elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 1*. Editor: Peter K. Austin

The opportunity and challenge of language documentation in India

E. ANNAMALAI

Cite this article: E. Annamalai (2003). The opportunity and challenge of language documentation in India. In Peter K. Austin (ed.) *Language Documentation and Description, vol 1*. London: SOAS. pp. 159-167

Link to this article: http://www.elpublishing.org/PID/013

This electronic version first published: July 2014

This article is published under a Creative Commons License CC-BY-NC (Attribution-NonCommercial). The licence permits users to use, reproduce, disseminate or display the article provided that the author is attributed as the original creator and that the reuse is restricted to non-commercial purposes i.e. research or educational use. See http://creativecommons.org/licenses/by-nc/4.0/

EL Publishing

For more EL Publishing articles and services:

Website:	http://www.elpublishing.org
Terms of use:	http://www.elpublishing.org/terms
Submissions:	http://www.elpublishing.org/submissions

The opportunity and challenge of language documentation in India

E. Annamalai

1 New needs and forms of documentation

Documentation is creating a record of the past or the present for the future. The nature of documentation as regards content and form depends on its purported use and its perceived users. The documentation of language has received academic and public attention in recent times as a response to the disappearance of languages accelerated by the modern economic and political structures. Languages, however, have been recorded from the earliest historical times on stones, barks, leaves, plates and cloth in the form of texts, and in some cases, in the form of grammars and glossaries. In recent history, colonial administrators and Christian missionaries documented languages for the purposes of governance and preaching respectively. Their documents were used also for scientific exploration of language by the linguists when their own recording of the language in the filed was not there or needed to be supplemented.

Documentation of languages in the context of language endangerment is new historically in terms of its purpose and scope. The purpose is preservation of language and it is two fold: to keep a record of a language that is to disappear and to aid to stop and, may be, to reverse the disappearance of a language. For the second purpose, documentation of language is a necessary, but not a sufficient, condition. This new documentation is also, like the earlier recording of languages, usable for linguistic exploration either at present or at a future date. But the scope of recording is vast, which is more than written texts. The linguistic exploration is to help reverse language disappearance by using it in education, for example, and to enhance knowledge about human beings through an understanding of the diversity of their languages.

When a language is documented, something more is made available besides the form of the language to the user of the document through the interpretation of it. It is the knowledge of the social system mediated by the language, the cultural values signified in the language and the cognitive pattern of interpretation of the social and natural environment codified in the language. The nature of the language materials documented will differ to become amenable for these different kinds of interpretation. The present day documentation of language, thus, has a larger canvass and is for multiple purposes.

There is a new kind of language documentation at present using the digital technology, which is called creation of language corpus. The corpus is being created for the major languages of the world. It is huge running into millions of words. Mostly it

consists of written texts and as such the texts are inherently edited or deliberated. A corpus may document historically earlier texts of literary compositions that give the earlier form of the language, or spoken materials, but they are not many. The goal of the large corpus is to record the actual use or manifestation of the language. Since the language use has wide variation and multiple contexts, the size of the corpus is naturally large to be representative of the wide spread use of the language. The purpose of the corpus is to provide data for analyzing naturally occurring language, making dictionaries of finer distinctions of contexts of use and meaning, producing natural sounding pedagogical materials and aiding development of tools for automated language processing for practical applications and translation non-creative texts of languages. The corpus aims to push linguistic analysis to go beyond the elicited and self-generated language data and to theorize language competence from language performance. If there is any element of preservation in language corpus, it is the recording of language change and language variation that could be discerned in the future.

Documentation of endangered languages differs from language corpus creation with regard to the status of the language selected, to the purpose of the user and to the size of the material. The grammar and the dictionary are not an integral part of the corpus, but are potential products from it and are produced for their commercial value. Language documentation, on the other hand, is for the sake of preserving the language through its texts (overwhelmingly oral), grammar and dictionary and the diversity of human cultural and intellectual heritage.

Language documentation has the advantages of the new technology. It can be done digitally in cyber space and is not constrained by considerations of storage space and perishable material. The documents can be easily duplicated and transmitted globally. New tools help quick and precise recording in the field (provided there is a source of power) and help quick and efficient organization and analysis of data. The technology makes it possible to have multimedia documentation with sound and image that give authenticity and animation to the data. The texts in documentation may remain open for later additions with ease, by the same or other linguists as well as by the native speakers themselves, when they are acquainted with the new tools. The documentation at one site can provide links to other sites of documentation of the same language. Thus the limits imposed by the earlier medium and techniques of documentation are overcome. Nevertheless, since the technology changes fast, the data recorded, for example, in a CD now may not be readable after some years or decades, like the data in punch cards are unreadable now. It means that language documentation must keep transferring the data to new medium as it is developed by improved technology.

