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 / TiCC - Tilburg University / Radboud University Nijmegen

CLARIN Conference 2108 Hotel Galilei, Pisa, Italy October 8-10, 2018

Towards TICCLAT: Intro

- 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 researchers to actually achieve all these steps on their own.
- Column 1 introduces major extensions to our OCR post-processing system TICCL
- 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

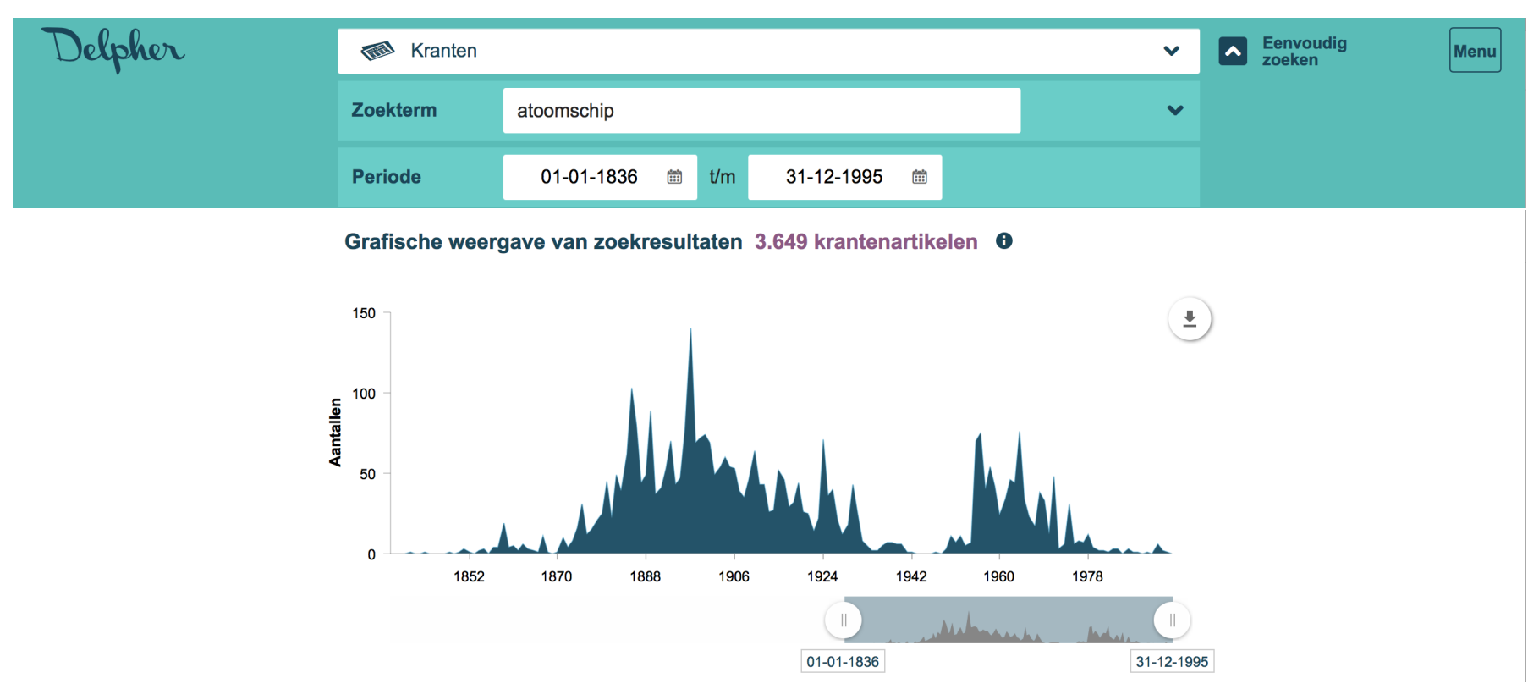
- The Text-Induced Corpus Clean-up system TICCL has now been ported from Perl to distributable (in both senses of being shareable and being parallelizable) C++ code. It has been rethought to be multilingual and diachronic.
- TICCL uses:
  - a large lexicon consisting of validated as well as background corpus derived uni-, bi- and trigrams
  - exhaustive word variant look-up up to a given Levenshtein distance due to anagram hashing
  - a combination of corpus-induced ranking features to determine the most likely correction candidate

