Originally published in: 41. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Universität Bremen, Rahmenthema: Kontrast und Opposition, 06. – 08. März 2019. - Bremen: Deutsche Gesellschaft für Sprachwissenschaft, 2019. Pp. 301

OCR post-correction of the Royal Society Corpus based on the noisy channel model

Carsten Klaus[1], Dietrich Klakow[1] & Peter Fankhauser[2]
[1]*Saarland University,* [2]*IDS Mannheim*
cklaus@lsv.uni-saarland.de, dklakow@lsv.uni-saarand.de, fankhauser@ids-mannheim.de

Linguistic analysis of historical texts are posing substantial challenges for researchers. Old documents are often of inferior quality hence errors occur during the reading process. Such errors can be severe disruptive factors for further processing and analysis. In this contribution we introduce the *Noisy Channel Spell Checker*, an approach for automatic detection and correction of optical character recognition (OCR) induced misspellings in historical data, with a particular focus on the *Royal Society Corpus*. This corpus is a collection of scientific texts from 1665 to 1869 published in the journal *Philosophical Transactions of the Royal Society of London*. It comprises about 10.000 documents with 35.000.000 tokens in total. Due to the old material words have been recognized incorrectly thus leading to files corrupted by thousands of OCR misspellings. This motivates a post processing step. (UdS Fedora Commons n.d.). The current correction technique is a pattern-based approach (Knappen 2016). Misspellings are corrected by applying a replacement mechanism for substituting the errors with their corresponding correction. Due to its lack of generalization it suffers from bad recall. For that a new approach is required. The *Noisy Channel Spell Checker* is based on the noisy channel model (Shannon, 1948) which is able to estimate the most likely correction. The special characteristic of the tool's model is that the training is completely corpus-specific. It does not require any annotation of training data since only the Royal Society Corpus itself is used.

For the purpose of evaluation we extracted a subset of documents from the corpus and corrected it manually to create a ground truth. Then we applied our tool and the pattern-based approach separately. With an F1-Score of 0.61 the *Noisy Channel Spell Checker* significantly outperforms the pattern-based state of the art which only accomplishes an F1-Score of 0.28. Enhancing the denoising of the Royal Society Corpus will promote further investigation of the historical data. Thus it is conceivable to replace the current correction technique permanently with the approach presented in this contribution.

References: • Kermes, H., S. Degaetano-Ortlieb, A. Khamis, J. Knappen & E. Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). • Knappen, J., S. Fischer, H. Kermes, E. Teich & P. Fankhauser. 2008. The making of the Royal Society Corpus. *ListLang@NoDaLiDa*. • Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*. • UdS Fedora Commons Repository (n.d.) The Royal Society Corpus (RSC), https://fedora.clarin-d.uni-saarland.de/rsc/. (last accessed 29.03.2018)