Tools and Technologies of Language Documentation

Prof. Bornini Lahiri and Prof. Dripta Piplai (Mondal)

Department of Humanities and Social Sciences

IIT Kharagpur

Week-08

Lecture-36

Lecture 36 : Language Technology for Endangered Languages

Welcome to the 36th lecture of the course, Tools and Technologies of Language Documentation. Today, I will talk about language technologies related to endangered languages. By now, you know that, language technology can actually help in language maintenance, language documentation and also in revitalization of a language. The topics that I will cover today are use of language technology for endangered languages, difference between computational linguistics, natural language processing and language technology, different types of language technologies, developing language technology for minor languages and steps of designing apps or related applications and also requirements for those. And for an example, I will show you Mundari gaming app which we designed here and how we have designed the app and also, I will talk about some other examples. So, now, you know that language technologies have the potential to dramatically accelerate and facilitate efforts in language documentation and revitalization because nowadays, we are using technology everywhere.

We are living in a digital platform where we have lots of technological apps related to language. So, when these are available for the dominant languages, but not for the minor languages then, there becomes an indirect pressure to shift to the dominant languages. So, with technological help, we can actually add prestige to the smaller languages and also, give them a new domain, where they can use the language. And technology can help us in both language maintenance, preservation and also in data collection, glossing, transcription and translation, which I have already talked about.

There are various apps which are actually helping in documenting a language. Similarly, language technologies are also used for the communities where it helps in revitalization of a language. or for creating various types of applications related to pedagogy, gaming

app, health related apps. So, there can be various types of apps created for the smaller languages or various types of other technologies can be created, so that the language can be represented in the digital platform. So, when we talk about language technology, we think about technologies related to machine translation, information retrieval and extraction, digital assistance and for educational technology; we know about various language learning apps, which are present there for many of the major languages.

There are also applications like grammar and spelling corrections which you even find in your smart mobiles and you can also find in various other applications. Like Grammarly and other applications, where they are actually correct your language, correct your grammar, correct your spelling or spelling predictions are available. Generally, what we see is that these are available mostly in the major languages of the world. But when we talk about the smaller and endangered languages and specially so, when we talk about the Indian endangered or smaller languages, we hardly find any language technologies. Though right now, we see that lots of groups and institutes are working for building language technologies for the smaller Indian languages.

So, now let us look through these terms because they are important. We often interchange these terms and we use it. So, computational linguistics, NLP, LT or these terms are also in fashion and people use it without knowing actually about it. So, it is good to know a little bit about it, just a brief definition. So, NLP is taken as a part of computational linguistics.

So, computational linguistics is broader and NLP is a part of it. It applies processing of natural language to enable a computer program to understand human language, both in written and oral form. So, that is what comes under NLP or natural language processing. NLP was earlier taken as an engineering part and computational linguistics as the theoretical part of the whole process of making computers react and respond to human language or natural language, as they say. But now, we see that the terms are also used for the same concept as well.

And language technology is a comparatively newer term, which actually covers all, and which involves any type of technology related to language human language or natural language, which is created for aiding speakers or language users. So, basically they help us in various use of language. It can be the result of any NLP or can be result of any computational process. There are various types of applications like you know machine translation. We often use machine translations; there are chat bots which are very common nowadays, natural language interface, sentiment analysis, spell and grammar checkers, text prediction, handwriting recognition, So, also we know about image recognition.

So, there are these type of applications which are mostly found in the major languages. Now, when we talk about technologies, we should also know that, like we represent language in both written and spoken form, technologies can be accordingly designed either for the written language or for the spoken one. So, for written languages, we have spell and grammar checkers, spelling predictions, handwriting recognition. So, all these are related to writing of a language, but as you know, many Indian languages are just oral languages; we do not have script for these languages or they are not widely written. So, for them what type of technologies are more useful, which are related to spoken language, which are related to speech like speech language technology, or what we can say is text-to-speech or speech-to-text.