Documentation of endangered languages in different parts of the world, as different from the field notes and publications of linguists, has been taking place since a decade. The professional linguists are involved in this new enterprise, but, unlike in the past, the language communities are taken to be their partners. The documentation work at different

places and by different people differs in the extent of the use of technology and in the professional orientation of those who do the documentation. I shall describe the opportunities and challenges for the documentation of languages in India, which has a long history of inscriptions in languages used by the royalty.

2 Response in India

If one goes by the names people give to indicate their mother tongue, there are 10,400 mother tongue names returned in the 1991 census of India. This number is systematized to be 3,372 in the census by eliminating the spurious names. Of these, 1,576 are linguistically identifiable names, but are not necessarily distinct from others linguistically. The remaining 1,796 have not been identified with any language variety and they become listless in the census being brushed aside as other names (Bhattacharya 2002). The mother tongues with identifiable names that have more than 10,000 speakers are grouped under a language (as viewed by linguists, not necessarily as perceived by their speakers). In essence, they are not given language status. The mother tongues with less than 10,000 speakers are also grouped under a language, not individually but under the amorphous category 'other mother tongues'. They do not count for anything other than adding to the number of speakers of the language under which they are grouped, or becoming invisible being not grouped under any language.

The 1,576 mother tongues, which are expressions of social identity rather than grammatical codes, are classified into 114 languages (Indo-European 20, Dravidian 17, Austro-Asiatic 14, Sino-Tibetan 62, Other 1). The question raised for language documentation by the census report on languages is whether documentation is concerned with endangerment among the abstracted 114 languages alone. The answer is obvious. By the size of the speaker population, which is one of the factors that indicate endangerment, the mother tongues not grouped under any language are the ones endangered. In other words, the languages not in the census list are endangered. These are almost as many as the listed languages[1]. One, therefore, does not go to the current census report to find endangered languages, but goes to the field. The filed linguists, for example, claim that there are 26 Dravidian languages (Krishnamurti 2003). When compared with the census figures, nine of them are endangered by the population criterion.

With regard to mother tongues grouped under one or the other language, they do not count for documentation, if the purpose of documentation is to record variant representations of human cognition as coded in the languages of the world. The mother tongues (they are not dialects, which are constructs of linguists) of a language are superficially different in their grammar. There are, however, mother tongues that differ

[1] The 1961 Census that abstracts languages from mother tongue returns without reference to the number of their speakers lists 200 languages. In the absence of information on languages with less than 10,000 speakers in the later Censuses, one will have to rely on the 1961 Census to identify endangered languages.

substantially in their grammar, but are grouped under a language, or there are mother tongues that have same grammar but are listed as separate languages, for religious or political reasons (Brass 1974). If the purpose of documentation is to record language diversity based on the community construct of language difference, they count for documentation.

The job of identifying endangered languages falls on individual linguists when a survey like the census is not helpful. The linguists in India are concerned about the loss of languages. They select for description languages that have not been described so far; but the fact that some of them are endangered is an incidental one; it is not the primary criterion for selection by linguists. Language documentation is a facet in field linguistics and not a distinct field of professional work. As such, there is no specialized training in language documentation in linguistics departments other than training in field work, which is also provided in many departments in the class room by bringing speakers to them. Any work in actual field is done as part of doctoral work by students. Such work has become less attractive to students after the emergence of Chomskyan linguistics, which makes possible the merger of the analyst and the speaker of a language into one and the emergence of sociolinguistics, which makes the study of dialects -variation within a language- the focus of field work. To top it all, linguistics lately does not attract motivated students in India. All these result in documentation of endangered languages not being a top priority in the discipline.

The institutional arrangement for describing un- and under-described languages, that includes endangered languages, rests with the government. There is no major infra-structure and support building efforts for documentation by academic bodies like University Grants Commission and Indian Council of Social Science Research or by private foundations. The Tribal Welfare departments in State governments have a wing for work on tribal languages, but the work is on teaching them to tribal students in schools and government officials posted to work in tribal areas. Any documentation work such as elementary grammar and word list is not of professional standard.

There are three institutions with the Central government, Language Division of the Census, Calcutta, Anthropological Survey of India, Calcutta and Central Institute of Indian languages (CIIL), Mysore, which do linguistic work of different kinds that have a bearing on language documentation and preservation. The Language Division is responsible for the analysis and classification of language data of the Census and it is bound by administrative and political decisions of the government about the nature of the data reported to the public. One descriptive work it does is establishing a mother tongue as a language by its grammatical distinctiveness, which process is called language identification. This identification work enables any mother tongue with less than 10,000 speakers to be called an endangered language instead of being subsumed under another language and ignored for official purposes. The Anthropological Survey, as part of its recording and study of tribal

cultures, documents tribal languages. The linguistic work is secondary in importance in terms of human and budgetary resources of the Survey.

