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-40

Lecture 40 : Summing up

Welcome to the last lecture of the course, Tools and Technologies of Language Documentation. So, this is the last lecture and, in this lecture, I am just going to sum up all the things which I have talked about in the course. So, I started with language endangerment and I talked about that what is language endangerment, how are the languages distributed across the world, what are the causes of language endangerment. And also I talked about relation between mental health and languages, linguistic diversity and biodiversity. And also you came to know about different factors related to language death, stages of language endangerment. By now you should know that languages do not die all of a sudden, there are various stages of endangerment through which a language passes before it dies.

And then also there are different degrees of endangerment as marked by UNESCO and also Fishman. So, you also know about various degrees or various ways in which we mark languages depending on the vitality status of the language and that is the way we access language endangerment. So, by now I hope you all know about the various features of language endangerment. It was important to know language endangerment to understand language documentation or the need of language documentation.

Why do we need language documentation or what are the factors which affect language documentation? And that is why it was important to know language endangerment which we talked about in our module 1. In our second module, we talked about variousattributes of language documentation. What is language documentation? We introduced you to the idea of language documentation, targets of language documentation as well as the outputs of language documentation. We also talked about various steps involved in language documentation and what is a good documentation corpus. What are the qualities of a good documentation corpus.

We also learned about designing questionnaires. These questionnaires, which you can use in both language documentation work and also in other field linguistics related works, where you can use these types of questionnaires. In tools and techniques of data collection, we talked in details about data elicitation and designing questionnaires. So, when we talked about designing questionnaires, you must remember that we talked about various methods of data collection or there can be different types of questionnaires. Picture books can be used, videos can be used, of course, there is translation method and then there are other methods like role playing or there can be other methods like just asking a questions about telling names of fruits, animals things, etc.

Whatever method we adopt that depends on our target, also depends on the target language as well as the target of a particular domain that you are aiming. So, if you are planning to collect data related to body parts then you can use picture books that can be a way of collecting where arrows are pointing towards different parts of bodies. But if you are planning to collect kinship terms, then you might not use picture book because that might not be very helpful. So, depending on the particular domain in which you are planning to work, you can adapt a particular method and similarly depending on the language as well. So, before we start designing our questionnaire, it is very important for us to know the speech community and that is why generally, we do a pilot survey, where we know about the basics of the community and likewise we customize our questionnaires.

So, that the questionnaire should be suitable to the speech community. And when I talked about customization of questionnaires. If you remember I talked about how you can change the names which are more commonly found in the speech community, how you can change the objects like names of the objects for example, names of the fruits or animals which are more common for the community. rather than keeping some names which are more foreign to them. Those are certain ways through which we can customize a questionnaire.

There are also different guidelines for each of them again, depending on the target there are merits and demerits related to it there are benefits and limitations. We talked in details about the limitation of translation method. Though translation method is the most common method which is used regularly and which we can use when we are starting when we are beginners we do start with translation method, but then translation method has its own limitations. So, I talked about that and I am sure when you are designing a questionnaire, you will keep all these points in your mind. And accordingly, you can design your questionnaire before starting your work.

And then a very important point which I discussed which generally, we do not hear

much about is remote data collection. And nowadays, with the development of language technology, we have seen that remote data collection has become very important because it can be made more convenient with the use of language technology. And also for the use of language technology, we need huge amount of data which can be collected through remote data collection method, crowdsourcing or other methods. When we are talking about remote data collection, there are certain general platforms, certain general digital platforms like social media, other type of YouTube channels and also those can be used to collect data of course, with proper permissions and copyrights. Sometimes, data is crawled upon from those types of websites, comments are collected and people work on those.

That can be done even for the smaller or endangered languages. Speech data can be collected from the YouTube channels for developing some speech related technology for a language. Those can be used and again there are certain specific digital platforms. We see that certain apps are being created across the world where these apps help you to collect data through remote location. These apps can be downloaded and used by the language experts or by the community members and then they can record their own data.

You have to just provide them with certain guidelines and you need to explain them the task and they can easily do it whenever they are free to do it. So, that way you do not need to visit the field and at the same time you are getting your own data. And so, that is a very convenient way of collecting data which lots of projects are nowadays adapting. These are some of the benefits of remote data collection, but all of these methods have their own limitations. In remote data collection also we find lots of limitations and one of the most important one is that we do not know our speech community like the rapport is not built with the community members.