If something is written, it can read out. So, when we were interacting with a community, we came to know that most of the speakers of that community could not read or write. So, if there is an application which can read the materials for them, whatever is there, may be some news items, some government policies they want to know, whatever is available, if that is read out to them, then that is actually very helpful for them. So, what type of applications can do that? Text-to-speech can actually help in doing that. So, the text is there, it will read it out.

Or we can also have speech-to-text, where I am saying something and it is getting written. So, that you find in Hindi, but also in some other Indian languages. So, these type of things, which can also be used in developing other apps- apps, entertainment related app, health related apps. These apps can use the techniques of TTS or STT. So, all these can actually be incorporated in developing an app.

So, one app can have different features of all those things. So, that can be done. Lot of research work, corporate and government funded, focus on the use of computational methods to support and revitalized minor and endangered languages. For example, for Mundari, we created an app with the help of funding from Microsoft Research India. They actually funded the project and under that, we collected data and we tried to develop an app.

So, it helps the speech community to use the language in various digital platform but there are also various challenges related to smaller or endangered language. One of the basic challenges that we see is that when I am talking about, for example, Mundari; by now, you know that Mundari is spoken in Odisha, West Bengal and Jharkhand. So, now which variety should we choose? If we choose one particular variety then others might not feel good about it, might feel that they have been left out. So, there can be things like that. Which variety to choose and which script to choose, if we are also developing the

app in some textual manner.

If there are certain things, which are in text and certain things which are in sounds. If we are writing the text, which script should we choose because the language uses more than one scripts, so depending on the region. Now, that becomes a big challenge; which particular variety to choose and which particular script to choose. Secondly, we know that there is no availability of the data for these smaller and minor languages. So, basically we have to start the work from the basic level like collecting data.

And when we are talking about building language technologies, it means huge corpus. So, just collecting say few 100 sentences does not help for creating language technologies, we need a good amount of corpus. So, that takes again lots of time. And for these languages, generally, we do not even find printed materials available. It is not that already, some textbooks are available which we can incorporate in the apps.

So, we have to actually develop everything. So, basically nothing is there. So, that takes lots of time and there are other logistic issues related to it. So, that becomes a very challenging task and availability of data and again, the choosing of a particular variety or a particular script. Now, when we are planning or when we start to design say an app, or anything which is related to language technology, if we are developing something, what we do is that, for these smaller languages, we generally start our work with collection of primary data.

Because as I mentioned, for any technologies, we generally need a good amount of purpose and we do not find that for the smaller or endangered languages. Now, when we have received the primary data or when we have collected primary data, we have worked on it, like we have done the glossing and the basic thing on the data, we try to know or create an app which is needed for the community, which the community will like or which is their requirement and then, we use it for the user experience. Now, collecting primary data and working on it and then user experience, where we are actually asking the user, how do they feel using the app, whatever app; it can be any entertainment app, gaming app, health related, any news app, whatever we have created or if you have not created any app, certain keyboards or certain type of translation system. So, whatever has been created, how are the users feeling about it. So, you will always get some feedback.

So, maybe they like it, but then they need certain modifications, certain changes should be there. So, when we use it or validate it through the users, then we get a better version of it and then, according to their feedbacks, we modify it. So, modification is needed. So, it is not like we have prepared something and it becomes the final. Unless and until, the users use it, no one can tell whether it is good or not or what are the modifications which

are required.

So, it is tested through the community members and then, a modified version is brought and then, based on the input, certain modifications are done in that app and then comes the final version of it. So, lots of steps are actually included. So, you have primary data and creation of the app or other technologies, whatever you are creating, then coming to the field and trying to see what are the reactions of the users and then modifying it, according to the comments or inputs you have received and then, comes the final version. So, these are the various stages of designing. Now, we also look at the requirements.

So, when we are planning to prepare something, maybe, you are planning to create such type of technologies for a language, you need to think first what do you want to create. So, it is it cannot be like "ok let me collect data and then I will see". What is done is that, at the very initial step, it is decided what has to be done or what has to be created. Whether it is a certain type of app or maybe grammar checker, spell checker whatever you want to create, that has to be decided at the very first step, because data collection depends on that. If you want to create something related to speech data, you will collect more speech data, but if you are trying to create something related to spellings, grammar checkers, spelling checkers, you will collect more written data.

