INDIAN JOURNAL OF APPLIED LINGUISTICS
VOL. 42, NO. 1-2, JAN-DEC 2016

The Carriage of Indian Languages Corpora: And Miles to Go Before We Stop

NILADRI SEKHAR DASH
Indian Statistical Institute, Kolkata, India

ARULMOZI SELVARAJ
University of Hyderabad, India

MAZHAR HUSSAIN
Jawaharlal Nehru University, New Delhi

ABSTRACT

Exactly after 25 years (1991-2015) of the first pan-Indian effort for generating digital text corpora in all the Indian languages included in the 8th Schedule of the Constitution of India it is perhaps necessary to look back at the road that we have traversed in the last 25 years to review how bumpy or comfortable the journey was, what are the milestones achieved, and how far we need to travel before we reach our destination. This paper, in the form of a narrative of this journey, likes to focus on these issues with some specific goals: to report on stock verification, to highlight the milestones achieved, to shed lights on the soggy slopes of the road, to look into the state we stand, and to reset the targets for our future journey. This paper, thus, is an engaging window through which, for the first time, we are in a position to know the state and status of language corpora generation activities in Indian languages, as well as to evaluate our achievements and aspirations, successes and failures, and our dreams and realizations. Here lies the relevance of this paper in the wider spectrum of the discipline that tries to bring out the Indian languages from the closet of resource-poor state to the arena of resource-rich status to put them in the larger canvas of "Digital India."