Towards TICCLAT, the next level in Text-Induced Corpus Correction
Martin Reynaert, Maarten van Gompel, Ko van der Sloot and Antal van den Bosch
Meertens Institute - Amsterdam / TICCL - Tilburg University / Radboud University Nijmegen

- Having worked in Dutch CLARIN projects for going on for a decade, we want to give a brief overview of what we have achieved.
- Our work centers around facilitating the building of text corpora, around improving the lexical quality of text corpora, around enhancing search and retrieval and around providing necessary infrastructure for research to actually achieve all these steps on their own.

Column 1 introduces major extensions to our Column 2 provides information about our latest project TICCLAT.

- Column 3 outlines how TICCL is embedded in a larger corpus building system called PICCL and lists how the systems are made available to the larger community.

Multilingual and diachronic Text-Induced Corpus Clean-up
- The Text-Induced Corpus Clean-up system TICCLAT has now been ported from Perl to distributable (in both senses of being shareable and being parallelizable) C++ code. It has been rethought to be multilingual and diachronic.
- TICCL uses:
  - a large lexicon consisting of validated as well as background corpora derived uni-, bi- and trigrams for context information for resolving split or run-on words and short words. This results in better ranking of Correction Candidates (CCs) overall.
  - Chaining: “my friends’ friends are my friends”: the CC with best-first ranked variants within the set Levenshtein Distance (LD) that act as CCs for further variants beyond this LD (and so on for even greater Lds) is directly linked to these larger LD variants.

TICCLAT: recent developments
- Recent developments in TICCL
  - Language recognition: performed on the level of text paragraphs
  - Focus list: TICCL used to evaluate all words, whether from corpus to be corrected or from larger background corpus. This is now limited to
  - Harnessing various OCR-versions of the text
  - Word bi/ trigram correction: utilizing local word context information for resolving split or run-on words and short words. This results in better ranking of Correction Candidates (CCs) overall.
  - Chaining: “my friends’ friends are my friends”: the CC with best-first ranked variants within the set Levenshtein Distance (LD) that act as CCs for further variants beyond this LD (and so on for even greater Lds) is directly linked to these larger LD variants.

TICCLAT builds on the Nederlab corpora
- The Nederlab project has brought together major collections of digitized texts relevant to the Dutch national heritage (c. A.D. 800 – present) consisting of terabytes of data in a unified format, i.e. FoLiA XML.
- The focus in Nederlab has been on incorporating the vast digital text collections of the Koninklijke Bibliotheek (KB or Dutch National Library) as well as the contents of the Digitale Bibliotheek voor de Nederlandse Letteren (DBNL - The Digital Library of Dutch Literature).
- KB text collections comprise newspapers from 1618 to 1995 and the mainly 18th century Early Dutch Books Online or EDBO (http://www.dbnl.nl/), as well as the Staten-Generaal Digitaal (1815-2013).
- All results of large digitization programmes have in common that they are riddled with OCR misrecognition errors.
- These texts spanning twelve centuries present a wealth of diachronic and regional spelling variation, besides the even wider OCR-variation.
- TICCLAT is to learn from, incorporate and re-apply all this variation.

Acknowledgements
- Martin Reynaert’s work is supported by the Netherlands Organisation for Scientific Research (NWO) and the Netherlands Academy of Arts and Sciences (KNAW) in cooperation with CLARIN and the Netherlands eScience Center. More info: reynaert@uu.nl

This poster online: http://ticclops.uvt.nl/TICCLAT.CLARIN2018poster.MRE.pdf