Tools and Technologies of Language Documentation

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

Department of Humanities and Social Sciences

IIT Kharagpur

Week-07

Lecture-32

Lecture 32 : Creating Dictionaries

Welcome to the 32nd lecture of the course, "Tools and Technologies of Language Documentation." Today, I will talk about designing dictionaries. So, how do you create dictionaries? By now, you know that creating a dictionary can be a very useful outcome of language documentation. It can be easily created, and at the same time, it can be very useful for both academicians and community members. But then there are various challenges and aspects to consider when creating dictionaries. I will talk about lexicography.

The sentence is already grammatically correct. It is all about creating dictionaries; how do we make dictionaries for endangered languages? What are the different types of dictionaries and the structures of dictionaries that we discuss in microstructure and macrostructure? And basic dictionary entries, along with challenges related to them. So, I will talk about them. Lexicography is, in fact, the activity or occupation of compiling dictionaries.

And studying the science of making dictionaries and the problems related to it. There can actually be various challenges related to creating a dictionary. So, there are various practical sides to it, as well as theoretical aspects. Both are studied under lexicography and are referred to as practical lexicography and theoretical lexicography. So, if you want to study lexicography, it is actually a sub-branch.

So, you can learn all about making and creating dictionaries in lexicography. When we talk about lexicography, why do we use the term "lexicography"? It is derived from the words "lexeme" or "lexicon." So, a lexicon basically means a mental dictionary; that is, we all have a set of words in our minds, right? So, we use it according to our needs, and

this is called the mental dictionary. We have a collection of words for all the languages that we know, and these words are in their basic form. We add other information as needed.

Now, in one of my classes, I talked about morphemes. So, you can think about that for a little bit. Without the morphemes that contain information, there is extra information, or without the bound morphemes. There are also free morphemes, which are the base forms, as you can see in these examples. The root form of words like 'rah' for 'rahna' in Hindi, which means "to stay," is correct.

Lexemes are those words, those root forms without any extra morphemes, or you could say, Those basic forms, or root forms, carry the meaning, and when other suffixes or prefixes are added to them, That creates or provides extra information. If I say 'rahti', it means I am referring to the feminine gender. So, that is why it becomes "rahti." Or if I say "rehta," that means I am talking about the masculine gender. So, by using '-ti' and '-ta,' you are getting extra information while the base meaning of 'to stay' remains the same.

So, similarly, you can see other examples. If you talk about the English examples, such as 'cook' and 'cooks'. So, "cook" here means that the verb "cook" refers to "to cook" something, right? But when I say "cook," that means, as you will understand, I am referring to the third person singular: "He/she cooks." So, the bound morpheme is added to provide that extra information. So, those are inflections, suffixes, and affixes.

So, they will come under inflection. So, inflections are those that provide additional information about the same root verb. Those can be related to numbers, such as book-books, or they can be related to verbs, such as '-ing'. So, those are over there. Now, you can also see that 'cook' can be entered as a verb in a dictionary.

Similarly, 'cook' as a noun should also be included in the dictionary because we have 'cook' as both a verb and a noun. So, when I say "cooks," I mean those who can cook. So, we know that there are chefs in the hotels. So, cooks are those who actually cook the food. So, the meaning actually differs slightly.

So, that is a derivational suffix. Similarly, we can also have words that involve compounding. So, two words are joining to create another meaning, like "handbags." So, the 'bags' that we carry in our hands are called handbags. In Hindi, you can see examples like 'bacha', and from 'bacha', we get 'bachpan'.

So, 'bachpan' is not the same as 'bacha', but it is derived from 'bacha'. An example in

Bangla is: Natok-Natokyo. So again, from 'natok', 'natokyo' has been derived, which means drama and dramatic; however, the two are different. Though the meanings are similar, one is derived from the other. Yet the meanings are not the same, like 'cook' and 'cooks,' singular and plural—it is not like that.

