The role of linguistic knowledge in Sanskrit OCR

Oliver Hellwig

Universität Heidelberg

16 May 2013

Digitizing Sanskrit texts poses several challenges to an OCR engine. Most importantly, recognized text cannot be validated using a simple dictionary lookup strategy, because Sanskrit is able to generate a potentially unlimited set of linguistically valid strings. The presentation discusses the problems involved in validating recognition output of a Sanskrit OCR engine, and sketches possible solutions to this problem.