**Tools and Technologies of Language Documentation**

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

**Department of Humanities and Social Sciences**

**IIT Kharagpur**

**Week-02**

**Lecture-07**

Lecture 07 : Language Documentation

Welcome to 7th lecture of Tools and Technologies of Language Documentation. Today I will talk about what is language documentation. So till now you have talked only about language endangerment, language vitality and all. So from now onwards, we will actually talk about what is language documentation. We will talk about targets of language documentation, outputs of language documentation and steps involved in language documentation. And of course, qualities of good documentation corpus because when we are talking about documentation, we should also know what are the measures to check the quality of the corpus that has been collected.

So we begin with this definition which I think everywhere, wherever language documentation is done this definition is referred to. Language documentation is a lasting, multipurpose record of language. So why this definition is important? Because it talks about lasting, multipurpose record. So 'lasting' means it has to be long lasting.

That is very very important and this definition focuses on that. Because when we are talking about language documentation, we want to preserve the language samples. You know, we generally document languages which are endangered. Of course, we can document languages which are not endangered, but the focus is more on endangered languages because they are dying. So we want to save it through documentation and revitalization.

So when we are trying to save languages and we are collecting samples, then it should also be long lasting, it should not die out or it should not get ruined. That is why it is important to store it properly and that is why, long lasting is very very important feature of language documentation. Secondly, it is multipurpose record. There are various outcomes of language documentation data. So the data which we collect or the corpus which we build can be used for various purpose.

So what are those things, we will talk about that. So basically, language documentation is a lasting multipurpose record of language. He identifies the following as important features of documentation. (Himmelmann, 2006). So he focuses on primary data.

So what do we mean by primary data? Primary data is the data which is actually collected from the speakers. It is not collected from any book or any journal. Again why it is important to collect primary data? Because when we are talking about secondary data, that means they are already being documented somewhere. They are in a form of book or they are in a journal, they are in some YouTube channel, somewhere they are already there. That is why it is important to collect data which is not documented.

That is the primary data which is actually being used by the speech community. So if someone is collecting data, one should be accountable for that. It should not be like I have collected some data, but if someone challenges me and I say I do not know about it; I am not accountable for it. It should not be the case and when we talk about accountability, that means we should also know the details of the data, that is we should know about the metadata: from where the data was collected, when it was collected and who were the speakers who were involved in the process. So everything should be noted and those also come under the part of language documentation.

Because if we do not take notes for that, we might forget and as I have mentioned, with time language changes. So after 10 or 20 years, one might challenge my data. One might think that the data I have collected is not accurate because language has changed. But if we have the date, time, place, everything recorded, we know that the data is 15 years old and that is why there are some changes in the language. So those who are collecting data should be accountable for that.

Concern for long-term storage and preservation of primary data. So if I am collecting data or if I am documenting a particular language, I should also be concerned about its storage because as I mentioned, the data should be long lasting. So how can the data be long lasting? That means, we have should have a proper storage of the data. The data which one collects should be preserved properly, so that it does not get destroyed. Another important feature is working in interdisciplinary teams.

So when people work in the field of language documentation, generally, the work is done with an interdisciplinary team. So there are people from anthropology, there can be people from computer science, there can be people from different disciplines. So it is not only about linguists who are documenting language, there can be scholars from different fields, who actually contribute in language documentation. Generally, when we talk about

language documentation, it is done by a team. Individuals can do, but it is not language documentation, it can be a part of a language which can be documented like folklores, proverbs, kinship terms.

So certain parts which you think are getting endangered can be documented by a particular individual. But if we are talking about whole language documentation, then generally we need a team. There are lots of equipments involved, there are lots of process involved, not only collecting data, but also preserving it, glossing it, meta data. So lots of works are involved which cannot be done by an individual. So what are the targets of language documentation? When we talk about language documentation our first target is of course, collecting or documenting linguistic practices and traditions.

How the speech community actually practices language, how they talk in their day to day life. So it is not about talking, but it is also about how they perform various rituals and practices. So when we document language we not only document their day to day talking, but also we look at the rituals, how do they perform. We can also look at how marriages are performed or how various other religious performances are performed. So all this actually comes under linguistic practices.