But now, when we talk about lexical entries, what will we do? We enter all the words. We enter "cook." We also provide the extra information that they are either a noun or a verb. And how we can form the plural by adding '-s', and then we also provide another entry for the new lexeme information.

Like cooker. So, "cooker" will be another entry. We cannot say that we have already included "cook." So, we will not include 'cooker'; it has become a new entry. So, it will be there again. So, each lexical item is equivalent to one word, and all the words of a language should be included.

In that dictionary, as far as possible. New words can be created either by derivation or by compounding. Where the two words actually come together to give you another meaning. Like "armchair" or the different types of compounding that you see in various languages. Now, there are dictionaries of endangered languages.

For endangered languages, the most useful and relevant research outcome can be dictionary resources. The sentence "Why?" is already grammatically correct. Because dictionaries can be easily circulated among community members, they can gain access to them. Dictionaries help in creating various types of pedagogical materials, picture books, and other resources. Dictionaries also help in the standardization of the script.

So, if the script has been used with proper orthography or according to a rule for spelling, Then, that will be standardized and followed across the community. As you know, when we are confused about the spelling of certain words, we consult dictionaries. Or we are confused about the pronunciation of certain words. We do consult dictionaries; these can be online dictionaries as well, but we do consult them. So, when community members forget their language, they forget certain words.

They can consult the dictionary to find out the exact pronunciation and the exact word. That way, dictionaries can actually help maintain the language. Similarly, younger children can learn from the dictionary; they can use it, and dictionaries also help in. As I mentioned, the goal is to standardize the orthography and spelling rules of a language. So, dictionaries are very useful and functional texts; they are always used and referred to.

They can be used in various domains; for example, they can be used to create

educational materials. They can be used for creating various types of apps, including gaming apps, which are actually a collection of words. They can be used for various types of language technology, such as morphological analyzers and others. So, for all those, dictionaries can also be very, very useful. In recent years, there has been a trend in endangered language lexicography to produce small dictionaries focused on semantic fields.

So, maybe you are working on a specific aspect of a language. So, maybe you feel that the names of birds in a particular community are becoming endangered. And you just want to create a dictionary for the names of the birds or the names of the flowers. Or just the kinship terms, food items, or anything else that you want to create.

That can also be created. So, it is not that dictionaries have to include all the words of the language; they can also be domain-specific. So, you might have seen dictionaries, such as a dictionary of physics and other dictionaries on various topics. So, there can be dictionaries for particular semantic domains, such as body parts or kitchen items, or anything else. So, dictionaries like that can also be created, and they are easier to create. It will take a lesser amount of time.

So, you are planning to document the entire language and then produce some outputs. So, compared to that, it will be easier and will take less time if you focus on one particular domain of a language. You work on it, and then you create a dictionary for that particular domain. Which can be very useful for community members. And if you think that particular domain is endangered, then circulating that dictionary among the community members would be beneficial.

It can actually help revive those terms. So, these are particularly suited for maintaining a language. These are called mini dictionaries, or you may have heard the term pocket dictionaries. There are smaller pocket dictionaries available. These provide the speech community with quick access to a dictionary of its language.

For use in schools and the community in general. So in general, these words or these mini dictionaries can be used. So, these mini dictionaries can be for a particular semantic field, or they can also be for overall language. So, maybe some people are working on a language. They have collected a few words, such as 10,000 or 8,000, and then they come out with these mini pocket dictionaries.

These dictionaries can also be very useful for an endangered language. When we are creating a dictionary, it is very important for us to know for whom we are creating it. Who is the target audience of the dictionary, and who are the target readers? That is a

very important format, and the mode of compilation of the dictionary depends on it. How we compile the dictionary depends on our readers and the region. Because when we talk about dictionaries, we should generally think about the community members.

Dictionaries are very easy things to refer to, and if you give them back to the community, they are very useful for that. Now, when we think about the community members, we should know about the region where it will be used. If you are planning a digital dictionary and we know that there are community members, They do not have access to electricity or the internet; then what will they do with a digital dictionary? How will they access the digital dictionary? So, that will be of no use to them. So, we should understand the users and plan our theme accordingly. We can create or design our dictionary in a digital format and then print it out for circulation among the communities.

