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-38

Lecture 38 : Language Documentation Across the World

Welcome to the 38th lecture of the course, Tools and Technologies of Language Documentation. Today, I will talk about language documentation projects those are happening across the world. There are various projects going on and what are the trends that we see in progress that is what I will talk about . We know that language documentation has evolved as a new field or sub field which is known as documentary linguistics. There are major documentation activities related to that which I will talk about and various projects which are running across the world and as I mentioned, we see a change in the approach. So, we see that trends are changing and I will also mention about that.

This overall lecture is based mainly on Austen's 2014, I have given the reference and to know it in a better way, you can go through the article which will be very helpful. So, when we talk about documentary linguistics, we see that in last few years, it has developed as a sub field of linguistics. And this field incorporates lots of things not only documenting a language, but also the outputs and also various things like rights and needs of language speakers of smaller languages, how they have the rights to use their language in different platforms, whether they can do it or not. So, all of those things are also involved in it.

There was also a concern for supporting speakers and communities who wish to maintain their language. So, it was not only about when a language is dying, if the community members also wish to maintain their language, but they cannot do it due to various reasons, then also documentary linguistics can help them. So, by providing documentation that could feed into revitalization efforts like creating materials or creating ah technologies. And also use of language in different domains, how that can be

done. All of those have become part of language documentation .

So, there are major documentation activities which we see across the project. Whenever we are looking at different projects, these five are the basic parts of it. These are the activities which are directly related to documentation or you can say that these activities are part of documentation or to have a successful documentation, we need to go through these steps. Recording- we record data including both audio video recording of the data. It should also have the metadata recording depending on the context and all.

Then when we have the data recorded, we transfer it to the management environment that means, somewhere it can be preserved. As you know preservation is an important aspect of the documentation. So, if we are talking about documentation, it is not only about collecting data, it is also about preserving it for a long time. So, it needs to be transferred to a place where it can be safely preserved. Then adding value to it, how do we add value to a documentation work or how do we add value to a data, which we have collected by adding more information about the data like transcription, translation, annotation, glossing, other contextual meanings and all.

So, when you provide more information related to the data, the data becomes richer and more valuable. So, adding value to the data and if no value is added to the data, we cannot use any outputs from the data. If it is only collection of words and sentences without any glossing, without any translation or transcription then, that data cannot be used. So, for using the data for outputs we need to add value to it, then comes an archiving. Archiving is very important because the data which has been collected or the materials which were created or the process which the data has passed through; you have done translation, transcription related to the data, everything needs to be archived in a proper manner.

So, archiving not only stores the data in an organized manner, but also gives access to others. And also talks about various copyright issues. Those who can get access to the data or those who cannot. Those type of things are dealt in archiving thing. And archiving can always help in giving access to the data; others can access data, they can use the data for various outputs.

So, that is important. And then the final and the most important aspect of documentation, mobilization. Some people nowadays think that documentation is not relevant or not useful unless and until we have some outputs. Because documenting a language keeping the data stored in an archive is of no use until and unless it is used for revitalization or maintenance of the language. So, when we are talking about the smaller or endangered languages we need to create some output.

it can help the language to be maintained. So, creation and distribution of the materials. Again, when we are talking about creation, we cannot only talk about like creating grammars, dictionaries or other materials and then keeping it to the researchers or only creating things for linguists and other researchers. We need to create those things, so that it can be circulated among the community. So, that the community members can use it for their benefit.

So, that is why mobilization is very important where you mobilize the speech community with the outputs. So, that they are motivated to use their own language and that can improve on the vitality status of the language. So, that is an important aspect of documentation, which was earlier not considered to be part of documentation, but now we see it has also been incorporated as a part of documentation or documentary linguistics. Different projects that we see across the world are: 1. the ELDP project by SOAS, where we see that ELDP has funded around 300 documentation projects and the endangered language archive holds around 100 collections of materials.

You can explore this and see lots of materials available, some are accessible and for some, you might need some permission. So, all of these are there. So, it continuously funds projects related to language documentation. We also know that lots of Indian languages have been funded under this fund and Indian languages have been documented using this ELDP fund. So, you will also find lots of Indian languages in ELDP project and of course, there are other languages as well.

The Volkswagen Foundation DoBeS project which was established in 2001 and was funded 80 research projects to a value of over 60 million euros. They are also funding till date and its funding came to an end in 2013. So, before that, they have funded 80 projects. So, we can see that it was a huge funding source which funded lots of projects related to language documentation. So, this project other than only funding people for or teams for documenting languages also created lots of projects, technology or lots of materials for documenting languages.

