A Speech Corpus of Dharamsala Tibetan

Introducing the Project

In 2016-2017, a 10-person team worked for 3 months with the goal of creating a multi-use, balanced corpus, similar to the Brown Corpus (BROWN Corpus search online). The speech section of the completed Nanhai Corpus—named for its sponsors, the Nanhai Nunnery of Taiwan—is a 289,497 word corpus of collected, transcribed, and word-split natural speech of local Dharamsala Tibetan (རྡོ་རྗེ་སྟོན་སེམས་)

Data Collection, Transcription, & Segmentation

There were three types of speech collected: 1) Dialogs (prepared speech); 2) Studio Recordings (prompted speech); and 3) Natural speech. For the final category, participants from the community wore lapel mics as they went about their day. The data was saved in MP3 format. These were cut into manageable file sizes by conversation and conversation type (link: Nanhai Corpus - Codes for Topics & Types).

The transcription process tracked common spelling errors, new spellings for speech words (link: Nanhai word list). Segmentation was done first by an early version of Esukhia’s PyTib (a pre-cursor to PyBo), a rules-based tokenizer programmed in Python. The splitting was then proofread and corrected by the team, while tracking the corrections (link: Nanhai word list).
Outputs

The Nanhai Corpus

The Nanhai Corpus is a 1.2 million word corpus of modern speech and writing, freely available online at Esukhia’s Github corpus repository (https://github.com/Esukhia/Corpora). The speech section accounts for approximately ⅓ of the corpus; the remainder is made of modern news, literature, stories, blogs, and other web content. Users may download the text files and analyze for their own use using AntConc or a similar corpus analysis tool.

Frequency Lists

A series of frequency lists were compiled and analyzed (link: Frequency Lists -- Natural Speech) by Esukhia's R&D team in 2017-2018 during the planning for the new A0 beginning reader for students learning Tibetan (https://esukhia.xyz/a0-cover). Frequency from the corpus data informed decisions such as 1) which letters to introduce in which order; 2) which vocab to focus on first; and 3) other frequent phrases and sentences that are foundational to speaking Tibetan.

Text Editor

These frequency lists are also pre-loaded in the beta Text Editor & NLP tool suite. Linguists, authors, and material developers may use the tool to analyze texts directly in an editor, Word-like editor (https://github.com/Esukhia/dakje). Users can highlight text, by word and frequency, to determine the level of the vocabulary; view the Tibetan text with or without word spacing; and view and edit POS tagging.
Who?

- **Organization**: Esukhia – esukhia.org
  - Founded 2011
  - Based in Dharamsala, HP, India
  - Immersion & Online Tibetan language education; Research & Development for language & education; Projects for translation, resources, etc.
- **Staff**: 10-person team of native speakers, led by
- **Sponsor**: Nanhai Nunnery, Taiwan

Who?

To briefly introduce the organization, Esukhia was founded in 2011. The school is based in Northern India, in the foothills of the Himalayas, and offers immersion-based, Tibetan-language education.

In concert with that work, we have also been involved in research and development work for adapting language pedagogy, translation, and materials for anyone who works with the Tibetan languages.

The bulk of the work for the Speech Corpus itself was accomplished by a 10-person team of native speakers, sponsored by the Nanhai Nunnery of Taiwan, with the goal of improving textbook materials.
<table>
<thead>
<tr>
<th>What?</th>
<th>Where? When?</th>
</tr>
</thead>
<tbody>
<tr>
<td>What we call the &quot;Nanhai Corpus&quot;, then, aimed to be a balanced, general-purpose corpus (similar to the famous Brown Corpus, one of the first of its kind, compiled in the 1960s). It contains 1.2 million words (it is word-spaced) representing both modern speech and writing. The speech section is made of sections (representing 4 major regions, Lhasa, Central, Amdo, and Khams). There is also a section of Children's Literature.</td>
<td>Meanwhile, the speech section was recorded in Dharamsala, India in 2016 and 2017.</td>
</tr>
</tbody>
</table>
Why?

- Tibetan language education
  - Frequency lists for grading curriculum (L2) & stories (L1)
- Reference / Resource
  - ala the Brown Corpus: https://www.sketchengine.eu/brown-corpus/

Why?

A modern, balanced, general purpose corpus — especially with an aim for creating language-learning resources for Tibetan language students (like https://esukhia.xyz/a0-cover/).

By creating material based on speech corpora, we can begin bridging the gap between “how Tibetan is spoken” and “how Tibetan is written”.

With extensive reading in speech-like writing, beginning readers (who can, or are learning to, speak Tibetan) can begin making connections between sounds, symbols, and meaning, supporting them on the track to sophisticated literacy.

How?

- Speakers wore a clip-on mic during the day
- Recordings were collected
- Transcribed and edited
  - This includes word-splitting editing
  - ...and standardizing spelling for colloquialisms

How?

To capture natural speech, speakers wore a clip-on mic during the day; the recordings were collected, transcribed, and “edited” — not for language use, but for word-splitting (putting spaces between words) and for standardizing spelling of colloquialisms (spellings for words that aren’t usually written down).