John Benjamins Publishing Company

Anju Saxena and Lars Borin (eds). 2006. *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology*. Berlin: Mouton de Gruyter [Trends in Linguistics. Studies and Monographs 175]. viii+386 pp. (ISBN 978 3 11 018976 6)

Reviewed by Robert D. Eaton (University of Texas Arlington and SIL)

This recent monograph in the Trends in Linguistics Series (175), edited by Anju Saxena and Lars Borin, is a compendium of 17 articles that grew out of a desire to explore ways in which modern information technologies could be used in support of endangered languages, particularly in South Asia. The book is divided into three main sections: (1) language policy, (2) case studies of lesser-known languages, and (3) information and communication technology (ICT) use.

In the introduction, Saxena highlights the fact that, due to forces of internationalism and globalization, many of the smaller language communities in the world are being "embattled" by politically and economically dominant communities and their languages. One estimate (Krauss 1996) suggests that half of the world's 6000 languages will die in this century if something is not done to stop the trend. With the recent discussion in linguistics on documenting endangered languages, this book provides a good summary of the situation as it stands in South Asia as well as suggestions for the way ahead, i.e. the use of information technology to both document and revive lesser-known language varieties. Several of the articles also discuss the pitfalls encountered in the process of documenting languages, so this book is especially encouraging for those of us involved in the documentation task who desire to learn from the experiences of others.

Language Policy

In the section on language policy, there are three articles on how the social and political situations in India, Nepal, and Pakistan either support or degrade the viability of lesser-known languages. In India, Udaya Narayana Singh points out that while the government has created different language policies and constitutional provisions towards promoting and protecting linguistic diversity, these positive steps are inhibited by various complicating factors which have resulted in less progress than could be hoped for. The inhibiting factors, interestingly, are mostly social rather than political and include population, economics, and educational issues.

Studies in Language 32:1 (2008), **245–253**. DOI 10.1075/sl.32.1.15eat
ISSN 0378–4177 / E-ISSN 1569–9978 © John Benjamins Publishing Company

The case study by Bettina Zeisler on why Ladakhi ought not to be written is a good case in point. She discusses the phenomenon common across the Tibeto-Burman world in which Classical Tibetan — closely associated with the Tibetan script — has almost no similarity with the modern "common languages" currently in use. Certain reform-minded people and organizations would like to revise the Tibetan script and begin using it to write the different modern vernacular languages. They are concerned with the current situation in which only a fraction of the Buddhist monastic elite truly understand Classical Tibetan. The common people are taught the Tibetan script, but since it is for a language they neither know nor use, it is not effectively internalized.

From the Buddhist scholars' perspective, they see the attempt to write the vernacular languages with the Tibetan script to be anti-Buddhist and lacking in traditional scholarship. Since the Tibetan script has been around since the 7th century and was used primarily for the codification of the sacred Buddhist scriptures, the "script and classical orthography have become sacrosanct and should not be altered even for lay purposes." (p. 178). To do so, it is thought, would result in the disintegration of pan-Buddhist identity and is even seen as treachery to the Tibetan cause.

Ultimately, Zeisler's conclusion is that there is little possibility of convincing the established Buddhist scholars who simply do not want to be convinced and that the only hope for development of Ladakhi as a written language lies in the younger generation developing the interest and necessary freedom of thought.

In Nepal, Mark Turin shows the recent shift that has taken place in language policy over the last several decades. During the Panchayat rule (prior to the restoration of democracy in 1990) — when the focus was primarily on unity rather than diversity — the state promoted the idea of "One nation, one culture, and one language" (Nepali). Specifically attempted to eradicate the use of "local dialects and tongues" from public life. Since then, however, the government has made some progress towards recognizing the diverse ethnic and linguistic situation in Nepal even going so far as to suggest that "All the languages spoken as the mother language in the various parts of Nepal are the national languages of Nepal" (from the Constitution of Nepal).

Similar to the situation in India, however, the case study by Michael Noonan argues that the rise of **ethnic consciousness** has not necessarily led to social or political action taken on behalf of preserving languages. Since language has such a central role in the ideology that underlies the rise of ethnic consciousness, this result is somewhat unexpected. After all, a proper ethnic group should have its own language and the group should have rights with regard to that language.

Noonan argues that what accounts for this lack of social or political action to preserve the language is the lack of a clear connection in peoples' minds between the official status of their mother tongue and economic and other forces affecting them. All the evidence suggests that the use and promotion of indigenous languages does not rank highly on peoples' list of concerns; mostly because the case has not been successfully made that the use of Nepali is causing a serious disadvantage for these ethnic groups. As a result, language issues have not made it to the forefront of the political discussion.