So, for video annotation Elan was created which is a software which can help you intagging or glossing of videos. So, basically you can if you have video recorded some sequence from the field you can bring it and you can gloss it, transcribe it, translate it using Elan. There are different tiers and it can be easily done. It has also created lots of recommendationswhich can be used as IMDI tools. There are questionnaires, there are various materials for data collection.

They have also developed those type of tools for documenting languages. Then there is

foundation for endangered languages. So, endangered language fund and UNESCO have this and they have again funded lots of smaller projects for documenting smaller languages. They also conduct conferences where they give platforms for presenting works related to endangered and smaller languages. So, this is foundation for endangered languages you can go into the link and look at it.

So, there are various grants given to the individual scholars to the teams where they can together work for languages. This has also been given to many of the Indian researchers and many Indian languages were also funded under this. We can see Documenting Endangered Languages Interagency Program of National Science Foundation. and the National Endowment of Humanities which has funded 100 projects and awarded approximately 30 million dollar in grants. So, that you can see and now it has become a regular program where National Science Foundation funds projects.

Living Tongues is again another you can say institute which documents languages; they also create various types of outcomes with these documented materials. They have also worked on many Indian languages. So, they have created dictionaries and grammars for various Indian languages as well. So, you can explore and know more about these. So, these were the major one of course, there are lots and lots of them, but I have not mentioned them.

And I am also focusing more on the global platform, because about Indian scenario, Professor Dripta has already talked about. Google's The Endangered Language Project is also a good funding where they give platform to preserve languages, language data and they have also funded some of the projects; they interact with different speech community members and various researchers and they work on language documentation. Microsoft also has projects like Elora, where they again try to create technologies related to minor languages. They are not collecting data as such, but they are collecting data for creating language technologies. Their aim is to provide some language technologies to the smaller languages.

They have worked on three Indian languages like Gondi, Mundari and Idu Mishmi. which is a language spoken in northeastern parts of India. Gondi is spoken in the central India and Mundari you already know. These languages they are working on. What we see is that they want to provide technologies for these languages.

So, all of these projects they are working and again you can see there are different dimensions to it. Some are creating platforms or tools for documentation like creating different types of softwares or questionnaires or different other types of toolkits for documenting languages. In other cases, we do see that they are creating technologies

directly which can be used by the community members. There are different types of funding and there are different dimensions to those. Now, we also see some of the efforts which are put by the community members.

So, community members they work for their revitalization of the language or for the maintenance of the language. In various cases, we do see that community members form their own group or they form their own organization, where they translate major text videos and they promote it in their own language. If there are books written in English or other languages,, those are translated into the native language which is used by the community members, videos are translated. In some of the languages, I have seen that the nursery rhymes which have been translated into an indigenous language and were uploaded in YouTube, so that the users can use them or listen to them. So, if the children start listening to their language, then of course, the vitality status of the language will improve.

They also use at times various types of digital platforms like social media. We often see lots of groups or pages where formed by the community members. So, they use these groups or pages for interactingwith in their own language. Also, they publish materials related to their language or any information related to their community in those pages. So, what happens is that the language gets a digital presence, the language is used in the social media and when few people do it then of course, it encourages others as well.

So, others also get included in that and they also start using the language and anything if published in that language is also posted. So, they know about that. So, that way also the language gets a platform where the language is used. We also know about various community schools. So, what happens are that the communities run some schools on a particular day or parallel schools where they teach the same subjects in their own language in their own native language so that, children can understand the concepts and at the same time, also practice their own language.

So, both way it helps the children. These are some of the efforts that we see by the community members who are very motivated to save their language or protect their language and these are again seen across the globe. We see these type of scenarios in Indian case, we also see those outside India as well. So, these type of efforts can be seen across the globe. When we look at the various ways or various projects which are working on language documentation, we see that there has been a shift. So, earlier recording, analyzing, preserving the records of language in use in ways that can serve a wide range of constituencies particularly the language communities themselves.

Earlier it was more about documenting and keeping and preserving the language, but

now we see that there are also more focus on the outcome thing. Earlier when the focus was more on language documentation, how to collect data, where to collect data, these were the main focus. So, the concentration on defining a model of language documentation was there. Specifying best practices, so what what the best practices to be done in the field, what can be the best tools to use, so these were the focus area where the process of collecting data was studied more and various ideas came related to that where documentation or the process of documenting the language was in focus. But now as we saw, there is has become a shift, we see that the goals or outcomes have become more important.

So, identify diversity of goals methods and outcomes of language documentation and the introduction of social models of research these have become quite important. So, what do we see is that when we are talking about goals, we know now that there can be different types of goals; there can be diversity in that and depending on the goal, the methods of collecting data can be also different. So, we cannot just say that these are the best practices of collecting data. Of course, we can say for certain cases, but then depending on our goal our method can be different. So, if earlier before going to the field, we know that we need data for this particular thing, this is our goal, maybe the goal can be creating a dictionary or the goal can be creating somelinguistic app.

