5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Developing Speech Resources from Parliamentary Data for South African English

Febe de Wet*, Jaco Badenhorst, Thipe Modipa
Human Language Technology Research Group, CSIR Meraka, South Africa

Abstract

The official languages of South Africa can still be classified as under-resourced with respect to the speech resources that are required for technology development. Harvesting speech data from existing sources is one means to create additional resources. The aim of the study reported on in this paper was to improve the harvesting and transcription accuracy of a corpus derived from parliamentary data. This aim was achieved by improving on the text normalisation process and pronunciation modelling as well as by iteratively training more accurate in-domain acoustic models. In this manner, more data could be harvested with higher confidence than using baseline pronunciation dictionaries and out-of-domain speech data.