So when we are talking about language documentation, we of course, collect linguistic practices; we also go beyond that and that is metalinguistic knowledge. What do we mean by metalinguistic knowledge? Sometimes languages are used in certain way, in a certain context because we know language is context dependent. So if we meet someone, in Bengali culture people might say "Bhalo?" which means "Are you good? Are you fine?" which is a way of saying "Hello" actually. But, in some places, in Telugu they will say "Have you eaten your breakfast?" which again is a way of saying "Hello" which actually they mean "Are you well fed? Have you eaten in your home?". So that is a way to greet.

So there can be different ways of greeting. In some parts of Uttar Pradesh, you will see they will take names of some gods. So that is just a way of greeting each other. People might say Ram Ram, Radhe Radhe, anything that they like. So there can be various ways of greeting.

And this knowledge is actually metalinguistic knowledge, why? Because one needs to know when to use it and why it is being used, it is not only related to the word and its meaning. So if I say in Bangla "bhalo?", "bhalo" means "good", but to use it in a particular context, when you are meeting someone for the first time in the day, you can say that person "Bhalo" or when you are meeting someone after some time and you are greeting the person by saying "Bhalo" which means "Are you good? Are you fine?" then that has a meaning. So one who knows Bangla or one who has learned Bangla may know

all the syntactic rules and the words, but when meet someone and the other person says "Bhalo", this person might get confused; why all of a sudden this person is saying "good" to me, what does that mean? But if he has that metalinguistic knowledge, he will know that that means he is just greeting the person. So like when I was in my Mysore, some people used to ask me "Are you well fed? Have you taken your breakfast?" in the morning and I used to feel why do they ask me about food every time they meet me in the morning? So actually I could not understand it, but then after some time when I stayed there for a longer time, I understood that is a way to greet. They are actually greeting me; they are meeting for the first time and they want to know that whether I have taken my food properly or not.

So this is a metalinguistic knowledge which does not come from the words or the syntactic rules. For that, one needs to know the linguistic practices of the community and that actually varies from region to region. So languages might be same, but sometimes these practices can be little different. So when we are talking about documenting a language, we not only need to collect the words and the sentences and maybe the oral literature. We cannot be restricted to those only.

We also need to document the metalinguistic knowledge which is little tricky because one can go with a questionnaire, get the words translated, get the sentences translated or just get narrations and all, but to collect metalinguistic information is little trickier because you cannot get them; just you go with a recorder and you can collect it, it is not like that. But if one stays with the community and one documents the language, stays there slowly one can actually document it. And we have seen lots of such cases where people can actually document the metalinguistic knowledge as well, along with the linguistic practices. So these are the targets of language documentation. The other one is systematic recording and transcription.

So the recording should be systematic and then there has to be transcription to it. So the audio files are transcribed in IPA. So that is also part of language documentation. It is not only about one going to field, recording lots of data, lots of speech data and just filing that in somewhere, because they also need to be arranged properly, so that it can be used for various other purpose. So as I mentioned that language documentation corpus can be used by various other disciplines, so it is also important to keep it in a systematic manner.

And when I say that the data is kept or arranged in a proper way, they have been transcribed in IPA, we also need translation and analysis. And when I am talking about analysis, we are not going to adapt any particular theory. In language documentation, data is not looked through a lens of a particular theory because that can create bias, rather there are very basic analysis. Translation is important because we do not know what is

being said. Only those, the team which has visited the field will know what is being recorded.

So when one comes back, one needs to transcribe it and translate it so that the data can be used by other scholars as well. The data also needs to be analyzed, basic linguistic theory is used so that means, only a very basic analysis is done for the data. So that one can use one's own theories if one wants, one can use the data for other purpose like developing dictionaries or pedagogical materials. So for anything the corpus can be used. That is why the documentation data does not actually use lots of different types of theories to analyze the data.

So the outputs of language documentation, as I mentioned when the data is collected it is transcribed and translated and a basic analysis of the data is given, so that there can be outputs created from this data, so that it can be used; the corpus built through language documentation can be used for various other purpose. So, linguistic and metalinguistic data are stored for future study, interdisciplinary studies can be done. So, sociological work or anthropo-linguistic works are done based on these data and then there can be production of various things like dictionaries, grammars, teaching tools, collection of folklores, revitalization programs can be done based on the documented data. So if we have proper documentation of a language, then materials can be created which can actually help in the process of revitalization. And that is actually the ultimate goal of documentation, if the data is documented and stored somewhere and then no one is using it, then it is useless.

