**Tools and Technologies of Language Documentation**

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

**Department of Humanities and Social Sciences**

**IIT Kharagpur**

**Week-03**

**Lecture-15**

## Lecture 15 : Methods for Remote Data Collection

Welcome to the 15th lecture of Tools and Technologies of Language Documentation! Today, I will talk about a very different topic, that is methods of collecting data through remote ways. So, remote data collection. So, we do not generally see people talking about this much, but nowadays, we see lots of people actually practicing it. So, I thought I will talk about it. Today, I will cover the following topics, like what is remote data collection? What do we mean by that? Methods of remote data collection. So, how can we do that? We can use general digital platforms which are already available.

We can also use some specific digital platforms, which are actually designed for data collection, for linguistic data collection or linguistic documentation. And I will talk about the benefits of doing remote data collection, why is it helpful for us and the limitations related to it. So, let me begin with what is remote data collection. We know when we talk about primary data collection, we believe that researchers or a team of researchers visit a particular place, they meet the speech community and start collecting or documenting the language samples.

So, these are generally trained linguist, anthropologist or other technicians, who visit the field together as a team and start collecting or recording data, using various methods that we have discussed. So, that has been a traditional practice and it is still practice today; that is fine, but nowadays, we also see what we do is that, we use lots of digital medium, we use digital recorder, digital video and audio recorder and all those things. Earlier people used to write on their notebooks, but of course, now we have shifted to digital medium. We also see in certain cases, that people record in their smartphones, if they have a good recorder. So, that is also done.

So, why to travel to collect data of course, it has its benefit, but sometimes due to certain constraints, people can actually ask the language experts to record in their smartphones using various applications. So, that is remote data collection where the researchers they do not travel, they do not need to go to the specific speech community. Rather they can stay in their place and ask the language experts or train them and collect data from them. So, that can be possible. It is convenient for financial constraints.

So, if the researchers they do not have much money to travel to the place, to stay there for a week or a month, and when you are travelling, you also need to carry your equipments along with you, So, there are also logistic issues. So all these can be avoided, when one shifts to remote data collection. Secondly, during pandemic, we saw that there was no social mobility, people could not travel, but during that period also, data collection works were going on. And then how was it done? It was done through remote location. So people, the researchers they did not travel to the exact location, but they collected through remote ways, using various digital medium.

So, it mainly also bloomed due to language technology development, because when we talk about endangered languages or smaller languages and we are talking about preserving them, of course, we talk about language documentation, we also talk about language technology building. So, when we are building language technology for these smaller languages that means, we are giving a domain, opening a domain for the language users to use the language. If the language users can find their language in various technologies like machine translation, text to speech, speech to text and all those applications are available in their own languages, then they can use their own language; they do not need to shift to the other dominant language. So when we are talking about language technology, we need huge amount of data. So, this huge amount of data, like say 50 hours of speech data, this means huge data and which cannot be collected by the researcher going to the field for a week or 10 days; that is not possible.

So in those cases, what is helpful is crowdsourcing, when we are collecting data through remote location. So we cannot stay there, we cannot go there, but we ask them to provide us with their data. So, because of digital presence and because of building language technologies, remote data collection has become a very important aspect of data collection, which are very useful for the under resourced languages, when we are thinking of building some applications for the under resourced languages. So, what can be the methods of remote data collection? There can be the general digital platforms. So, we all know about Google forms, we use Google forms for various other objectives.

So, what we can do is that people actually create Google forms as a questionnaire. So, they put their questions in the Google form. And it can be very helpful for collecting language attitude related data or for collecting language vitality related data, where actually we know Google forms are used more by the younger generation, not by the older people. So, this younger population, they can actually tell about their language, whether they use it or not, whether they use it in certain domains, in which domain they can use it, in which domain they cannot. In fact, in Google forms, we can insert pictures and ask them, what do you call this in your native language or what do you can you talk about or write about a song which you sing in your own language.

So, various data related to attitude and vitality can be collected through Google forms, which we see people are doing. Then there are other mediums of online video meetings and conferences like Google Meet, Teams, other means. So through these what happens, you actually meet the language experts, video meeting is being created and you are asking several questions related to the language and collecting information. Again, this method is useful for collecting language attitude or language vitality related issues. So, because how many sentences or how many words can you collect through video meetings or through Google forms; that is not possible.

