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

Kalliopi Zervanou, Job Tiel Groenestege, Dennis de Vries, Tigran Spaan, Wouter Klein, Jelle van der Ster, Peter van den Hooff, Frans Wiering, Toine Pieters

♦ OCR-error correction

- ***Optical character recognition (OCR)*::** required for making scanned document texts machine-readable
- **But::** OCR text output often contains a lot of error, especially when the scanned document is old
- **Solution::** ***Post-OCR Error Correction Methods:*** combine corpus statistics with lexical/rule resources

♦ Spelling-error correction

Tools developed for a variety of (mostly) modern languages

➢ **To what extent can we use an existing spelling correction tool for OCR-error correction?**

♦ TICCLops [Reynaert, DATECH 2014]

- Uses a variety of modern, historical & semantic resources
- Applies on a range of file formats (text, FoLiA, hOCR…)
- Candidate corrections ordered by *Levenshtein distance* (LD)
- ***TICCL-indexer:*** indexes *all* possible character confusions given LD
- Does not fully correct hyphenated words
- Resources complemented by input corpus TICCL-indexer

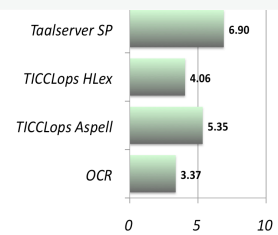
♦ GRIDLINE Taalserver Spellchecker

- CGN lexicon
- Applies on tokenised text (GRIDLINE XML)
- Candidate corrections ordered by edit distance
- Typical errors are taken into consideration
- Corrects hyphenation, word boundaries, OCR noise
- Currently without historical lexical resources

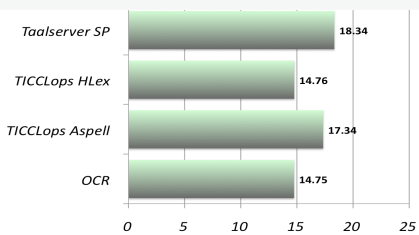
♦ Comparative evaluation

VU-DNC OCR corpus: 1950s KB newspaper articles

Character Error Rate

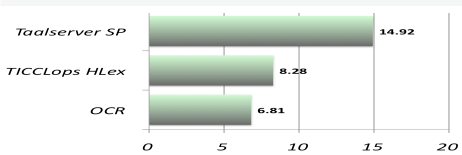


Word Error Rate

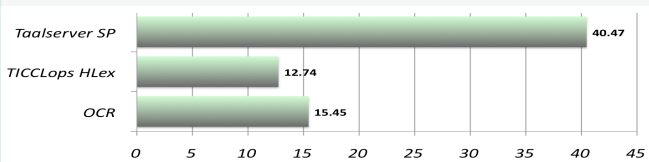


1800s EDBO DPO35 OCR corpus [Reynaert, COLING 2014]

Character Error Rate



Word Error Rate



Evaluation Setup

- Subset of VU-DNC dataset, 50 newsarticles from 1950s (Algemeen Dagblad, NRC, De Telegraaf, Trouw, De Volkskrant)
- EDBO DPO35 dataset: Martinet, J. F. (1789) *Kort begrip der waereld-historie voor de jeugd*. KB:DPO35.
- OCR evaluation tool: `ocrevalUAtion` [Carrasco, DATECH 2014]
- TICCLops with Aspell & Contemporary/Historical/NE (HLex) resources
- Baseline: plain OCR output

♦ Conclusions & Future Work

- Ground truth not comparable to OCR-error correction results:
  - token alignment, spaces, sentence splitting, hyphenation
- TICCLops best performance in older data (1800s)
- Taalserver Spellchecker performance comparable to TICCLops with basic modern language resources (Aspell) in the 1950s corpus
- Suitable lexico-semantic resources (historical, named entities) play an important role
- Tools performance on spelling "modernisation" requires further evaluation experiments
- Qualitative evaluation of OCR & OCR correction readability required

Universiteit Utrecht

GRIDLINE