But if the data is properly arranged, analyzed and all, then it can be used for various types of productions which can actually help in revitalization of the language or if the language is endangered, then it can be given to the speech community so that they can use their language. So if the language does not have dictionaries or grammars that can be produced out of it. Recently a grammar of Akabea was written. So Akabea is a dead language; it is no more spoken. The first grammar of traditional Great Andamanese language was written according to current linguistic standards.

So how was that possible? Because there were documented material for Akabea language and based on that documented material, the grammar was written. So now, one can ask why the grammar is important? The grammar is important because it helps us to understand the history and prehistory of the place, Great Andaman islands. We can know about the speech communities and that particular area through the grammar. So Akabea, as I mentioned died in the 1920s, but still there were two British administrators who during that time actually documented the data. During that time, we did not have digital recorders and all.

Still the data was so well documented and well preserved that through that, linguistic grammar was written and which was recently published. So we know, if the language is properly documented and preserved, we can actually use it after lots of years to create grammars or various other materials and that is why it is important to preserve it properly. So steps of documentations involve recording, which is the first step actually; of course, before that we need to know what are we going to record. So recording includes everything audio, video, image, text, everything. So recording does not only mean digital recording.

Of course nowadays we use digital recorders and digital recorders are there, but along with that we also collect text, text in the form of marriage cards, calendars, journals, books, anything can be there and those are also taken into account for language documentation. So various types of recordings, capturing of moving materials to digital domains like performances; performances are there and then there can be day to day rituals or day to day works like fishing, agriculture related, cooking, there can be different ways to that. So all of these are captured because all of these actually involve language. When we are doing anything in our life, in our day to day work whatever we do we always use language; there are terms for everything we see and do right. That is why it is very important to capture as many things as possible; those can be actions, those can be static object, everything is actually captured and then analysis of those things.

One can go to the field and record fishing and come back, but that also needs to be analyzed. One needs to know what is the person doing, what is it called in that particular language and all those things. So transcription in IPA and translation and annotation of the data, all of these are there and of course, metadata. I will talk about metadata in details after some lectures.

So metadata is basically data about the data. So as I mentioned where we collected, when we collected, who were the speakers or participants involved, so all those are noted down. So those all things are also very important in documentation. It is a systematic process, it is not just a random process that someone feels like I will document a language, goes to the field, records something and come back. It is not like that, it is very very systematic. One needs to know the steps, one needs to know what to document, what to record, then records accordingly, comes back, cleans the data, annotates it, transcribes it, translates it and then also creates storage for it; where to store it properly so that it is preserved.

And that is the fourth step which is archiving, how is the data to be preserved. So archives are being created where data is kept properly. So across the globe, there are

various projects going on related to language documentations and they have their own huge archives where data are stored properly. So that one can access it if required and these data are saved there. Then comes mobilization which is basically publication and distribution of the materials in various forms.

So, if the data are kept properly in the archives and they are locked there then of course, that will not benefit the language. So it is important to create publications or other materials from those data. There can be dictionaries, grammars, pedagogical materials, picture books, different things can be created out of those data which can be given to the speech community. So that they can use their language which can help in promoting their own language and that is why that is very important. Earlier documentation was taken as only documenting a language and archiving it, but nowadays we see a focus more on mobilization as well.

So what are we giving back to the speech community, so that the language documentation data becomes useful. So recording a good documentation corpus will include audio and video materials ideally recorded in authentic setting and in good environments. So which means that there should not be much noise and there should not be noise of people laughing or talking or wind. So there can be different types of noise. So it should not be that because when we are collecting audio records, it should be proper.

So that is important. Good equipments are needed with trained people to handle it; we cannot just record in anything. Firstly we cannot do that because some devices they are not good and they can actually not record the audio files properly and secondly it becomes a problem for storing these data when they are not in standard format. That is why it is very important to use good equipment. Interdisciplinary teams are there. So the team members who know how to handle the equipments, team members who know how to deal with the people, how to collect data, team members who know how to build archive or save the data, So there are lots of people involved in the process.