So,we just give a task or we give a guidelines maybe we know two or three of the members and we explain them the task and they explain the others and it goes on like that. So, we do not know our field, we do not know the environment, we do not know the how they are living and these are very important to know when we are talking about language documentation. Because we are not only talking about capturing some words or some sentences, but we are trying to capture their world view, world view of a speech community. So, when we are not looking at the world where they are staying, when we do not know about their living style their day to day routine, then how can we know about their world and their world view. So, for capturing the world view for documentation, it is also very important to go to the field and document the language individually, rather than just depending on remote data collection.

It can be a part of the whole process, but not the whole process for documenting a

language. And again, when I am talking about the documentation of a whole language, if it is only for collecting speech for some technology, then of course, we can depend only on remote data collection. But overall documentation needs personal visit to the field and personal interactions with the speech communities. I also talked about oral literature. So, when we are talking about oral literature, then also we can see remote data collection cannot be a method.

because when you are interacting with people then you can know about their proverbs, riddles and all of those. So, remotely it might not be possible to collect oral literature. And oral literature is very important to be collected because through oral literature, we know about the morals, we know about the sayings, we know about the relation between the community and the outer world, how they are looking at the nature, how they are perceiving animals. So, all of these are hidden in oral literature. And that is why it is very important to know oral literature and also lots of historical facts are hidden in this oral literature, that is why also it is very important to document oral literature.

I also talked about how to collect oral literature and various challenges related to it, what are the challenges which you can face while you are working on folklore or oral literature of a community. Whenever we are talking about literature, please remember that literature is not only those things which are written literature can be oral. So, when it is not written we do not find any document written about the literature, then also it is a literature part of the literature. So, oral literature has its own features which I have already talked about. In my module 4, I talked about methods, ethics and challenges in the field.

One of the very important aspect of language documentation is metadata creation. Metadata is all about the data. It is not about the linguistic data as such, but the information related to the data. That is very important so that we can preserve the data in a organized manner and also so that the data can be easily retrieved. It is very important to have a proper metadata.

In metadata also, we collect details of language experts that we do when we are in the field while collecting data. Prior to that, we also collect detailed information about the language. Also, we always maintain a log book where we carry details of each of the session. So, all these information are always recorded. Metadata is created some part of metadata like the basic information about the language expert or each session these are collected when we are in the field.

And when we come back then we process our data and then again another set of metadata is created like from which file a particular word has been sliced out. and all

those information. There are also differences between thick and thin metadata, I talked about the utility and use of metadata and I also talked about some metadata models which are used across the world in various projects. So, these were all about metadata. Metadata is very very important aspect of language documentation.

So, I also talked about history of language documentation, how the approaches of language documentation is changing with time. How earlier the focus was mainly on documenting a language, but now slowly we see there is a shift and people are talking more about the outcomes. And scholars have also realized this that outcomes can be of diverse types. Earlier outcomes were imagined to be limited like grammar, dictionaries, primers all those things, but now we know there can be innumerable types of outputs like there can be picture books, there can be technological aids, there can be apps related to gaming. There can be various types of things, there can be keyboards, there can be different things created for these smaller languages.

Because the outcomes have become so diverse people are talking about more outcomes and that is also so because when there are outcomes given to the speech community, then only there are chances of maintenance of the language, then only the speech community will get something and that getting of something helps in adding prestige to the language which can save a language. Because when we are talking about language documentation, it is not only about preserving a language in a digital format, it is preserving it within the community. So, the outcomes are very important. So, we see there are different approaches to documentary linguistics. I also talked about the timeline of different types of field works and different types of field works depending upon the community and its relation to the land.

So, what can be the challenges faced in the Indian field while documenting a language, that also we discussed and there can be challenges related to linguistic area and there can be challenges related to technical issues. I mentioned those there are different types of disadvantages which we can face there can be different types of challenges that we can face which can be related to the location, again gathering of the unnecessary people, there can be problems created by those self proclaimed leaders, but then we should know how to tackle with them and move ahead with our work. These can be some of the challenges which we should be ready mentally prepared before going to the field that ok these type of challenges we can face and we should know how to tactfully handle those without offending the community members or without hurting their sentiments. In Lecture 20, we had an interesting discussion to show you what are the good practices of conducting field work.