You can collect a sample, but not much of data. Secondly, when you are talking about Google forms, we cannot collect linguistic data because that is in written form and we assume that most of our language experts are not good with IPA unless and until they are trained linguist. So, you cannot ask them to type the word or write the sentence, but you can ask other questions related to attitude about the language through Google form. In video meetings also, maybe you can ask certain words or certain sentences, but not much because of time constraint. So, data collection through other social media.

So, we know that there are various social media pages available, there are various blogs being created by the speech communities. and specially so for the smaller language groups. We see there are lots of Facebook pages, blogs for the smaller languages, where people actually interact in their language or they also post some important message related to their language, may be some new book has been published in their language or things like that. So, what can be done is that one can collect those discourse from the social media. In that way, one can know how the language is used in social media or how the language is used in the written form.

So, whether they are shifting to English or some other language or they are using the same language in social media, you can know about that. Little bit of linguistic data can be collected and again lots of language attitude related and vitality related information can be collected through these blogs and pages. When we talk about YouTube channels, you can actually collect speech data and these speech data are actually important for building various applications like text to speech, speech to text, because this is the actual speech that they use. So, there can be speech related to certain description of the place or there can be song channels, there can be certain smaller channels where they teach their languages. So, these channels can be helpful in collecting speech data as well.

So, these are some of the general digital platforms which we use in our day-to-day life for various other purpose, but we can also use it for collecting data. And then there are also specific platforms. So, what are these specific platforms? These are actually being designed for collecting linguistic data. So, we have seen methods of data collection through Google Forms. As I said, various questions mainly related to attitude and vitality can be asked through Google Forms.

This is helpful for the language experts. Why? Because they can actually sit in their free time and fill up the data. So, you are not hampering their daily routine. It is not like you are in the field. So, they have to leave their own work and sit with you.

At night, when they are free they can actually fill the form. And at the same time, you are collecting data from different places. So, if a language is spoken in 5 and 6 different districts, you are collecting data from all these districts at the same time. So, a Google form is circulated and you give like 10 days duration within which they have to fill it. So within this 10 days, you are collecting data from whole country, you can and even beyond that you can do.

You can also get data from the diasporic residents who are staying outside their own village. So, that you can know about their own attitude towards their language. So, that can also be done. Secondly, Google forms are very useful for the younger generation, when they have migrated to other cities for studies or job, they can tell about their language or how much they remember their language, you can collect that information through Google form. And you can also use it for different age groups because sometimes we know senior people or elderly people, they cannot use these forms, but then there are younger people to access them.

So, through that also it can be collected, but when we are talking about Google forms, we also need to have a well description or instructions written in the form. So, that they can understand the questions and reply accordingly. So like any other ways of collecting data, in remote data collection also, the questionnaires and the prompts are designed according to the target. So, if you are planning to collect say, language attitude related data for a specific age group of people, the questions should be like that. And again, if you are collecting data for say some ethno-linguistic information, then the questionnaire will of course, be different.

So, depending on the aspect the questionnaire will vary. Socio-linguistic data can be collected through video conference, we have seen certain examples, this is a reference from a recent paper, where they talk about how they actually collected socio-linguistic information through video conference during COVID-19. So, language attitude, ethno-linguistic information, limited number of narrations and words can also be collected through video conferencing. According to the location of the community, decide which medium to use, because we know for video conferencing, we need good connectivity. But for Google forms, weaker connectivity can also be useful.

So, depending on the place where people are residing, what is their internet connectivity, how is it, what are the devices that they use, whether they use smartphones or they also have computers and laptops, depending on that, one should actually decide or choose the method. If the community does not have a proper internet connection, then video conferencing can be actually very disturbing and that they have reported in the paper, I mentioned. They have said how video connectivity issues actually hampered their work or created disturbance in their work. So, it is difficult and again for video conference, there has to be time-limit, you can sit for 1 hour, maybe 2 hours, not more than that. So it becomes a limit; in limited time, you have to collect data and within that limitation also, there can be several times of disconnectivity.

So, those issues are always there. But other than that, there are other methods where social media is used, you can collect anytime you want, you can crawl data automatically or you can actually collect manually, depends on you. So, those things can be done. So, for social media pages as I mentioned, from Facebook pages or YouTube channels, not speech data can be collected from YouTube channels, but there are several comments below each of the videos. So, those data can also be collected for various types of works. And to see how the language actually is used in social media or for commenting people, sometimes we will see that people are actually cursing the video.

