PICCL constitutes a complete workflow for corpus building. It is to be the integrated result of recent developments in the CLARIN-NL project @PhilosTEI which ended November 2015, and further work in NWO ‘Groot’ project Nederlab, which continues till end 2018 and in CLARIAH, which will run till 2020.

Introduction

CLARIN activities in the Netherlands in 2015 are in transition between the first national project CLARIN-NL and its successor CLARIAH. In our paper we give an overview of important infrastructure developments which have taken place throughout the first and which are to a further level in the second. We show how relatively small accomplishments in particular projects enable larger steps in further ones and how the synergy of these projects helps the national infrastructure to outgrow mere demonstrators and to move towards mature production systems. We present a new corpus building tool called PICCL. This integrated pipeline offers a comprehensive range of conversion facilities for legacy electronic text formats. Optical Character Recognition for text images, automatic text correction and normalization, linguistic annotation, and preparation for corpus exploration and exploitation environments. We give a concise overview of PICCL’s components, integrated now or to be incorporated in the foreseeable future.

Main Work Flow Components for corpus building

The term ‘philosophical’ in the system’s title should be understood to denote: ‘well-considered’, i.e. we aim to integrate into the workflow only the best pick of available tools for the various jobs to be done.

- Conversion: a choice selection of available open-source image and text converters has been and are to be incorporated in the workflow.
- Optical Character Recognition: Tesseract (further development supported by Google) is currently the OCR engine of choice in the @PhilosTEI work flow.
- Pivot format: the format of choice central to the whole workflow 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.
- Book collation: A digitised and post-corrected book can finally be delivered as a single tome in TEI XML format whatever the number of input files, whatever their original format.
- Linguistic enrichment with lemmata, POS-tags, Named Entity labels.
- Indexing towards online availability, cf. the WhiteLab interfaces of OpenSoNaR.

The Text-Induced Corpus Clean-up system TICCL has now been largely 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.

- We have incorporated into TICCL the largest extant historical lexicon for Dutch and its accompanying historical name list. Both were developed at INL (http://www.ini.nl/), the Dutch Institute for Lexicology. They were deliverables of the European project Impact and are available through the Impact Centre of Competence (http://impact-centre.eu/).
- The new implementation is easily adaptable to other languages and older language varieties by plugging in special purpose lexicons.
- TICCL uses:
  - a large Dutch historical lexicon and name list
  - recursive word variant look-up to a given Levenshtein distance
  - a running list of known historical character confusions
  - corpus-induced ranking features to determine the most likely correction candidate

Legacy diachronic text and challenges for Digital Humanities

- All projects dealing with diachronic text face the same challenges, whatever the actual language under consideration.
- The @PhilosTEI work flow provides the layman with facilities for building his own digital library. With PICCL all will be able to build their own special-interest corpus according to today’s best practices and standards.
- Automatic normalisation of diachronic text into more modern text will enable to re-use tools developed for modern language varieties.
- FROG, the major Dutch lemmatiser and PICCL(ops) are to be incorporated in English and German.

Further PICCL Functionalities

Output text is in FolIA XML. The pipeline will therefore offer the various software tools that support FolIA. Language categorization may be performed by the tool FolIA-langcat at the paragraph level. TICCL – Text-Induced Corpus Clean-up – performs automatic post-correction of the OCRed text. During handling of numbers of input/output files, taking e.g. x PDF input files apart into y (where y ≥ x) image files to be sent to the OCR engine Tesseract, then presenting the y OCRed files as a single batch to TICCL which eventually corrects the y FolIA XML files to be collated into a single output FolIA XML and also, if the user so desires, a TEI XML output e-book.

- The user-friendly system will be made available as a large black box to process a book’s images into a digital version with next to no user intervention or prior knowledge required. It will more than well be equipped with the necessary interface options to allow more sophisticated users to address any submodule or combination of submodules individually at will.

Acknowledgments

The authors are supported in their work by CLARIAH, the Dutch project within CLARIN. Martin Reynaert’s work is further supported by the Netherlands Organisation for Scientific Research (NWO) in project Nederlab.

The text was prepared by the CLARIN Interoperability and Applications Research Team.

More information: http://tisclope.uni.nl, TICCL and http://philos.tei.clarin.inl.nl