I hope you remember this lecture. So, I and Professor Dripta were talking about our own

experiences our experiences of different fields and what are the problems challenges that we faced and what were the good things that we experienced we were talking about all those things. We were sharing our experiences in the hope that you will learn something from our experiences. And from this discussion we also came to know that how we should adjust to a situation, how we should know to adjust because we have seen certain people, certain field workers who go to the field and then they complain because they are not getting their home comfort. So, they might complain the place is very hot, there is no AC, no fan or no electricity or certain type of problem the food is not of their choice. So, these types of complaints we cannot do when we are working on a language in a field.

We should be prepared well before, we should have our equipments ready, we should have like rechargeable batteries which are charged. We can have power bank and every equipment ready with us before going to the field and we should plan our sessions properly, but then there can be situations which we do not know about and depending on that we should also be flexible with our plans. So, that if there is certain emergency or if our language experts are not feeling well they cannot talk to us we should leave them, we should be flexible, we should not be like we have to complete this word list in one day and so, we are just doing that. And we should give have time limitations for the language experts like we should not make them sit for more than 2 hours. So, if they are or they have some important work we cannot force them to sit with us and talk.

And most important thing that is needed for good language documentation or field work is rapport building. So, when you are in the community you should be create a rapport they should be your friends in some time. So, that you both the community member and the researcher together can work for the language and that is where we can get good results, where both come and work together and then rapport is built up. And when you are back also, you can call them you can know about their problems their day to day life you can chat with them and that is how rapport is built. And we have seen for most of the good field workers or those who are working on language documentation they have very good rapport with the communities where they are working.

So, because those help in the work. module we talked about different dimensions of language documentation. We talked about how to be ethical in the field and after the field. So, when we talk about ethics, there are certain ethics which we need to maintain in the field and also when we return from the field and we are giving outputs, we are writing something, then also we need to follow certain types of ethical practices which are very important. And it is not only important for the report or it is not only important for the sake of document, but it is also important so that this rapport that I was talking about is also maintained. So, when the practices are ethical then generally the rapport is maintained, if it is not so then somehow the community members they feel hurt, they feel

cheated and then of course, this rapport will be broken.

I also talked about Sapir-Whorf hypothesis to show you how language affects our worldview, how we perceive the world is shaped through our language or you can say both effect each other. The world surrounding us affects our language and the language we speak also affects the world which we perceive. So, those things I talked about in Sapir-Whorf hypothesis. In Lecture 23, we shared experiences of community members from Mahasui and Sylheti community. So, they talked about their experience and we found that Mahasui is an endangered language and how documenting folklore can be difficult in this language.

And at the same time, we saw that Sylheti is not an endangered language, but some parts of it is becoming endangered like kinship terms or food related items. We also saw how prestige was related to Sylheti in certain parts, but not in the other parts. So, where Sylheti is often clubbed with Bangla and then speakers of Sylheti also wants to identify themselves as Bengali speakers. So, when prestige is related to the language, people do identify themselves with the language, but when they feel that there is no prestige, then sometimes, there is a tendency of shifting their identity to the major or the bigger language. During this discussion, these are the points which we found that the to find the correct location where people speak the language is very important.

So, when we are talking about documenting any aspect of a language, we need to know the exact location where people use the language. Because, when we are talking for the smaller or endangered languages, what we see is that sometimes people claim that they are part of for example, say Mahasui community or they are part of Sylheti community, but they cannot speak the language, they cannot use the language. then if you go to that location or you go to those families who identify themselves with the community, but cannot speak the language then you cannot get any data. So, that is important. Secondly, when sometimes people can speak, but they want to hide it.

If you are speaking Bangla, then maybe a Sylheti speaker will reply in Bangla instead of Sylheti. So, that will also you will not get the data. So, the language used by theresearcher is also very important on the responding language. Some aspects of non endangered languages can be danger like kinship terms, traditional food items, language documentation should be more inclusive. So, at times we see that the community members themselves will say that do not go to that woman, she is illiterate or do not go to that man, he does not know anything he cannot read or write.

So, he cannot tell you about our language, but that is not the thing; literacy has nothing to do with the way we speak our native language. or we should be more inclusive, we

should not think about gender or we should not think about literacy, we should be more inclusive and we should try to include every type of gender, every type of people, people from different backgrounds. So, our data will be more rich if we can involve everyone with different type of backgrounds. I also talked about script and language, how script and language is related and how when a language has its own script it adds prestige to the language. So, one of the outcomes of language documentation can be giving a script through the language.