In certain dictionaries, a CD or a pen drive is also provided along with the hard copy. So, if they have access, they can listen to the sound; otherwise, they can read the hard copy. The health of the language and the degrees of endangerment. So, if the language is seriously endangered, then a small pocket dictionary is also helpful.

Depending on that, we can use it. If it is just for general terms, then it is very small, but it is still very useful. The community's attitude toward language literacy and learning is important, so we should consider it as well. So, if the community is trying to develop its own script, If a new script is available for the community members, which has already been established, If it has already been standardized, then it is always good to include that script in the dictionary. Because that will also help spread the script, which is being newly designed for the language. So, those different aspects should also be considered before creating or designing a dictionary.

The target audience also includes community members, professionals, learners, and others. Sometimes, some dictionaries are created only for beginners or learners, while others are created specifically for children. So maybe, if you feel that in a certain community, children are not learning the language. Or maybe children do not know the names of the local flora and fauna, so a dictionary is created for them. Then, of course, it should look very attractive to the children.

It should have colorful pictures, and there should be example sentences that children can read and understand. The target audience should always be kept in mind before designing a dictionary. When it is created only for academics, there are different types of information in it. There can be more detailed grammatical information available. So, all of those can be present when it is created for both the academics and the community members.

Then what do we do? We include both IPA, which can be useful for academicians and others. The corrected sentence is: "And also the local script in which the language is written." (Note: The original sentence is already grammatically correct. However, if it's intended to be part of a larger sentence, it may require additional context or connection.) So that the community members can read it, only IPA may not be understandable to them.

So, incorporating all those features, we can design a dictionary with our target readers in mind. Now, there are different types of dictionaries. (The original sentence is already grammatically correct.) Before addressing that, I will also tell you about this book, where you can see a picture of a woodpecker. This is, once again, a Great Andamanese dictionary written by Professor Anvita Abbi and her co-author.

So, what they have done is that they have only collected the words related to birds in the dictionary. So, there are many bird names in this dictionary, and the names are provided in Great Andamanese as well. Their biological names are given along with their descriptions, including how they look and what they do. When are they found, when do they migrate, and what are all the details? So, in that way, they have tried to actually save or protect the names related to the birds. Because the bird names were endangered, people were actually forgetting the names of the birds that were found in that area.

So, it is just a dictionary dedicated to bird names. Here in this picture, you can see a dictionary of the Kurmali language. Here you see that it includes many of the words that have been collected from the field related to the Kurmali language. You can also see Bangla, Kurmali, English, and Hindi written. So now, the words are not only in Kurmali but also in Bangla, English, and Hindi.

So, the meanings are written in Bangla, English, and Hindi. So that speakers can read it. Why have both Bangla and Hindi been included? Because Kurmali is used by both people from Jharkhand who are more accustomed to reading Devanagari script and Hindi, They also use Kurmali, and again, we see parts of Purulia. Where Kurmali is used, people more often use the Bangla script or are more fluent in Bangla. So, both Hindi and Bangla have been included, along with English, as we know that people from anywhere can read.

That is why English has already been included. That means meanings are provided in English, Hindi, and Bangla, and there are also words written in IPA. So, the IPA is also present. So now you know that dictionaries can have several types of entries; it depends on how one wants to design them. So, there can be a monolingual dictionary that consists

of the language. So, only the word—the main word, known as the head word— The word is given, and its meaning is written.

So, it is monolingual; the meaning is also provided in that language. There can be bilingual ones where you see that the English word is given, and the meaning is provided in Hindi. Or in any other language; those are bilingual ones. And then again, I talked about this Kurmali dictionary. This is a multilingual document, where we see the Kurmali words alongside their meanings written in Bangla, English, and Hindi.

