A general purpose spelling correction tool in the post-OCR error correction task: comparative evaluation and feasibility study

A general purpose spelling correction tool in the post-OCR error correction task: comparative evaluation and feasibility study

Kalliopi Zervanou
Utrecht University
k.a.zervanou@uu.nl

Job Tiel Groenestege
Gridline
job@gridline.nl

Dennis de Vries
Gridline
dennis@gridline.nl

Tigran Spaan
Gridline
tigran@gridline.nl

Wouter Klein
Utrecht University
W.Klein@uu.nl

Jelle van der Ster
Utrecht University
jellevdster@gmail.com

Peter van den Hooff
Utrecht University
P.C.vandenHooff@uu.nl

Frans Wiering
Utrecht University
F.Wiering@uu.nl

Toine Pieters
Utrecht University
t.pieters@uu.nl

The digitisation process for text documents that were not born digital typically entails document image scanning, often followed by optical character recognition (OCR), so as to make the text machine readable for subsequent information processing. Despite the progress in OCR systems software, OCR text output still often contains so much error, that both human readability and computer processing is impaired. That is especially the case for old documents, where page quality is degraded and font style often follows obsolete typographic conventions. A solution to this problem is provided by post-OCR error correction methods which often combine corpus statistics with lexical resources (Reynaert, 2014a) and/or additional linguistic information (Baron et al. 2009) with human rule pattern input (Vobl et al. 2014). In this work, we investigate the feasibility of using a general purpose spelling error correction application, the Gridline Taalserver (TM) toolsuite, the knowledge resources of which have been optimised for post-OCR error correction. For this purpose, we comparatively evaluate the Gridline Taalserver toolsuite to the latest version of TICCLops (Reynaert, 2014a), a purpose built post-OCR error correction tool, using two evaluation corpora, the 1800s EDBO DPO35 OCR gold standard corpus (Reynaert, 2014b) and a subcorpus of 1950s newspapers from the VU-DNC corpus (VU DNC). The results of our comparative evaluation show that both approaches present advantages in reducing error and could be applied in combination.