And when there are no scripts for certain languages, often people take it as a dialect or a bully. They do not consider it to be a language, but it is not a fact. We as a linguist, we know that every language is a language irrespective of the fact whether the language has a script or not. Script adds prestige to the language and also helps in a maintenance and revitalization of the language because when a language has its script, then various materials can be created written in that script so that is important. In our sixth module, we talked about processing the language data.

Till now we talked about various things related to field how you can collect data. It was all about collecting data and then how can you process it for various types of outcomes. Data organization and cleaning that is very important when you have data you need to keep it organized. And then of course, you need to delete the unnecessary pauses, unnecessary noise. So, maybe a dog was barking or other type of sound you can delete those.

So, those things are needed. I also talked about data transcription where we do use IPA and when we are talking about IPA and IPA chart, we also need to know little about human speech sound. And I talked about various types of human speech sounds, manner of production, place of articulation, how to create phonemic inventories. We discussed that and I think that that can be helpful though that was very limited and you should go to further studies for that, but this was to give you a basic introduction about the sound systems of human languages. I also talked about glossing rules and POS tagging. We follow Leipzig glossing rules to gloss our data which is again very very important.

Whatever work we want to do further on the data, we need to either POS tag our data or do a basic glossing mostly we do that. For some cases like speech data we might not do, but mostly we do it for other types of applications or for writing grammar, writing dictionaries. for all those things we need basic glossing. And of course, transcription, glossing these are important equally important is translation of the sentences, translation of the narration narrative text that you have collected. So, all these things are important and are part of processing of the data.

Then coming to language maintenance and revitalization which we talked about in seventh module, we talked about what is revitalization and what is language reversal. So, these are important terms because when we are talking about language documentation and outputs of language documentation those are often used to revitalize a language. So, that the language is saved or learned by the younger generation. What are the factors for language revitalization and Fishman's model of revitalization those I discussed. And also, I discussed how revival process can be affected by inclusion in the age schedule of a language.

If a language is included in the age schedule, then how it adds prestige to the language. And then we see lots of fundings are there for the language, the language is introduced in the academic area and slowly the language can become healthier its vitality can improve. So, there can also be various initiatives of revitalization from the community itself. So, if the community is well motivated to revitalize its status, then we can see that at times, there are lots of revitalization programs within the community. So, they practice their language, they try to teach the language to the youngsters.

Again creating digital dictionaries, writing grammars, designing pedagogical materials, these can also help in revitalization of a language. I have talked about all this ah. In the eighth schedule ah I talked about the case studies. You have seen various projects which are going in India which Professor Dripta talked about and I talked about various projects which are going across the globe related to language documentation. Use of language technology for endangered language was also discussed, difference between computational linguistics, natural language processing and language technology was discussed and then I also talked about different types of language technologies.

So, developing language technologies for minor languages and why it is important. There are various stages involved in designing an app. what are the requirements for those. So, those were discussed with an example from Mundari gaming app which we have developed.

And I also talked about certain other examples. I just brought one or two examples, but there are various such examples where you can see that organizations are working for building language technologies for the smaller or minor languages. And then again there are different types of technological aids which are being created. So, all those things were discussed. I also talked about documentary linguistics as a sub field and major documentation activities, different projects across the world and change in overall aspect of language documentation.

So, these things I talked about. Of course, this course was on a basic introduction to

language documentation, there are various other aspects which I could not talk about this is the limitation of the course. And I am sure if you are interested, you can explore that and this course is not only about learning a course, getting some marks, its applicable and it would be a very good effect of the course if you can document certain aspects of a language. So, that is very interesting thing to explore and I am sure you can use your knowledge from this course to do the actual work in the field. Again certain limitations of this course were that we could not talk about various types of softwares, we could not do hands on on that. There are various softwares like Lexipro or Flex or Elan which are used as part of language documentation for processing the data which you can explore.

Again I did not talk about monolingual field work, how do you collect data when there is only one language used by the community members, they do not know the linked language that you are using. I have given you a link in the suggested reading. which you can watch and know how you can collect data for in a monolingual situation. And these are again certain suggested readings you can explore these to know more about this domain of language documentation. I hope you enjoyed this course. Thank you all the best.