Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects

Martin Reynaert

Induction of Linguistic Knowledge, Tilburg University, The Netherlands

Abstract. This paper proposes a non-interactive system for reducing the level of OCR-induced typographical variation in large text collections, contemporary and historical. Text-Induced Corpus Clean-up or TICCL (pronounce 'tickle') focuses on high-frequency words derived from the corpus to be cleaned and gathers all typographical variants for any particular focus word that lie within the predefined Levenshtein distance (henceforth: LD). Simple text-induced filtering techniques help to retain as many as possible of the true positives and to discard as many as possible of the false positives. TICCL has been evaluated on a contemporary OCR-ed Dutch text corpus and on a corpus of historical newspaper articles, whose OCR-quality is far lower and which is in an older Dutch spelling. Representative samples of typographical variants from both corpora have allowed us not only to properly evaluate our system, but also to draw effective conclusions towards the adaptation of the adopted correction mechanism to OCR-error resolution. The performance scores obtained up to LD 2 mean that the bulk of undesirable OCR-induced typographical variation present can fully automatically be removed.

1 Introduction

This paper reports on efforts to reduce the massive amounts of non-word word forms created by OCRing large collections of printed text in order to bring down the type-token ratios of the collections to the levels observed in contemporary 'born-digital' collections of text. We report on post-correction of OCR-errors in large corpora of the Cultural Heritage. On invitation by the National Library of The Netherlands (Koninklijke Bibliotheek - Den Haag) we have worked on contemporary and historical text collections. The contemporary collection comprises the published Acts of Parliament (1989-1995) of The Netherlands, referred to as 'Staten-Generaal Digitaal' (henceforth: SGD)[1]. The historical collection is referred to as 'Database Digital Daily Newspapers' (henceforth: DDD)[2], which comprises a selection of daily newspapers published between 1918 and 1946 in the Netherlands. The historical collection was written in the Dutch spelling 'De Vries-Te Winkel', which in 1954 was replaced by the more contemporary spelling

[1] URL: http://www.statengeneraaldigitaal.nl/

[2] URL: http://kranten.kb.nl/ In actual fact, this collection represents the result of a pilot project which is to be incorporated into the far more comprehensive DDD.