You will also know that we talk about ideal informant very often. So in field linguistics also, people talk about ideal informant, but in language documentation, when we talk about ideal informant or as I mentioned, there can be ideal setting for language documentation that is very very difficult to find. Because, when you are there in a speech community, you are collecting data, you cannot ask everyone to be silent. So generally these are rural settings and they have a community feeling.

So lots of people are there. If you are there, they see a visitor in the village and lots of people gather, you cannot ask them to just leave and record one person. So it becomes

difficult and when people are gathering, they will talk, there are animals, there are children crying. So you cannot actually make everyone stand still. When you are in a village you will have to adapt to the setting of the village. But still, one can try for the best environment, one can take one person in isolation and then try to collect.

One can try to collect inside a room, not outside in a very windy place where there are because there will be noise for the wind. One can avoid places where there are lots of animals because animals will make noise. So one can actually try little bit or one should not consult an informant or language expert who is drunk or chewing pan or something. So we can try our best though we cannot reach the ideal situation, but then one should of course, try for the best. Sometimes there are certain situations where you can just record them.

So you cannot create; there can be natural thing happening like maybe fire or marriage or some ritual. So there can be different types of things happening in a village and you need to document those of course, then you cannot think of an ideal situation. In marriages, there will be lots of noise and all thing. So what one can do is that one can document all whatever one is getting and then, nowadays we have good technology, we can always delete noise and make our recordings little better, but of course, we can never manipulate the data. So speaking and writing are conceptually different activities and we know that the way we speak we do not write in the same way, the ways are different; varieties are little bit different and so, that is important for us to collect both oral and written forms of data.

In creating an authentic literature that can be rooted in oral tradition, the researcher can encourage and assist the people to find their own ways of developing new modules of expression rather than just following the written dominant language model, which is often there. People try to follow the dominant language pattern, but we cannot do that we can always try with some new modes. So, when I am talking about written data, we can go for registration papers, cards, newspaper, journals, various types of books, marriage cards and all those things. So they also help us in looking at the language; how actually marriage cards are written or how calendars are actually prepared.

So all those things can be documented. So when we talk about good documentation data, Woodbury talks about diversity. Diversity is very important because language is a thing which we use in every domain of life. So it is very important to collect as many domain as we can. Language in a home domain will be used very differently than the same language being used in a very formal setting or in rituals. So one should try to collect as many domains as possible, so that the data is diverse.

We can know more about the language. The data should also be large in quantity; that means, we should build a large corpus out of it. Because with few sentences I go somewhere, I collect 500 words or 500 sentences. Through that, we cannot produce anything; that way we cannot save a language. That is why the data collection or the corpus should be large or huge. It is an ongoing process, data can be added to the corpus from whatever sources that are available and can be expanded with new materials becoming available.

So the one can document certain language from the field and then one comes back and sees that there are new journals which have started, then one can again document those. So, it is an ongoing process. So that should be there and why it has been said to be opportunistic because sometimes you can see something is happening there in the speech community, you cannot wait to take their permission maybe, you can document at that moment and of course, then you should take permission afterwards and without permission of course, we should never document anything. The data should be transparent; there should not be anything hidden or something which has been documented without permission that should not be there. Data should be preservable and of course, portable; nowadays we carry data, we can save it somewhere, download.

It has to be portable and as I mentioned, it has to be very very ethical. We cannot actually document anything with a hidden recorder or hidden camera that is unethical. Secondly, we cannot manipulate data whatever we have collected. So the whole process has to be very very ethical. So these are some of the features of good documentation corpus. So language documentation is all about documenting every aspect of language, which also includes metalinguistic information.

We need to collect or document every usage of a language; wherever the language is used we need to document it. It is not only about collecting the data, but also about preserving it properly so that it can be long lasting and so that it can be reused it can be used and reused. Language documentation is a multidisciplinary work which involves trained scholars from different disciplines like linguist, anthropologist, computer scientist, folklore studies and biologist as well. So we see that people or scholars from different disciplines come together when we talk about language documentation, because again, language is a thing which enters every discipline and that is why scholars from every discipline can contribute in the process. So these are the references and some of them are easily available and freely downloadable.

So I believe you will read this and I hope you enjoyed today's class. Thank you.