So, based on the need and requirements of the community, it is decided what needs to be created for the community. So, whether we need words, sentences, whether we need it in audio format or textual format, may be, for certain cases we need both. Pictures, relatable to the community. So, maybe if you are planning to create some games, where you also want to put pictures related to the community, you need to collect those as well. Then also, the functional requirements: what things are actually needed by the users.

So, designing app based on the users, how can users actually use it. So, if you are thinking about something some device, where it can be used, so that the device should be available to the users. So, if the users they have android phones, then your application should be based on that. If your users are targeted so, mainly for the Mundari app, our targets were 5 to 11 years of children. So, the material which we are putting in the app should be appropriate for the children, it should be at their level, it should not be like very difficult or boring.

So, for the children we could put pictures in that, so that it is more interesting. And whatever material we are putting in the app or however we are designing it, we need to know our target user. So, if the users cannot read it, then you cannot put lots of text in it, rather you will have to give voice instructions. So, depending on their functional requirements, these things are decided. Then of course, the system requirements like on

what types of device, they will need to use it.

So, if you are designing something which can only be used when one is online and then, the village may not have internet connection always, then how will they use it. So, you need to always know their your target user and decide accordingly. So, something which does not need internet connectivity continuously, those type of things can be planned. And if you think that they have good internet connectivity, they have a proper phones and all, then those type of things can be decided. So, like for Mundari app that we created, we knew that the children (5 to 11 years), they were given android phones by the government during the COVID period for educational purpose, but then they had no educational material available in Mundari.

Of course, they were using it for learning through Bangla and English medium. Those materials were there, but then that is why we thought of creating some gaming type of app which can also be used as a pedagogical app, created for the audience which actually had Android phones. They were using the smartphones they had, but they did not know how to use it for Mundari because no material as such was available in that. And then the app was designed in such a way that it did not require internet connectivity because they do not get internet connectivity always.

So, once they download it, they can use it. The game was actually designed in that way. You should know about the target users. You can look at this Mundari app, you can see here, there are instruction written. So, when the app is opened, you choose the language of instruction. So, the options are many like Mundari, English, Bangla, another language, Mahali was also included later.

So, which language the user want to play in? So, we also had Bangla and English, because if someone wants to learn English through Mundari or Bangla, then also this app can be used. Then choose the types of games. There are various types of games like text based games, picture based games, these were basically MCQs. So, they have to choose the correct option. And then for these MCQs, there was different levels and these levels were based on the semantic domains like natural objects, body parts, fruits and vegetables, birds and animals, like that.

So, based on those semantic domains, there were questions which were in the way of MCQs and then, they have to choose the correct answer and they could choose any semantic domain of their choice to start the game. And this instruction was whatever is shown is actually in Mundari, but using Bangla script. Why Bangla script? Because this data was primarily collected from West Bengal and based on that, this game was created. And we also use this game with the community members in the field and we tested it on

the Bengali Mundari people. So, basically those who were staying in West Bengal.

The children were very excited to see this game because for the first time, they were seeing their own language in the mobile phones. They could read their language, they could listen to their language in the device; that is why they were very happy. But there were limitations like, we had only few levels and then, when children played these levels very quickly and they wanted more levels to be included. And when we were testing, we realized that all children could not read Mundari, but they can understand.

So, we also put the voice instructions there. The instruction which you can see or read, you can also listen to that by clicking on the voice instruction. So that way, one can know if the children cannot read it, they can know what they need to do. That was more convenient for them. So, the user has to choose the correct answer by clicking the buttons which are available. So, you can see that though this is written in Bangla script, but it is Mundari and there are 4 options given and there is also a countdown timer to make it more exciting.

Within that given time, they have to complete the task. Like, within 30 seconds, they have to complete it, otherwise its marked wrong and they do not get point for that. When they are actually clicking the right answer, they are getting one point to create some excitement and then, they can go to the next question. Now, there are also picture based questions. So, there is this picture where the arrow is shown and they have to name the object What do you call sun in Mundari.