Noonan's conclusion is that the only way these indigenous languages will ever be written on a regular basis (i.e. besides the documentation that is being done mostly by foreign linguists) is if the languages are used as the medium of instruction in schools. Educational reforms promoting the use of minority languages in the early years are the best way to encourage literacy in these languages, which in turn promotes their use in writing. Interestingly, as this section was being written, I received notice of a project proposal in which the Nepal government requested Finland's Ministry for Foreign Affairs to provide technical assistance for a program of Bilingual Education for all non-Nepali speakers. The proposal included providing education in the mother tongue for communities up to the primary level. If this proposal is accepted, that should bode well for Nepal's lesser-known languages.

By contrast, Tariq Rahman paints a fairly grim picture of the future of lesser-known languages in Pakistan. He discusses the lack of governmental policies that encourage the protection of lesser-known languages. Instead, the policies that do exist give privilege only to Urdu (the national language, but the mother tongue of only 7.57% of the population) and English. Similar to the earlier Nepal situation, the rationale given for this decision was that Urdu is a symbol of unity helping to solidify Pakistani identity and is very widely known, especially in urban areas. It also prevents any of the regional variants from attaining higher status, which could cause division in the federation.

One result of this favored status for Urdu has been resistance to it by the different ethnic groups particularly because of the perception that it has been promoted, often in insensitive ways, by the ruling political elite. Nevertheless, though there were several language riots in the early '70s, the population remained pragmatic: they knew that without learning Urdu, they would never be able to advance economically. The result was the loss of importance of their own languages. So much so, Rahman argues, that today many see the idea of reversing language shift as "fatuous or sentimental nonsense" (p. 83). The indigenous languages are seen as markers of backwardness or symbols of ethnic resistance to the center and are therefore not taken seriously. Rahman states that very few people see this as a problem and as a result, there are no policies about preserving the linguistic diversity of the country.

In fact, similar to the way in which Urdu has risen in status above the other indigenous languages, English has managed to climb to the top of the preferred language hierarchy in Pakistan. Another email I received this week announced that as of November 2007, English will replace Urdu as the language in which science and mathematics is to be taught in all state schools in Pakistan. This is a reversal of an earlier policy introduced under the military regime of General Zia ul-Haq (1977–88), whereby English was not to be formally taught in state schools until a child reaches the age of 10.

One contrasting perspective is given in the case study by Kohistani and Schmidt on the Shina language spoken in the Karakoram and western Himalayan Mountains in Pakistan. The authors' conclusions about Shina are that, in rural areas, due to the rugged and inaccessible terrain, languages other than Shina have not made significant inroads and there is little chance of language death in the near future — especially in areas where modern education has only recently become available and among women who are mostly uneducated and monolingual.

Information and communication technology

On the use of Information and communication technology (ICT) to benefit the situation of lesser-known languages, the book can be divided roughly into three main topics: developing resources for language research and documentation, language standardization, and language promotion.

The article by the co-editor, Lars Borin, entitled, "The promise of language technology," sets the stage for the discussion about ICT. He starts out by defining several terms used in the field of language technology and discusses how he envisions such tools could be used to come to the aid of lesser-known languages.

He goes on to discuss some of the problems facing those who hope to use ICT for the benefit of lesser-known languages. For example, most of the presently available resources were written with English in mind, where minimal morphological complexity, relatively fixed word order, etc., is taken for granted. However, this is not the rule among most of lesser-known languages in South Asia. Much of the work that has been done to improve the accuracy of these tools for English will likely need to be redone when applied to these significantly different languages.

This reality was demonstrated in the article by Hardie, et al., describing their efforts in developing a spoken and written text corpus for several South Asian languages. For example, one of the problems they encountered in collecting texts was the plethora of unique encodings used by the different sources. Because of this, it was decided to re-encode all of the texts in Unicode. Unfortunately, not all of the software they were using for data management was Unicode-compliant. So they first had to extend the software to support it. Consequently, it was not only the research tools that needed to be developed for working in these languages, but even the framework for data management had to be modified.

On topic of documentation, Jens Allwood's article on "Language survival kits" focuses on the issue of **levels of survival** for endangered languages, with the two interesting categories being: (1) for-the-record (i.e. museum) survival and (2) maintenance of functional viability. The former can be done by collecting written, spoken, and gesture data of lesser-known languages, categorizing it and storing it in such a way as to be available for future study.

The general guidelines he discusses for the survival kits are: (1) they should take into consideration both written and non-written languages (so they should include audio/video rather than just text), (2) they ought to be free or low cost (so a preference for Open Source resources, such as Linux), and (3) they should be based on standards (e.g. Unicode) to reduce the complexity in accessing the data. In addition to encoding and storage considerations, there was an appeal for consistency in annotation, in order to facilitate, as much as possible, "automatic computer-based analysis". The article was a bit thin on the details for the kits other than the concept itself. For example, a review of markup standards for data storage would have been useful, but otherwise, this article is a good first step.

