Background

One of the major objectives of Science and Heritage Initiative (SandhI) at IIT Bombay is to create a large data-base of Sanskrit texts (to begin with that dealing with Science and Technology) in searchable format. To meet this end, we decided to resort to OCR techniques. Given the limitations of different tools available, some efforts were launched to synthesize the available tools and thereby increase efficiency of the process.

The problem

The languages rich in inections, with Sanskrit at their peak, suffer the problem of text curation in applications like OCR, Speech Recognition, Machine Translation etc. since the vocabulary size is not easily scalable for a variety of reasons including the seemingly countless number of compounds that can be in principle formed in this language.

Our approach to mitigate the problem

Improving the different kinds of vocabularies or auxiliary sources can help improve the con_dence level of recognized words, or equivalently reduce the sample space of recognized characters/words. Especially in Sanskrit language, it can also help in the analysis of compound words. To this end, we are coding the rules given in Paninian grammar as explained in Siddhantakaumudi of Bhattoji Dikshita, and typing their root forms to synthesize the compound words using already available in dictionaries: Sabdakalpadruma, Vacaspatyam and Amarakosa. We also use such rules to detect the erroneous words in Sanskrit OCR Documents and generate the suggestions for erroneous words using standing pop-up mode. For this, we have developed the Sanskrit OCR Document Spell-Checker application, wherein the various auxiliary sources like general dictionaries, domain words, regional words, OCR character confusions, words from multi-OCR systems etc. are successfully exploited in conjunction with the rules given by Panini. Such auxiliary sources are also updated on the y leading to a the tangible reduction in human efforts. We also provide suitable color coding in our application that enhance readability. To facilitate the user to minimize efforts in typing, and avoid the unicode re-ordering issues we have adopted typing in SLP1 format.

Prof. Ganesh Ramakrishnan, Computer Science Engineering
Prof. K. Ramasubramanian, Humanities and Social Sciences, ram@hss.iitb.ac.in