So, these are multilingual dictionaries. (The sentence is already correct.) Now, regarding the information, when we talk about dictionaries, we do not only provide the meaning. We also provide a little bit of information because there can be certain words for which there may not be a one-to-one meaning, right? So, there may be various words that cannot be translated into a single word in English or any other language. So, there can be encyclopedic information that is actually detailed about the headword.

So, the headword is the main entry in the dictionary. So now, if there are "clay horses." So, if a word is given in Kurmali and its meaning is written as 'clay horse' Then, certain encyclopedic information is also provided to it, such as: When are these clay horses used, and in which festival are they used? Many dictionaries also provide information on pronunciation. Pronunciation is also provided, which is mostly written in IPA. Any dictionaries you open nowadays, including online ones, will generally show that pronunciation is written in IPA.

You will see how exactly the word is pronounced. Some dictionaries exclusively provide the historical origins of words and how they have originated. From where they actually come, those meanings or things are already mentioned. Sometimes, some words are archaic; they are no longer used. So, those mentions are already present. And then again, as I mentioned, there can be specific domain-specific dictionaries, such as a legal dictionary or a medical dictionary, depending on that.

There can be dictionaries that are talking dictionaries; that means you can actually listen to the sound. You can also interact; there can be sentence examples, or there can just be simple hard copy dictionaries. Nowadays, we generally see both versions of dictionaries. The same dictionary can be found in both hard copy and soft copy.

There are various structures involved in the dictionaries. One falls under macro-structure, while the other falls under micro-structure. So, what do we mean by this? In microstructure, we discuss the overall formation of the dictionary. How is the dictionary organized, or how will it be presented? So, that includes the front matter. Sometimes,

some dictionaries provide you with guidelines on how to read the dictionary.

So, maybe they will give you the phonetic chart for the language. Which sounds are present, and which are represented in IPA according to those guidelines? There can be guidelines related to various punctuation marks used in dictionaries. There can be several types of guidelines. (The original sentence is already grammatically correct.) There can be an introduction to the language and the speech community.

Some dictionaries also include sketch grammars. Here is the corrected sentence: "So, a very basic grammar of the language will help you understand the words more easily." And the various processes, like inflection and derivation, that I mentioned can be easily understood. A very basic grammar is provided in the introduction. How is the body of the dictionary going to be? So, like the entries and definitions, how will it be organized? The organization of the dictionary will be alphabetical. So, most of the dictionaries that you find for English are alphabetic, but there can also be thematically organized dictionaries.

There can also be dictionaries that start with Z instead of starting with A. So, those things can be there. "Thematic" means that there may be dictionaries where body part terms are listed first. Then other terms related to cooking, kinship, or animals come up thematically. What will the end matter, such as appendices and additional information, be like? The names of the volunteers or the language experts who have provided you with the data may be.

Map the location where the language is spoken and where the data has been collected. There may be names of places where different varieties of the language are found. So, all that information goes in the macro thing. So, it is the bigger aspect of the dictionary, the broader one. And now, when we narrow down the topic, the minute details, such as the internal structure of the dictionary, emerge.

What will the headword be? So, the lemma form needs to be looked up. So, this main entry is called a headword or lemma. So, which script should be followed when you are writing the headword? Will you write it in IPA, in local script, or in some other script? So, that needs to be decided. Are you going to write the pronunciation in IPA or not? Variance: This is very important if you are creating a dictionary for, say, Mundari. Which version of Mundari will you include? The Mundari language, which is spoken in Odisha, or the Mundari spoken in West Bengal. The same is true for the Mundari spoken in Jharkhand, as it is for many other languages we observe.

So now, which Mundari are you going to examine? Which Mundari word will be your head word? Of course, the language is the same; there will be many words that are the

same. But then, there can be words that are different. Differences in pronunciation. So, you have to decide which variety you are going to use as the headword.

Because only one entry can correspond to the headword. And sometimes, it is mentioned that this word is pronounced like this in Odisha. But it has other variants that are spoken in West Bengal and Jharkhand. So, that variant information can also be entered in the dictionary. Morphological information, such as whether it is inflectional, derivational, or compounding, is important.