So, you can know how aggression is being shown through the language. So, all these aspects can be collected through comments. So, that shows about the language different dimensions of the language. So as I mentioned, sometimes these data can be crawled automatically using various applications sometimes, you can do it manually, but whatever you choose to do, you should always cross-check the copyright issue; that is a very very important thing. So, sometimes we see that copyrights are open and you can use it.

So, there are creative commons copyright, but then in certain cases, there are copyright issues and you need to take permission from the owner. Sometimes you can easily get the permission from the owner for various types of academic or research works. So, that also needs to be remembered, we cannot forget about the copyright things that is very very important. So, then coming to these specific platforms which are created for data collection. So, since remote data collection has become very common nowadays and especially with the language technology needs, so nowadays, what we see is that there are various platforms which are specially being designed and created for collecting data.

These platforms can be used in the smartphones or in your laptops. And these are helpful for both the data provider and the one who is collecting data. Of course, both of them needs to be trained in the application, one needs to know about the application and then one can use it very efficiently to collect data. So basically what happens, these platforms give you a chance of collecting data through crowdsourcing. So, lots of people are actually joining to your project and supporting you with their data.

So, both the data annotator and the provider can be trained and other people like annotator, translator can join the project simultaneously and all can work together with a huge amount of data. So, let us look at some of these type of applications; one is Linguistic Field Data Management and Analysis System. So, this is a system where you can store, manage, annotate, analyze and share your linguistic data. So, you can collect through remote locations and also, you can store your data which you have collected from the field. So, what happens this applications also give you a storage.

So, your data is safe, you need to preserve your data. So if you remember, in my first class I was saying that language documentation is also about preserving the data properly. So, this gives you a scope to preserve it in a safe place manage your data. So, that they are not all hoch-poch, they are managed in a proper manner. You can also use the annotator to annotate the data, analyze the data, translate it, transcribe it.

So, everything is basically done in the given platform. And this platform is useful because anyone can download it in one's smartphone and use it. So, you can actually train the language experts and ask them to download the app and they can download it and they can record their own speech or record according to the prompts given into it. So, these applications also give you the scope to upload your questionnaire. So, your questionnaire will vary depending on the field.

So, you can upload those type of questionnaires, what are your needs from the language experts, what do you ask them to do. So, there can be different types of prompts like sentences, you can ask for narrations, you can give picture prompts, you can also upload smaller videos and ask them to narrate it. So, basically whatever you are doing personally going to the field, you can actually do that all through this app. So, it is just that you are not going in person to the field, rather over the telephone you are instructing the person and then that person is working in the app. Or sometimes what is done is that, workshops are held where one or two members from the speech communities are invited and they are trained in the application.

So, basic things like the recording should be done in a silent place, the voice quality should be good. So, these basic trainings are given to them and when they return back to their speech community, they tell or pass on this information to the other members. And then everyone, the whole speech community can gather in giving you the data, can record their data. And that can be done at their ease of time, whenever they are available they can record their own data; whatever data they want to record, they can record it. And in some cases, in some projects, when there are options for remuneration, according to the data which has been provided by the data provider, remuneration is given to the person, who has provided you with the data.

So, that is how the apps work. We see a similar app like Karya, which also works on crowdsourcing, it is again an android app used for data collection. So, basically Karya app is again downloaded and it is shared by both the data collector and the data provider. So, they are again trained with the basic use of the app and they can sit in their home, in a peaceful environment and record their data. Whenever they are available, whenever they are free, they can record the data. So, Karya also actually has another aim, which is, for this smaller communities, we have seen that, women they do not have source of income or some of the people they do not have much source of income, so what they can do is that, from their own home, they can sit and provide you with the data and at the same time, for the data, they can earn something out of it.

So, that way also, Karya wants to help the smaller communities. So, these basically LiFE app and Karya app, they are important for us because they are Indian originated ones. So, they are being developed and designed in India and they are specifically designed for Indian languages and they are used in Indian speech communities. So, they are developed in our own country. There are other tools like Robson, Kobo Toolbox, these are again some options through which data can be collected and they are not the Indian ones, but they are used across the globe; they are also available.

So these are some of the specific applications, which are there for collecting remote data. So benefits of collecting remote data: huge amount of data can be collected in a limited time. So if you want to document a language, you can actually collect different aspects of the language, you can collect huge amount of the language in a very limited time. It is cost effective because you are not travelling to that area. You do not need the fare of travel, you do not need the accommodation fare, you are saving all this money actually.

