TEXT-INDUCED CORPUS CLEAN-UP: INTRODUCTION

TICCL for TYPOS and OCR-errors

- Tool to perform large scale, unsupervised spelling correction of corpora
- Spelling correction = reduction of lexical variation caused by typos, OCR-errors, historical orthographical changes...
- Currently no direct text editing, but linking of variants to their most likely canonical form
- Prototype developed during a pilot project by invitation of the National Library, The Hague
- Production version developed according to KB specifications, second half 2008
## EVALUATION: Het Volk 1918

<table>
<thead>
<tr>
<th></th>
<th>Overall Score</th>
<th>Score per LD</th>
<th>Cumulated score per LD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FN</td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>types</td>
<td>1414</td>
<td>1671</td>
<td>134</td>
</tr>
<tr>
<td>tokens</td>
<td>11334</td>
<td>3163</td>
<td>2971</td>
</tr>
<tr>
<td>LD 1</td>
<td>312</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>LD 2</td>
<td>1039</td>
<td>60</td>
<td>108</td>
</tr>
<tr>
<td>LD 3</td>
<td>62</td>
<td>1569</td>
<td>16</td>
</tr>
</tbody>
</table>

- **TP**: True Positives
- **FN**: False Negatives
- **FP**: False Positives
- **R**: Recall
- **P**: Precision
- **F**: F1 Score
A demonstration project which will allow CLARIN users to submit their corpora for fully automatic spelling correction and normalization by TICCLops, the online processing version of our core component TICCL. This system should be widely applicable in all manner of curation projects and lexicographical work.

Start Date: 15 January 2010
End Date: 15 July 2010
TICCLops: PARTNERS

- **Coordination & Technology Provider:**
  Tilburg centre for Creative Computing (TiCC) - Tilburg
  - Martin Reynaert: Researcher
  - Maarten van Gompel - Scientific Programmer

- **User & Data Provider:**
  Koninklijke Bibliotheek (KB) - The Hague
  - Astrid Verheusen - Head Digitisation Department

- **CLARIN Center & Data Provider:**
  Instituut voor Nederlandse Lexicologie (INL) - Leiden
  - Remco van Veenendaal - Head TST-Centrale
Research Data

- Databank Digitale Dagbladen: Digitized newspapers: we also have a gold standard for Het Volk-1918 (http://www.kb.nl/hrd/digi/ddd/index-en.html).
- New collections of recently digitized copyright-free books and magazines
Demonstration Scenario

- Digitized copyright-free book and its gold standard
- All aspects of using the system will be demonstrated to prospective users on the basis of this book
- Aspects include pre-packaging the corpus (i.e. the book) as a compressed archive, choosing and setting the system parameters and receiving and interpreting the results
- Actual evaluation results returned will then teach the prospective user what to expect from the system and to choose which output formats best suit his own needs
KORT BEGRIP DER WAERELD-HISTORIE VOOR DE JEUGD, DOOR

44950.
Kort Begrip der Waereld-Historie, voord de Jeugd, door J. F. Martinet, Predikant te Zutphen, &c. Met Kaarten. Te Amsterdam, bij Johannes Allart. MDCCLXXIX.
Thanks!!

Thanks for your attention!

Papers about TICCL are available at:
http://ilk.uvt.nl/

TICCLops

Martin REYNAERT

Tilburg centre for Creative Computing

CLARIN-NL, Utrecht. 19 February, 2010