Large-scale language documentation in Nepal

A strategy based on SayMore and BOLD

Mari-Sisko Khadgi
SIL International
ICLDC-5, 4 March 2017
Background

• 1998 Himmelmann: distinction between descriptive and documentary linguistics
• Transcription of audio recordings is a real bottleneck
• Woodbury 2003:45: ‘respeaking’ of recorded materials and oral translations
• Simons 2008: BOLD (Basic Oral Language Documentation)
• SIL 2011: SayMore software
• SayMore's slogan is "language documentation productivity"
• So how productive can we get with the BOLD method and with SayMore software?
• a case study on how 4 nationals, without any previous linguistic training, have been taught to use this easy-to-learn software and the BOLD method very successfully in Nepal
So far achieved:

• recordings with oral transcriptions and oral translations in 14 languages of Nepal (two having less than ten speakers)

• 9 languages: more than 10-17 hours of recording with 70-100 speakers.

• Depending on the number of technicians and mother-tongue facilitators dedicated to a project, it takes approximately 7-10 weeks to complete one project.
How does it all work?

• 4 technicians (nationals)
• 2-4 mother tongue facilitators (or "local guides")
• Original recordings done in the language area
• The “respeaking” (oral transcription and oral translation) done in the capital city
How were the technicians selected?

• SIL working in partnership with Mother Tongue Centre Nepal (MTCN)

• Selection criteria:
  – young
  – willing to travel a lot and to remote locations
  – from an ethnic background
  – computer knowledge
  – interest to work in technical field (audio&video-recording)
How were the technicians trained?

• Why we want to do this
  – how you should introduce yourselves, how to represent well

• Informed consent

• Recording:
  – what is high-quality recordings
  – techniques in village setting (how to avoid background noise)
  – taking care of equipment

• Metadata, labeling collected data consistently
How were the technicians trained?

• What do we want to collect
  – Different genres
  – Different kinds of people (young–old, men–women, educated–uneducated, from different dialect areas)
  – Approx 10 hours of recordings per language

• Fairly little teaching needed on the actual BOLD & SayMore!
Plan and timing for a typical project

• language community approaches us and a plan for language documentation is made

• they select 2-4 mother tongue facilitators
  – time commitment for 2 months (full time),
  – willing to travel,
  – computer literate,
  – know their language well,
  – good relations in the community

• 2 technicians and 3 mother-tongue facilitators dedicated to a project, it takes approximately 7-10 weeks to complete one project
Plan and timing for a typical project

• 1-2 days training for mother tongue facilitators
• 5-7 days preparation
• 10-15 recording days in the language area
• 7 days data preparation
• 15-20 days data processing
• Distribution of recordings and photos to the language area (very important!!!)
• Archive
• (Create a website)
Size and scope of collected corpora

• Goal of 10 hours of original recordings per language (connected discourse)
• As many people as possible (70-100)
• Genres (handout explaining what each genre means, translated into Nepali)
• No word lists, semantic sets (numbers, colors, living things)
• Elicitation of basic sentences with different grammatical structures should probably be included.
Key factors of the success

• Taking enough time to train the national technicians!!
• Motivated language communities, motivated MT facilitators
• Going “local”:
  – All recordings done in the village settings,
  – by outside technicians who are Nepalis (not foreigners),
  – but with mother tongue facilitators also present
• Giving back to the community
Challenges

• Quality control, esp. at the beginning needs to be very high (recording levels), training Nepali technicians in attention to detail is a challenge

• Trained technicians may leave

• Genres difficult to define

• To publish some of the recordings in written format, a significant amount of work is required.
Further challenges

• Enough computer literate people in the MT group?
• No word-by-word translations, so before linguistic analysis can be started, major work is still required
• For an outside linguist to use the data is difficult, if they don’t understand Nepali
• Different archives getting more restricted, starting to charge (it is expensive to host the wave files and video)
• Funding...
Challenges with almost extinct languages

• Moribund languages: hard to get stories (only words and single sentences)
• Not any easier to get funding for almost extinct than to larger language groups
• How to respond when people are asking us to help with language revitalization
References


