Automated generation of standardized word forms for the Old High German text corpus

The project ‘Referenzkorpus Altdeutsch’ (Old German Reference Corpus) aims to establish a deeply-annotated text corpus of all extant Old German texts. One annotation layer is to contain a standardized version of all word tokens in addition to the versions of the printed editions that are standardized only to some extent, if at all. The generation and incorporation of these standard word forms from standardized lemmata and inflectional information is described by way of example for the Old High German part of the corpus. As is shown, in some cases, also the attested records are needed to determine as exactly as possible the shape of the standard word form to create.