TICCL: 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
- 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

TICCL: expected contribution



Kranten

Zoekterm: atoomschip

Periode: 01-01-1836 t/m 31-12-1995

Grafische weergave van zoekresultaten 3.649 krantenartikelen

Digitized Dutch newspapers Online:

- Dutch National Library web site `https://www.delpher.nl/nl/kranten/` query for 'atoomschip' (E. 'nuclear ship') in the digitized newspapers' section
- It should be obvious that any newspaper occurrence of 'atoomschip' before about 1945 most likely represents a real word error for 'stoomschip' (E. steam ship) due to OCR misrecognition of 's' for 'a'
- We hope that TICCL's further development in project TICCLAT will remedy this state of affairs

The TICCLAT project

- One of four projects awarded in the joint CLARIAH – eScience Center ADAH call
- ADAH = 'Accelerating Scientific Discovery in the Arts and Humanities'
- Budget: 300K euro
- Runtime: Started January, 2018. Ends mid 2019.
- Location: TICCLAT is hosted by the Meertens Institute, Amsterdam
- Principal Investigator: Martin Reynaert

TICCLAT's goals

- TICCLAT will extend TICCL's spelling and OCR correction capabilities with classification facilities based on specific data collected from the full diachronic Dutch Nederlab corpus: word statistics, document and time references and linguistic annotations
- TICCL as a corpus correction tool will be transformed into TICCLAT, a lexical assessment tool capable of deciding whether or not a particular character string constitutes a 'word' or not.
- TICCLAT's capabilities will also be evaluated in comparison to human performance by an expert psycholinguist

TICCLAT: Abstract

The Text-Induced Corpus Clean-up tool TICCL, integral part of the CLARIN infrastructure, is globally unique in utilizing the corpus-derived word form statistics to attempt to fully-automatically post-correct texts digitized by means of Optical Character Recognition.

The NWO 'Groot' project Nederlab has delivered a uniformly processed and linguistically enriched diachronic corpus of Dutch likely containing towards 20 million word tokens. We aim to extend TICCL's correction capabilities with classification facilities based on specific data collected from the full Nederlab corpus: word statistics, document and time references and linguistic annotations, i.e. Part-of-Speech and Named-Entity labels. These data will complement a solid, renewed basis composed of the available validated lexicons and name lists for Dutch.

In this, TICCL as a post-correction tool will be transformed into TICCLAT, a lexical assessment tool capable of delivering not only correction candidates, but also e.g. more accurately dated diachronic Dutch word forms, more securely classified person and place names. To achieve this on scale, the TICCLAT project seeks a successful merger of TICCL's anagram hashing with bit-vectorization techniques. TICCLAT's capabilities will be evaluated in comparison to human performance by an expert psycholinguist.

The data collected will be exportable for storage in a data repository, as RDF triples, for broad reuse. The project will greatly contribute to a more comprehensive overview of the lexicon of Dutch since its earliest days and of the person and place names that share its history. Its partners are the Dutch experts in Lexicology, Person Names and Toponyms.

NWO Groot project Nederlab: The corpora

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 (`http://www.kb.nl/en`) (KB or Dutch National Library) as well as the contents of the Digitale Bibliotheek voor de Nederlandse Letteren (`http://www.dbnl.org/`) (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.delpher.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 wilder OCR-variation.
- TICCLAT is to learn from, incorporate and re-apply all this variation.

CLARIAH / Netherlands eScience Center ADAH Call project TICCLAT

TICCLAT's embedding

- TICCLAT's broader environment will be in PICCL, a corpus-building work flow available in the CLARIAH infrastructure.
- In PICCL, the 'Philosophical Integrator of Computational and Corpus Libraries', TICCLAT is to supersede TICCL, largely extending and enhancing the current system.
- The name PICCL refers to the preceding project, @PhilosTEI. The main idea behind this prior CLARIN-NL project was to enable philosophers to submit scans of philosophical works they need for their work to the online system and to receive back an electronic version suitable for further processing into a critical edition.