So, depending on the goal our method of collecting data will vary. As we know that goals are diverse the methods can also become diverse. That has been accepted now. Nathan argued that despite the expressed concern by language documenters for recording language in its social and cultural context. Many researchers took an unscientific approach to audio recording.

As I said that the focus was on collecting the data, how was the process of documentation earlier. It was also seen that lots of people were working, but they did not know the right technique, they were not trained or theyadapted the unscientific way of. So, the basic smaller thing like where do you place the microphone or which microphone should you use all these things people were not aware of, but now, if you want to really know about this lots of documents are available which you can read and know. Again, earlier it was more specific to the field when you go to the field you keep the microphones in a particular position and you collect data, but now we are also talking about remote data collection where we are not using any microphone, the speaker is directly recording in his or her smartphones. So, things are changing because again related to the goals.

So, that way things are changing and of course, we always need to provide guidelines, we always need little bit of training for doing it, but we also can have variations in the

ways of collecting data depending on our goals. It was also a concern earlier where we used to see that speech communities or members of speech communities were objectified, they were treated as objects. So, like the researchers goes to the field collect some data from a person and comes back and this person was just like an object from whom some data was collected. But now, with the development of the you can say literature or language documentation, with the awareness trends are changing. We are nowadays talking more about involving the language speakers in the process of language documentation.

So, the process have become more inclusive. Of course, objectification was not appreciated earlier, it is not appreciated nowadays as well. Now, people have become more sensitive and they know that it is not the best practice, rather if speech community members are involved in the process, if it becomes an inclusive process, then we can get better results. In the earlier period of its development, there was a concentration on defining a model for language documentation and specifying best practices tools and analytical categories. Today, there is a recognition of diversity of goals, methods and outcomes of the language documentation and introduction of social models of research. We see outcomes at the community level which are very effective in language maintenance.

So, when we were earlier talking about language models of collecting data or what are the best practices, then again, we were focusing more on the researchers in a way. They were trained researchers who used to collect data, but now we can also see community members are working for their own language. We have become more inclusive together, we can work and that is why outcomes or goals can be there and as I mentioned, goals are focused nowadays. It is not about collecting data, it is also about what we produce.

So, this production can also be from the community. There are lots of literature being produced, lots of debate, conference, training program, workshops that we see. So, when we you see that there are lots of organizations working on language documentation or trying to preserve or help smaller languages, then we also see that lots of discussions and debates go on. We also see that funds are available for those who want to work on these aspects. And when there are more debates, conferences, workshops and training programs that also helps in overall language documentation, because when trained people will do it they can do it in a proper manner, when there are more debates and outputs then, that can also help in the overall documentation process.

Diversity in outputs that is what we see. There are lots of outputs which are diverse in nature. Some people are working on more writing grammar, text related things, others are working on creating pedagogical materials, while some people can create technology like

keyboards. So, there can be different types of things, outputs can be of different layers or different types. Earlier, generally, there was thought that outcomes can be like grammar, dictionary or maybe some picture book script, but now we see there are innumerable scopes and different things are being created after collecting data from the minor or smaller languages. I would like to conclude by saying the scope of language documentation has widened up in recent times.

We see that the scope has widened up so much that earlier it was only about collecting data, but now it is more about the outcome. When we were talking about collecting data of course, we focused more on our methods of our tools of the best practices, what should be done in the field what should not be done, but with the development of the technology we have seen that. Of course, we have lots of technologies for language documentation process like collecting data and also, we have seen that we can collect some data through cloud sourcing and remote data collection process. So, those have become also part of it, but of course, there are some ethics which stick to and then, there are certain good practices which we always follow. Earlier it was more focused on the process, but now we see the shift was the outcome.

A lot of government and non-government agencies are funding various documentation projects across the world. We see lots of government projects within our country and outside our country and also lots of non-governmental agencies are also supporting lots of smaller and endangered languages. Funding is helping in creating various materials and technologies for language documentation along with conducting training programs and workshops. When we see that people are funding or investing in language documentation thing or language preservation programs that helps in creating technologies and materials because of course, funds are needed anyway.

And those things can also help in training program. More people can get trained and they can work on language documentation and production of some materials for the smaller languages. If someone is just creating a small app in a minor or smaller languages, that is also a big step towards preserving the language or boosting the prestige of the language. So, that way also it helps a lot for the language. We know that when there are various projects across the world; the world is trying to preserve the smaller and endangered languages and that is a very good sign for the vitality of the smaller and endangered languages. Please go through these references. Thank you.