So, there are various types of morphological information related to the morphemes. Paths of speech information. (Note: The original sentence is grammatically correct but lacks context. If you meant to convey a more complete thought, consider rephrasing it, such as "There are various paths of speech information.") So, whether the word is a noun or a verb, or if the same word can be used as both a noun and a verb.

So, all those types of information. And the gloss, which is, of course, the meaning. Now, gloss as well; you need to decide whether it will be in the same language. Or will it be in some other language, or will it be in two or three languages? So, that also needs to be decided; it can be in English, Hindi, Bangla, or Tamil. So, in which languages will the meaning be written? Again, in some dictionaries, we do see the use of sentences. How the word is used in the sentence is also mentioned.

Corrected sentence: Along with the script of the sentence. So, if you are thinking of including an example sentence, will the example sentence be written in the script of that language? Or will it also be written in IPA? So, that also needs to be decided, along with the meaning of the sentences. So, if you are including three languages in the gloss, then you should also include all three languages for the meaning of the sentence. So, all these things need to be decided. This is an entry for the Lambani language, where you can see the word 'katmarlo,' which means 'family.

' Now, this is an entry that was made in a software called Lexipro. So, you can explore Lexipro; it is freely downloadable software specially designed for dictionary creation. Now, what you can do is enter words, include pictures, and add sound. So, you can see that there is a symbol for sound. So, when you click on that, you can listen to the sound.

You can see the meaning, a sentence example, and a picture, all of which are being entered here. So, there is various software, and Lexipro is very common; it is used for designing dictionaries. So, you can add historical notes, cross-references, still pictures or audio, and you can add semantic domains. You can add cultural information, sources, and other information as well. So, a lot of information can actually be added to the dictionary.

The more information is added, the richer it becomes, and it is very beneficial to include at least pictures in dictionaries.

Especially for flowers that are very unique to the community, as well as other related items. So now, when we talk about dictionaries, we also notice that there are certain types of words. Which are homonyms, meaning they are unrelated but sound similar. So, we should also be careful about this. We should also be careful about polysemy, which means related meanings. Now, for example, you can see that 'head' means a body part, which is actually the head, right? but then it also means head of an institution, it can also means one side of a coin head and tail, it can also mean upper end of the bed or table.

So, now 'head' has all these different types of meanings which might be related to a head, but then they are in use. So, how do we also incorporate all of these. So, that is also a challenge for thecreating a dictionary. So, these are some of the basic dictionary entries where we do enter the head word written in the script and also in IPA, there can be sound, main sound; how do we hear because nowadays, mostly we talk about digital dictionaries, digital dictionaries can be created and at the same time, hard copies can be taken out from those.

Parts of speech are entered. English meaning, example sentence, might be in IPA and the script, meaning of the example sentence might be in English and other languages, encyclopedic information, picture and again, video and other information can be used or can be entered in the dictionary. So, now, I would conclude by saying dictionaries can be an important output of language documentation. Language documentation can be very useful in creating dictionaries, where you can see that dictionaries help in language maintenance and also, in the process of language revitalization. Mini dictionaries take lesser time to be produced and at the same time, it can be easily circulated among the community members for their easy reference. Dictionaries can also be helpful in building various types of language apps and language games or pedagogical apps.

They can also be used to create various types of technologies related to language like predictive keyboard, morph analyzers and others. So, creating dictionaries are a very easy way or a very easy outcome, but then it also includes lots of challenges. So, when we are creating a dictionary, we should be very very careful about our entries. We should decide what are the entries those we are going to put in the dictionary. And when we are doing the work, we should be responsible for it, we should do it very very carefully because once we spread it or once we circulate it in the community, it becomes a standard.

Because people look at the dictionary as a standard, people refer to it and take it as the right way of writing a word or speaking a word. That is why this work should be done

very carefully and all the challenges involved should be taken care of properly. These are the references, please try to go through them. Thank you!