HP Labs Tamil Online Handwritten Word Data Set

hpl-tamil-iso-word

This document describes the details of the Tamil Online Handwritten Word data Set. The data set has been created by asking users to write Tamil words listed in the file ./Unicode_Words_Used_For_Data_Collection.html.

Data Collection: Data was collected from 131 users. The total words listed in ./Unicode_Words_Used_For_Data_Collection.html were divided into three non overlapping sets and a subset of users was asked to write the words. Each user contributed two trials for each word in the subset. The details of the subset of words and the subset of users are available in the following text files located in the same directory as this README.

./part1_symbol_lexicon_for_data_collection.txt

./part1-writers.txt

./part2_symbol_lexicon_for_data_collection.txt

./part2-writers.txt

./part3_symbol_lexicon_for_data_collection.txt

./part3-writers.txt

Symbol Ids: Each word is represented as sequence of symbol ids. The map of symbols and their ids is available in ./symbolIDs.pdf. ./symbol_lexicon_for_data_collection.txt is the lexicon of all the words mapped to the corresponding symbol ids.

Annotation: The data collected has been annotated at symbol level. Ink files corresponding to each user are located under ./usr*/ directory. Only the data that was marked as good after annotation is made available in this data set. The .SEGMENT SYMBOL string in ink file gives the details of the stroke number to symbol id mapping.

For example in `.SEGMENT SYMBOL 0,1 OK "3 0 __"`

`0,1` implies that strokes 0 and 1 form SYMBOL 3.