CIIL's work on tribal and other minor languages has pedagogical orientation. The grammars, bi- or trilingual dictionaries produced by the institute are primarily to aid production of school primers, training of teachers and language learning by workers from outside in tribal communities; the texts of folk tales recorded are primarily for providing cultural content to the pedagogical materials. Nevertheless, the largest number of tribal and other minor languages described at one place in a comparable format is at the institute. These descriptions have been archived there and are expected to be available digitally. The institute initiated a project specifically for recording endangered languages in a common format in collaboration with linguistics departments in universities, but the project did not go very far due to poor response from the linguistics departments and poor phonetic and grammatical training of their students. The institute has a programme of providing fellowship to students from tribal communities to pursue linguistic study in universities with the purpose of creating a pool of native linguists who will work on their languages. After graduation, however, many of them get into administrative jobs in the government.

The training for language documentation must be multifaceted, as there is no single documentation of a language. The kind of documentation varies depending on its use. It will differ from the choice of language to aspects of the chosen language with regard to the products aimed from it. If the aim is understanding of human history and the course of migration, language isolates will get preference for documentation. If the purpose is to know the language genealogy, the documentation will go for lexical data sanitized of words absorbed from contact with other languages. To know, on the other hand, about the past or contemporary interaction between communities, the documentation will value the mixed lexical and grammatical data. Documentation of a language of attrition will not yield a functioning language when its data are sanitized. The notion of purity of language for documentation does not serve the pedagogical purpose, which is important in language transmission between generations. If the language actually used in life should be the language of pedagogical materials, particularly for a bilingual education programme, it cannot be a purified variety. This is true also of efforts to revitalize an endangered language and make it function effectively in the modern times, since it will require words for new concepts, institutions and objects, which may come from other languages. When the linguist's ideal of a pure language for grammatical study does not match with the community's goal of functional use, the documentation must find a way to satisfy both.

The purpose of constructing the grammar of a language for a linguist may not be just discovering it for its own sake. Such a construction of grammar may not even stop with building the knowledge base of the professional linguists and their discipline. The grammar could be the means to understand the human mind, which manifests through the formal structuring of human languages, which differ from one another in many, but predictable, ways. The grammatical parameters of languages are as important as their

universal properties to understand the human mind. For this endeavour, theoretically informed in-depth grammars are necessary in as many languages as possible. The endangered languages that have grammatical structures that are unusual and different from the known languages get priority for documentation to be used for this purpose. It will be a loss to human knowledge if such endangered languages are lost without the study of mind being informed by them. The contribution of Warlpiri to linguistic theory regarding non-configurational nature of grammatical structure is a reminder to such a possibility (Hale 1998). This incidentally makes the point that there is no disjunction with theoretical linguistics when one of the purposes of documenting endangered languages is to have their contribution for the study of human cognition. The phonology of endangered languages may help to define the extent of the subset of meaning differentiating linguistic sounds out of the large number of sounds that humans are physiologically capable of producing. Without the knowledge of the endangered (Khoisan) languages with clicks in Africa, for example, the subset will have remained unreal.

Similarly, for appreciating human linguistic legacy, the documentation should incorporate typological orientation in the collection and organization of data to tell us how the grammatical diversity is not random, but is structured. This purpose suggests preference in selection for documentation for languages which have diverse grammatical structures. They will show how rich and plural the human linguistic legacy is, as is the artistic legacy.

One important value claimed for language diversity is that it allows for the diversity of experiencing the world and interpreting it through the medium of many languages and for helping to live a life in harmony with self and others by learning from the understanding of life by the users of other languages and cultures. To have this gain, the lexical documentation must be semantically detailed with information on the semantic range of words, their semantic classification and their metaphorical use in ordinary discourse.

The dictionaries that form part of documentation are commonly bilingual with the target language being an international language like English, which helps the linguists in their research. They, however, must be tri- and quadri-lingual with gloss in the regional (provincial) and national languages, if they are to help the speakers of endangered languages to improve themselves materially and to empower themselves politically. Addition of these languages in the dictionary makes them tools for learning these languages of power. This purpose of the dictionary for the community necessitates the entry of the endangered language in the dictionary using the alphabet developed for it. The writing system, grammar and dictionary will deny the common assertion of the dominant community that the speech of small communities without these is not a language and so is not fit to be used in public domains like schools. The denial of this assertion with such materials gives the community a sense of pride and self-worth, which is the first step towards empowerment. Dealing with the low socio-economic status of the endangered

language is a subsequent step to giving it the language status. To get the semantic information mentioned earlier, the dictionary must give gloss in the endangered language itself, as the dictionaries of major languages do, with detailed description of polysemy and contexts of use illustrated with citations. The bilingual dictionaries tend to give translation equivalents in the dominant target language for the words of the endangered language rather than giving a semantic description of the words within the framework of the semantic structure of the endangered source language.