The article by Boyd Michailovsky, on the other hand, provides good details of the electronic resources available to researchers of Nepalese languages. He discusses the different text corpora, sound recording-archives, online journals, and lexical resources available. For each resource, he gives an overview of the kind of data available, the storage format and data encoding used, software needed for accessing it, as well as internet addresses where these resources can be found.

He also gives a detailed overview of the Lacito Archive (Michailovsky 2006), which has an impressive storage and retrieval mechanism that is clearly using best practices of the current state of the art for digital resources. The data are all encoded in Unicode (for maximum future mobility), and stored in eXtensibile Markup Language (XML) documents in logical format. This means that anyone can create their own XSLT scripts for transforming the data into whatever "view" is useful to them. The Lacito Archive website has a number of useful XSLT transformations through which users can view the archived data in any of the following ways: straight transcription, interlinear glossed text, morpheme lists, lists of utterances containing a particular morpheme, and a concordance.

On the topic of language standardization, one of the important concerns raised by several articles is the issue of orthography standardization. The case study by David Bradley discusses the historical development of the different orthographies created for the Lisu language of China, Burma, Thailand, and India. It provides a good overview of the difficulties involved in developing a new/unique orthography for a language in terms of publication facilities, computing issues, promotion, etc. Regarding the "Fraser script" — the most widely used of the different writing systems — he points out that it was created with an *Einbau* rather than

Ausbau approach.[1] As a result, it does not represent any particular dialect of the language. Because of the large dialect chains in South Asia — where "the language changes every 12 kilometers" — this may, in fact, be the single most important factor facing those involved in standardization. One has to decide which variety to document: a specific one among many or a non-existent conglomeration of them all? Noonan's article highlights this as one of the obstacles to language development in Nepal. The process of standardization necessarily involves some decisions which may not be acceptable to everyone. Studies in identity politics have shown that people would often rather learn a national or international language before adjusting their variety to conform to that of a closely related neighbor.

In this age, clearly an important concern related to orthography standardization is operating system support for the writing system. On this topic, the article by Vasu Renganathan and Harold Schiffman discusses how script support has benefited the Tamil language. They note several developing technologies for representing Tamil on computers, including advances in printing, the Unicode encoding standard, smart shaping fonts and increased support for software localization. These advances have greatly facilitated the dissemination of the Tamil language on the internet (e.g. magazines, newspapers, Tamil-based software, etc).

One difficulty noted was that even once Tamil could be adequately displayed on a computer, another hurdle was the inability of people to type Tamil words since the keyboard sequences for certain letters were non-obvious. This resulted in a kind of "keyboard illiteracy" that had to be remedied.

Finally, whereas most of the articles in this volume are concerned with the documentation of endangered languages for (one might say) "posterity's sake" or for use by researchers in developing typological studies, the article by David Nathan and Éva Á Csató discusses a very different target audience: the language communities themselves. That is, the delivery of usable resources to the community — based on their perceived needs — that will help them maintain their language and culture before the language shift becomes too great.

Their suggested approach is to use multimedia technology to record cultural and linguistic events in context. These raw resources are then used to create a synchronic, multimedia encapsulation of the language and culture of the community: recordings of language events, linguistic descriptions and analyses, descriptions of the community's history, religion, literature, social structure, food, secular and religious music. As these resources are then delivered to and used by the community, they inherently contribute to language maintenance.

Because of certain email habits of the people among whom they work, which worked against language maintenance, the authors do not regard the internet as a good option for delivering the media. However, given the situation in South Asia of fewer computers per capita, the idea should not be dismissed too quickly. If the media were provided via the internet, people could access it from the now ubiquitous cyber cafes. In addition, internet practices like blogging and resources like Wikipedia, have been suggested as useful repositories for cultural and other linguistic information. This would allow anyone in the community with access to the internet, sufficient keyboard literacy, and motivation to contribute language content themselves. By including the community in the development of the resources, it would enhance their "ownership" of the content as well as enhance the prestige of their language and culture and, in the process, actually **maintain** their language.

Conclusion

Jens Allwood's article presents the arguments both for and against maintaining linguistic diversity. The arguments against are understandably based on various pragmatic economic concerns: hindrances to global trade, the waste of resources involved in communications, and a perception of the futility of the goal of language diversity preservation. The arguments for linguistic diversity are understandably more human: recognizing language as human beings' greatest collective cognitive achievement, though which we have been able to advance as a species; recognition that multilingualism increases mental capacity, flexibility, and creativity; and the insight that is provided into human nature itself by language. The grammar sketch of Great Andamanese by Anvita Abbi gives a good illustration of why language documentation and promotion is so vital. Those language groups are thought to be the last remnants of pre-Neolithic Southeast Asia and biological studies have suggested that they are descendents of the early Paleolithic colonizers of Southeast Asia. And sadly, their language and civilization are highly endangered.