That is what is written in Mundari. They can click the correct answer and again, go to the next level. At the end, they can get to know their scores. For all the questions, we also put the voice instruction. If they cannot read "what do you call this object" in Mundari, they can listen to it in Mundari and then choose the correct answer.

And at the end, they can see the final score. So, this final score, they can again try and then they can improve on it. They can also use this game to use learn Bangla or English, When they are learning English or Bangla, the questions will be in English or Bangla accordingly. So, that is how this was created. So, this gives just a scope for them to use it in the digital platform as well. So, when these children are playing different other types of game, they can use their own Mundari game and at the same time, another domain is getting introduced in their life.

So, they are using the language in the digital platform also. And simultaneously, they are also learning the words; if they do not know certain words, which were very culture specific, which there is a tendency that they are forgetting, they were asking their parents

and then, they were choosing the correct options That is what we saw, when we went to the field with the app to test it. So, when they are doing it, they are practicing and learning the words which they are forgetting otherwise. That helps in actually language maintenance and also can be used for language revitalization.

More examples like this can be seen across the world. We see lots of people are working on language technologies, they are creating various types of applications for the smaller languages. You can know about an app which were created for learning the indigenous languages of New Zealand, which is te reo Maori. So, because there was no one to speak the language, a chat bot was created, so that it can interact with the humans in the language. That way, humans can talk with the chatbot and learn the language. So, if they cannot find anyone to use the language with, they can use it with the chatbot.

So, that was very interesting way of or a scope is given to the user for learning their own language. So, the younger members, who do not know the language can use the chatbot and practice the language. So, that is a way. Then we can see another work where local communities archive linguistic data and provide teaching programs and apps through its First Voices platform. First Voices' latest innovation is a keyboard app that enables users to type in over 100 indigenous languages in any app in mobile device.

When there is the keyboard available for typing your own language or other smaller languages, it is also an encouragement for using the language because every time they are typing, they have to either type in English or some other language. So, it is so good if one can type in one's own language, the alphabets or the letters are available there. There can be different types of applications. So, I brought these two examples because these two are very diverse. In one, keyboard application is being created, in other, a chatbot is created and in another, a gaming app type of thing has been created.

What you can see is that there are diverse applications which can be created for these smaller and endangered languages, which can actually encourage them in using their own language. So, you can see that. So, the involvement of smaller languages in the development and implementation of technology driven initiatives is crucial to ensuring cultural sensitivity and maintenance of a language. When we see lots of applications being created in these smaller languages, that of course, helps in maintenance of a language and also creates cultural sensitivity. People know about those cultures, the languages also get exposed to the outer world.

So, when these languages are used in the digital platforms, lots of people who are from other communities also come to know about those languages about their culture. So, it gets exposed to the world, which is very important. A lot of companies are investing in

developing technologies for the smaller languages. We see lots of government organizations, lots of non-government organizations, lot of corporate companies and are also developing technologies for the smaller languages. Sometimes, it is done for the convenient of the customers, sometimes it is done also for developing or helping the smaller languages.

Language learning startups like Duolingo. You must have heard about it, it is very popular. People use it often to use learn new languages like German, French. Lots of languages are there which you can learn in a very easy and entertaining way, using this app. This app have diversified their product offering to contribute to the preservation process.

They have lots of languages which are smaller native American languages. So, they have also incorporated that into their app. One can also learn these languages using these apps. This is again technology building for learning the languages. Use of computational methods in the study of endangered languages have become a very relevant and popular topic nowadays. This title "Use of computational methods in the study of endangered languages" is a title given to a conference which regularly takes place.

So, you can see that lots of conferences are also giving platforms for presenting such works. We see lots of discussions and works going on for building language technologies for the smaller languages, but there are lots of endangered languages and very few people to work on it and then there are lots of challenges also. So, we of course, we are looking at the progress, we see it, but again, it is slow because of the various challenges which are there.

So, I hope you enjoyed today's class. Please go through these references. Thank you.