A language constitutes the social system in which it operates. The knowledge of how the language is controlled and used gives the knowledge about the social system. This knowledge could be obtained from information about language use and denial of use in actual situations. Natural conversations in different settings with different actors will give an idea of who speaks what to whom when and how from the cues like turn taking, hesitations, hedges, jokes, politeness expressions etc. To get these and other cues from language use to get an idea of differences in the social systems of communities speaking endangered languages, there must be documentation of ethnographically rich speech with detailed annotation of the contexts of use including the backgrounds of the participants in the speech event. This goes beyond elicited narratives.

All the above show that it is an extremely challenging task to arrive at the materials to be documented that are maximally usable for multiple purposes. Documentation for the use of both the linguists and the community is itself difficult. The problem of dialect versus language is a basic problem about selection for documentation. This distinction is a linguistic construct, which may conflict with the community's perception about their speech, as mentioned earlier. For them, others speak differently and they have no use in categorizing and labeling that difference. Another problem lies in the choice of the meta-language for annotation and description. The community has no meta-language different from their own language and the linguist is incapable of using the target language as the meta-language. Moreover, collection of different kinds of materials requires different techniques and expertise. The linguistics departments in India are not geared to provide training of such magnitude and complexity to take on language documentation in a serious way.

It is unlikely that funding and other support will be provided by the government and other bodies to linguistics departments in the country to train students in language documentation and take up that work on priority because there is no public awareness or demand for language preservation even to the little extent it is found with regard to preservation of endangered animals and plants. This is particularly true of people speaking major languages in India. The reasons for the apathy are many. First, there is a view that the loss of native language is the price inevitably to be paid for material progress. Second, there is a belief that the native culture could be practised and preserved even when the native language is replaced. Third, there is a hope that giving up small languages will enhance wider communication between social groups. There is a need to change these

erroneous public perceptions to gain popular endorsement and support to preserving endangered languages.

Linguists (such as Pandit 1979) have observed that societal bilingualism in India is stable and people build a repertoire of languages for themselves, rather than replacing one language by another. This historical fact makes people be complacent about language loss, particularly in tribal communities, whose mother tongue retention rate has fallen from 49% in 1971 to 42% in 1981 (Khubchandani 2002). Besides such shifts to the dominant languages of the region, there is also biological extinction of tribal communities like Andamanese (Annamalai and Gnanasundaram 2001). The question why the migrant communities in India maintain their language over generations, which Pandit posed, must be coupled with the other question why stable communities lose their language when participating in an industrial economy and centripetal politics of linguistic states in India. Raising the second question loudly will help create awareness about the developmental perils to multiplicity of languages and gain support to the traditional multilingual ethos.

Whether the society succeeds in arresting language loss or not, documentation of languages likely to be lost in India will contribute to knowledge about human capability and history, like the languages to be lost in any other country. The endangered languages in India may have unique grammatical and semantic features; the isolates among them, as found in Andaman islands and unitary members of language families, as found in the north-eastern hills, may have unique stories to tell about human migration. Given the transference of grammatical features between languages of different genetic families due to stable bilingualism as well as language shift that makes the grammar an areal phenomenon, the grammatical data of endangered languages, along with the data of other languages, will give knowledge to linguists about the process of grammatical convergence. This may perhaps be a special benefit of documentation of endangered languages in India.

References

Annamalai, E. and Gnanasundaram, V. 2001. Andamanese: Biological challenge for language reversal. In Fishman, Joshua A. (ed.) *Can threatened languages be saved?* Clevedon. Multilingual Matters.

Bhattacharya, S.S. 2002. Languages in India: Their status and function. In Itagi, N.H. and Singh, Shailendra Kumar (eds.). *Linguistic landscaping in India with particular reference to the new states*. Mysore: Central Institute of Indian Languages

Brass, Paul. 1974. *Language, religion and politics in north India.* Cambridge: Cambridge University Press

Hale, Kenneth. 1998. On endangered languages and the importance of linguistic diversity. In Grenoble, Lenore, A. and Whaley, Lindsay J. (eds.). *Endangered languages: Language loss and community response*. Cambridge: Cambridge University Press

Krishnamurti, Bh. 2003. *Dravidian languages*. Cambridge: Cambridge University Press

Khubchandani, Lachman, M. 2002. Language profiles of Jharkhand, Chattisgarh and Uttaranchal: A subaltern perspective of language development. In Itagi, N.H. and Singh, Shailendra Kumar (eds.). *Linguistic landscaping in India with particular reference to the new states*. Mysore: Central Institute of Indian Languages

Pandit, P.B. 1979. Perspectives on sociolinguistics in India. In McCormick, W.C. and Wurm, S.A. (eds.). *Language and society: Anthropological issues*. The Hague: Mouton