom does not get dowry from the bride.
- The bridegroom gives bride a new dress, a cloth to wear on her head and marry her by wearing a

Maangalyam called THALI in Malayan language. Then he takes her to home.

- A Malayan tribe man cans elopes a Malayan tribe woman whom he loves. It is not considered as wrong.
- They are excused and they are accommodated with Malayan tribe
- Remarriage is permitted
- Families are Nuclear

HOW WE ARE USING THE CORPORA

When we are collect the corpus form the field it will be in Audio and video form. We take the raw data form the field after that we have to convert the Audio form to word from using the phonetic transcription

- For phonetic transcription, we are using IPA (International phonetic Alphet)
- From the Phonetic transcription, we can analyze the language then we can do the language study on it.
- Phonological study
- Grammar writing
- Tir lingual E dictionary
- Document the language and socio-cultural practices of the people

MAKING OF THE E-DICTIONARY FORM THE COLLECTED DATA

An electronic dictionary is a dictionary whose data exists in digital form and can be accessed through a number of different media.

Electronic dictionaries can be found in several forms

- Dedicated had held devices,
- Apps on smartphones,
- Tablet computers or computer software,
- Functional built into an E- reader

The e- dictionary is made using the local dominant language, English and the concerned-endangered language, i.e., Malayan. Therefore, along with the endangered language three languages will be there for the readers, the specialty of the dictionary is all the collected words will be in audio form of the phonetic transcription of the concerned endangered language and with suitable picture which are captured form the field. Sub domains are also created according to the nature of the word so that all the people can easily use the dictionary.

CONCLUSION

This paper chiefly deals with how a corpus is built for endangered languages. My paper concerns with the description of the Malayan language, the people of the community, the area of their residence and everything related to the protection and preservation of the language and the community. The paper also takes into consideration the purpose and methods related to corpus collection and finally the benefits of such a corpus.

REFERENCE

[1] Corpus analysis and variation in linguistics. 2009 Edited by yuji kahlaguchi, makoto minegishi and secques. Durana

[2] Corpora and sociolinguistics investigating age and gander in female talk, Author : Brona murphy, university of Edunburgh, john Behiamins publishing company, 1996.

[3] Author : lynne Bowker and jennifer pearson Working with specialized language A practical guide to using corpora

[4] Questionnaire for Documentation, scheme for protection and preservation of endangered languages central institute of indian languages Department of Higher Education, Ministry of Human Resource Development, Government of India Hunsur Road, Manasagangotri, Mysore - 570 006