@PhilosTei

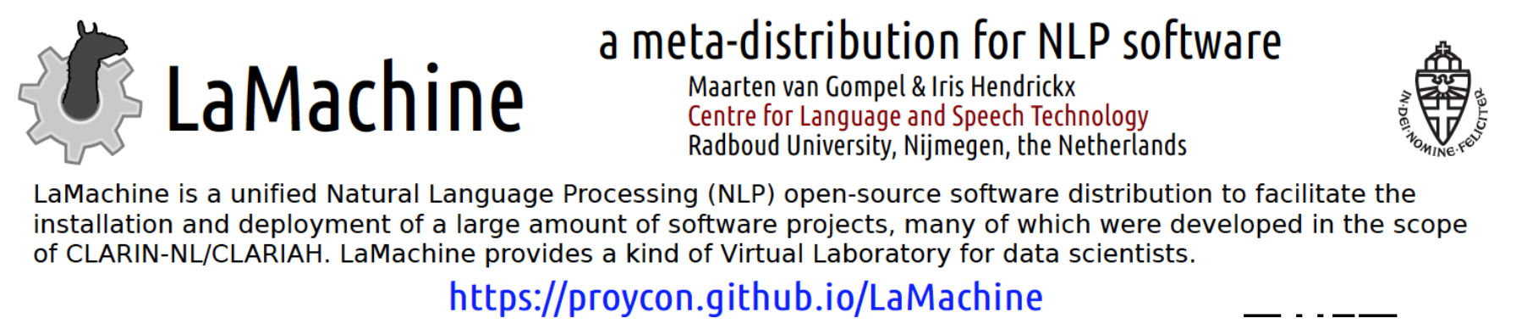
- The first of PICCL's stages are corpus ingestion, text conversion or digitization and text correction, normalization and/or modernization.
- The further stages of the work flow comprise linguistic enrichment by Frog and indexation by means of AutoSearch (courtesy of INT) towards online exploration/exploitation.

Main Work Flow Components for corpus building

- Conversion: a choice selection of available open-source image and text convertors have been incorporated in the work flow. The term 'philosophical' in the system's name should be understood to denote: 'well-considered'.
- Optical Character Recognition: Tesseract is currently the OCR engine of choice in the PICCL work flow.
- Pivot format: the format of choice central to the whole work flow is FoLiA xml.
- OCR post-correction: a new, modular and distributable implementation of Text-Induced Corpus Clean-up (as an online processing system) or TICCL(ops) provides diachronic and multilingual normalisation and transcription facilities.
- Linguistic enrichment: Text can be tokenized by Ucto or, if Dutch, further linguistically enriched by Frog.
  - Lemmatization
  - Part-of-Speech annotation
  - Named Entity labeling
  - Optionally: chunking, morphological analysis and dependency parsing
- The whole work flow gains parallelization capabilities through Nextflow.

PICCL: distribution

- PICCL is shipped as part of LaMachine
- Please see our colleagues' poster at this conference!



LaMachine

a meta-distribution for NLP software

Maarten van Gompel & Iris Hendrickx
Centre for Language and Speech Technology
Radboud University, Nijmegen, the Netherlands

LaMachine is a unified Natural Language Processing (NLP) open-source software distribution to facilitate the installation and deployment of a large amount of software projects, many of which were developed in the scope of CLARIN-NL/CLARIAH. LaMachine provides a kind of Virtual Laboratory for data scientists.

https://proycon.github.io/LaMachine

Acknowledgements

Martin Reynaert's work is supported by the Netherlands Organisation for Scientific Research (NWO) and the Royal Netherlands Academy of Arts and Sciences (KNAW) in cooperation with CLARIAH and the Netherlands eScience Center. More info: email: reynaert@uvt.nl

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