So, your cost effective, it is cost effective, it is hassle free. So, when you are traveling, you need to book your tickets like 2 months prior to your travel, you need to carry heavy equipments like recorder, tripod, camera, everything. So team member, they should be together, they should be available and free for going to the field in a particular time. And then when you are visiting a field, you have to also look at the climate conditions of the field. In certain regions it rains heavily, so data collection will not be possible if you are recording, it will be raining and the sound will be there.

So in certain places, there can be snowfall, you cannot visit during that time. So you do not need to think about this, when you are collecting through remote method. So, due to natural calamities or due to political unrest at times, we cannot visit certain places, we need some permissions; there can be various issues. So, we can avoid all these hassles by not going to that place. Language experts can record according to his or her convenient.

So, when we are visiting a particular field, we generally expect the language experts to sit with us and record the data, but that time may not be convenient for the language experts. May be it is their time for doing certain works. So, they may not be available. So, the lady of the house might be cooking or doing some job going out.

So, may not be available. So, how will you collect data? So, it is good if they have the applications in their smartphones and they can record when they are free, when they feel they have time, they can record. Again, they are recording, they are keeping it on pause, returning back after two three hours and then again continuing with the job. So, they have the liberty to do it; it is not like you are sitting with them and you asking for the data.

So, they have to sit and entertain you. So, they are free. So, these are some of the benefits of collecting data through remote. Limitations there are of course, lots of limitations of remote data collection. First and foremost is that no rapport is being created between the language experts and the researchers. So, you do not know them personally. When we talk about going to field, we interact with the language speakers, we talk to them, we try to know about their lifestyle and slowly, if you have experience field or if you are about to experience.

You will know that when you go to a field, you know about the geography, you know the people and slowly, they become your friend, you know them. You talk to them, they talk to you, sometimes you return back to your home and you keep the contact over telephone, you talk to them. They tell you if something interesting is happening in their village, if something exciting is happening. So when we go to field, we generally have that friendship with the language speakers, which is very useful for knowing the language, for knowing the community, but that is totally absent in remote data collection.

We do not know much about the language experts. We can know about two or three of them, whom we have trained, but most of the people, we do not know, we do not know the exact location. And as I mentioned, observation method is very very important in documenting a language; that is totally absent in remote data collection. You have not visited the village, then how will you know about the village? How will you know what is actually practiced in the village? So what happens is that sometimes, some people can record something, they can give you certain information, but that is not actually practiced in the village. It is not like the person wants to lie, but it is that the person is not conscious that they are not actually following it in practice. They are just giving you the situation which should be there, it is more prescriptive, but what is actually happening, the descriptive one is missing out.

So, that can be a issue with remote data collection. Depends on availability of the devices. So when we are talking about remote data collection, we generally expect to have the speech communities devices for collecting the data like smart phones, internet connectivity, but if these things are not present there, then how can we collect data through remote data collection process; it is not possible. Language experts have to be given proper instruction or training. Without proper instruction or training, they may not be able to provide you with the data or the data might be so noisy that it is totally useless. So, it has wasted time for both the language expert and also for you, but then the data is not useful.

So, that is why proper instruction is very very important. Authenticity of data can be questionable at times. So as I mentioned, that people can just say anything. In one of my lectures, I was talking about fake lores. So that can be a thing. People are just making up something and just recording it for the sake of say, money or for the sake of any political agenda or some other propaganda.

So in those cases, we do not know about the authenticity of the data. So, that can be a issue with data collection through remote methods. And secondly, you are not present there; we are not present with the language experts. So if they have any confusion, they might need to consult us over the telephone. We are not present at that situation to help them.

So, that can also be a issue in remote data collection. So, these are some of the limitations of data collection. So, what we see is that remote data collection has become very popular in the recent times and more so post COVID. In less duration time, we can collect actually a huge amount of data which is very useful, specially for building language technology. Different aspects of linguistic data can be collected through different specific applications like grammatical aspects, collection of foreclose, attitude data. So, different aspects can be actually collected through remote data collection method.

But still, a vast part can be missed out, which is necessary for overall documentation of the community because there is no observation method; we do not know about it. Meta linguistic information can also be missed out because that is not being recorded as such. So, lots of these aspects can be missed out when we are talking specifically for language documentation. We cannot only depend on remote data collection. Of course, it can help us for building language technology applications, it can help us with various other aspects, but overall documentation of the speech community is not completed unless and until we actually go to the field and record data, record metalinguistic aspects of the speech community.

That is why, we can balance both of our works with remote data collection as well as visiting personally to the speech community. So, I hope you enjoyed today's class. Please go through these references, these might be very helpful to you. Thank you!