However, even assuming that maintaining language diversity is a worthy goal, the solution remains unclear. Should it come from government policy or from the grass roots? In Pakistan, the language policy situation seems to be overall negative for lesser-known languages. Yet, the article on Shina suggested that language shift and death is not very imminent. By contrast, the language policy situations in both Nepal and India seem almost helpful by comparison. Yet social factors were foremost in working against language maintenance. Saxena hits on a key reason in the introduction when she quotes (from Winter 1993:311) the paradox of several people actively involved in the promotion of lesser-known languages, who nevertheless spoke only English to their children at home. Winter comments:

> What is to be observed in both cases is a conflict between wanting to do something for the language and wanting to improve the chances of the children to succeed in the macrosociety of which they're apart… does one have a right to blame the parents?

So if, as the majority of the articles suggest, such social factors are key, then the question is, can language technology help provide a solution? And to that, I think the book has convinced me the answer is "yes". The article on Tamil certainly takes the view that advances in technology and good operating system support have really benefited Tamil use overall. To a growing number of people in South Asia, "learning computers" is as useful for "succeeding in the macrosociety" as learning English. So if computers can be used to engage people in maintaining their language, then there is a chance for success.

Lars Borin began his article by quoting Nicholas Ostler's update to the old aphorism that "a language is a dialect with an army and a navy" with this new ICT age definition: "a language is a dialect with a dictionary, grammar, parser, and multi-million word corpus of text". If this is true, then there is hope for the lesser-known language communities as long as there are people willing to do the fieldwork necessary for such resources to become available.

Herein lies the biggest barrier I see to ICT being applied to lesser-known languages: the complexity involved. The article by Trond Trosterud is a good case in point. He encourages the development of morphological transducers and disambiguators — precursors to useful applications like spelling checkers, intelligent content searching, machine translation, etc. However, he mentions that his work in developing transducer technology for the first language he worked in took 2.5 years! Having worked with language technology tools in South Asia for 7 years myself, and having tried to promote them among colleagues, I've discovered that one has to be a fairly sophisticated computer user to work with the software, leave alone the data. This leaves out nearly all the members of the speech communities of lesser-known languages and that means that the work must be done mostly by outside researchers (and it takes a fairly well informed researcher to handle the tools too!) Whether or not this is possible for all of Krauss' (1996) 3000 endangered languages this century is not clear to me.

If it is to be possible, then more research is needed in ways to simplify the process. As Borin suggests, technologies are needed that will allow for **bootstrapping** ICT resources, such as, systems that can learn to categorize data automatically (e.g. learning inflectional morphology directly from an unannotated corpus). Studies are needed to determine what kind of resources the language communities themselves are interested in. Then templates for such applications are needed which can be filled in to provide language-specific implementations to make the development of such resources less cumbersome. ICT tools need easier-to-use interfaces that do not require an advanced degree in computer science; keeping in mind the goal of making these applications usable by the language communities themselves, so that community involvement can be leveraged.

In terms of orthography standardization, proposals to the Unicode consortium are needed urgently while the standard is still relatively “malleable”; we certainly do not have 100 years to make changes to it. For very small languages with unique orthographies, it may not be possible to have them added — especially if it does not happen soon. And any language that is not represented in the standard will have difficulty being represented on computers in 10–20 years.

Clearly, it is an uphill battle to stem the tide of language shift and death. If language preservation is a battle worth fighting, let us take the advice of Michael Noonan and see about encouraging policies that promote multilingual education programs. Let us heed the call of Bettina Zeisler that it is the next generation that must be engaged if there is to be a new openness for preserving ties to language and culture. Let us pursue the strategy of David Nathan and Éva Csató in developing materials that are not simply “for posterity’s sake”, but which can engage and empower the communities themselves to find reasons and resources for maintaining their language and culture.

And above all, let us speak our mother tongue to our children!

Note

[1] Einbau is the creation of a new dialect based on elements of various existing dialects. Ausbau (the more typical approach for standardizing an orthography) is the promotion and codification of a single existing dialect for use with several closely related dialects.

References

Krauss, Michael. 1996. Status of native American language endangerment. *Stabilizing indigenousllanguages*, Gina Cantoni (ed.), 16–21. Flagstaff AZ: Northern Arizona University.

Michailovsky, Boyd. 2006. Lacito archive. http://lacito.vjf.cnrs.fr/archivage.

Winter, Werner. 1993. Some conditions for the survival of small languages. *Language conflict and language planning,* Ernst Håkon Jahr (ed.), 299–314. Berlin: Mouton de Gruyter.

Reviewer’s address

Robert D. Eaton
c/o Sri. K. K. Jamwal
Tikka Aima, near Radha-Krishna mandir
Palampur, Distr. Kangra
Himachal Pradesh
India 176 061