**Building the Sinhala Conversational Corpus**

Spontaneous dialogue speech corpora are essentially important to model relevant features of spontaneous speech, such as pauses, hesitations, turn-taking behaviors etc. The following are widely used publicly available corpora which has have been used in most of the DA modeling.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Corpus** | **Utterance Count** | **Word Count** | **Distinct Words** | **Dialogue type** |
| SWITCHBOARD | 223 606 | 1 431 725 | 21 715 | Conversational |
| VERBMOBIL | 3 117 | 24 980 | 959 | Task-oriented |
| ICSI MEETING RECORDER |  |  |  | Conversational |
| MAPTASK | 26 621 | 152 705 | 2 502 | Task-oriented |

But, none of these corpora is in Sinhala and neither translation exists. This was a major challenge in carrying out our project. So, we tend to build a standard corpus from the scratch. We have tried out following different approaches.

* Translate an existing standard English corpus
* Sinhala chat tool
* Sinhala movie subtitles

Finding translators was not possible so we abandoned the first option. Then we deployed a Sinhala chat tool for public use and collected conversations. At the beginning this approach seemed promising but the process was slow because it was difficult to get volunteers and the volunteers were tend use to use English words in the middle of Sinhala utterances. Also they used slangs and urban words more often which makes the classification more complex. Although we understand that a dialog act recognition system should accept the existence of such non-standard words, this was considered out of scope for the current research.

Then we tried to extract utterances from Sinhala subtitles of English movies. The translation of English movies is a result of a community-based crowed sourcing effort. About 10 full-time translators are contributing to this under the trade name of “baiscopelk”[[1]](#footnote-1). In Sri Lanka, there is a large population that enjoys Hollywood movies and TV series. However, their low English literacy is a problem when understanding these movies and TV series. The aim of baiscope.lk is to provide Sinhala subtitles. The subtitle creation process is governed by a set of rules and regulations. The subtitles are almost in grammatically correct Sinhala.

One issue with this method is that some movies have frequent scene changes. This is problematic for extracting consistent conversations. To overcome this we had to manually select the movies that contained long consistent scenes. We collected about 1.8 million utterances using this method for 2306 movies.

Extraction and segmentation of utterances were done manually to build a more conversation-oriented corpus. Extracting the utterances from a subtitle file consist of several steps. First step is to omit the time-related information mentioned alongside utterances. Then the filtering out of advertisements and symbolic characters takes place. Finally eliminate the improperly used punctuation marks. Such as using multiple exclamation/question marks instead of using one right after an utterance in order to emphasize the emotion conveyed in the movie scene. Segmentation is done manually by checking each line for one statement broken into few lines in the subtitles. It is a result of a scene change in the middle of an utterance in the movie. If any such lines found, can combine them into a single line.

The final corpus contains 1.8 million utterances including tagged 12,000 utterances. It is publicly available[[2]](#footnote-2) under the name of “Sanwada”.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)