

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

Citation for published version (APA):

Zervanou, K., Groenestege, J. T., Vries, D. D., Spaan, T., Klein, W., Ster, J. V. D., Hooff, P. V. D., Wiering, F., & Pieters, T. (2015). A general purpose spelling correction tool in the post-OCR error correction task: comparative evaluation and feasibility study. In *CLIN 2015 : Book of Abstracts for the 25th Meeting of Computational Linguistics in the Netherlands : Antwerp 5 & 6 February, 2015* (pp. 76). CLIPS.

Document status and date:

Published: 01/01/2015

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

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.