PICCL : Philosophical Integrator of Computational and Corpus Libraries

Martin Reynaert[1][2], Maarten van Gompel[1], Ko van der Sloot[1] and Antal van den Bosch[1]

Center for Language Studies - Radboud University Nijmegen[1] / TiCC - Tilburg University[2]

CLARIN Annual Conference 2015 Wrocław, Poland October 17-18, 2015

CLARIAH project PICCL

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 taken 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

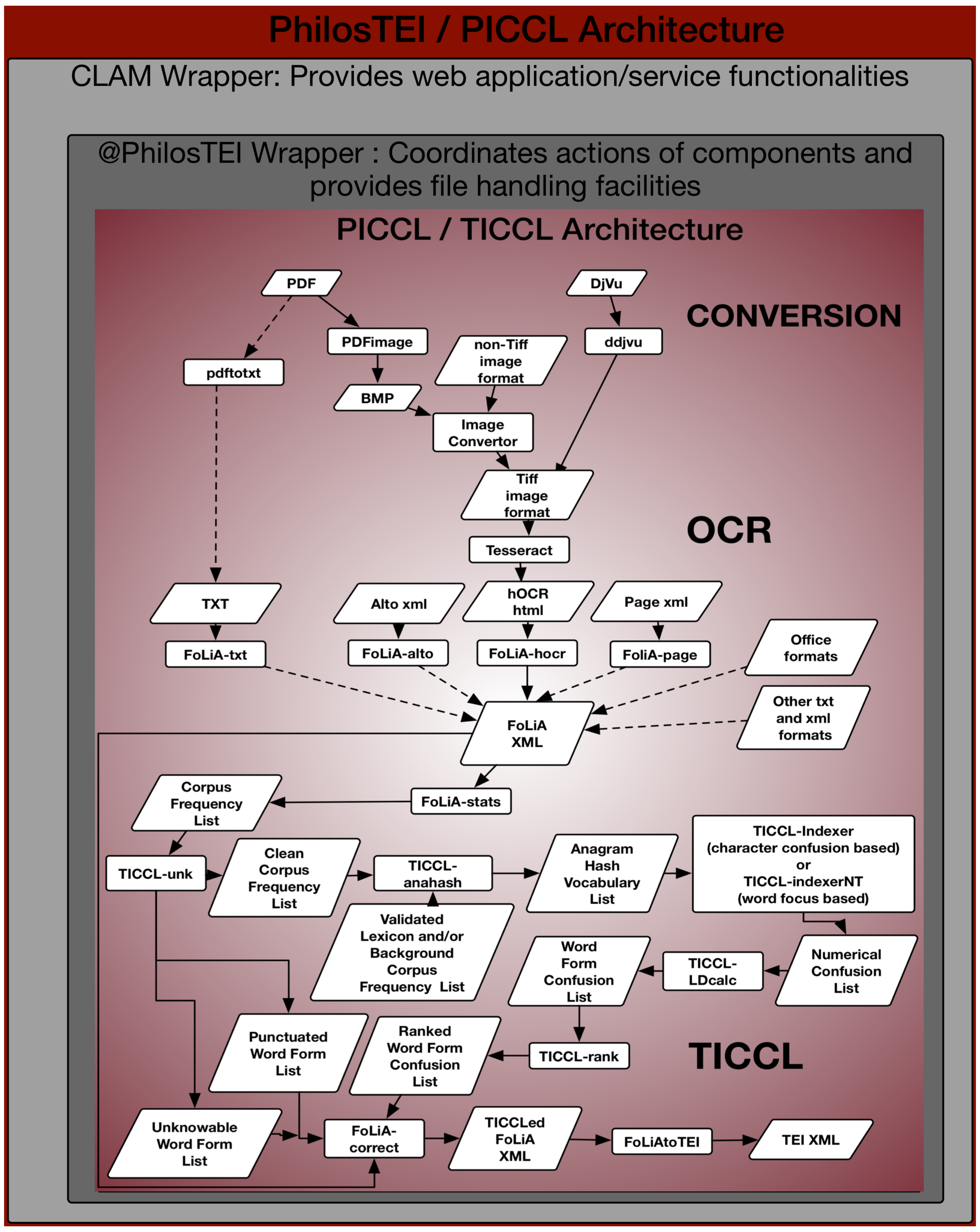
The term ‘philosophical’ in the system’s title should be understood to denote: ‘well-considered’, i.e. we aim to integrate into the work flow only the best pick of available tools for the various jobs to be done.

- Conversion: a choice selection of available open-source image and text convertors have been and are to be incorporated in the work flow.
- 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 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.
- 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.

Tilburg University

Tilburg University

The PICCL Corpus Building Work Flow


PhilosTEI / PICCL Architecture
CLAM Wrapper: Provides web application/service functionalities
@PhilosTEI Wrapper : Coordinates actions of components and provides file handling facilities
PICCL / TICCL Architecture
CONVERSION
PDF
DjVu
PDFimage
pdftotxt
non-Tiff image format
ddjvu
BMP
Image Convertor
Tiff image format
OCR
Tesseract
hOCR html
TXT
Alto xml
Page xml
FoLiA-txt
FoLiA-alto
FoLiA-hocr
FoliA-page
Office formats
Other txt and xml formats
FoLiA XML
Corpus Frequency List
FoLiA-stats
TICCL-unk
Clean Corpus Frequency List
TICCL-anahash
Anagram Hash Vocabulary List
TICCL-Indexer (character confusion based) or TICCL-indexerNT (word focus based)
Validated Lexicon and/or Background Corpus Frequency List
Word Form Confusion List
TICCL-LDcalc
Numerical Confusion List
Punctuated Word Form List
Ranked Word Form Confusion List
TICCL-rank
TICCL
Unknowable Word Form List
FoLiA-correct
TICCLed FoLiA XML
FoLiAtoTEI
TEI XML

PICCL as web application/service

- In contrast to e.g. the Taverna work flow TTNWW built in CLARIN-NL, PICCL is wrapped in a single efficient CLAM-based web service/application. This avoids network overhead and allows for better distributional use of the available hardware through load-balancing.
- The PICCL wrapper allows for flexible handling of numbers of input/output files, taking e.g. x PDF input files apart into y (where $y \geq x$) 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 equally well be equipped with the necessary interface options to allow more sophisticated users to address any submodule or combination of submodules individually at will.

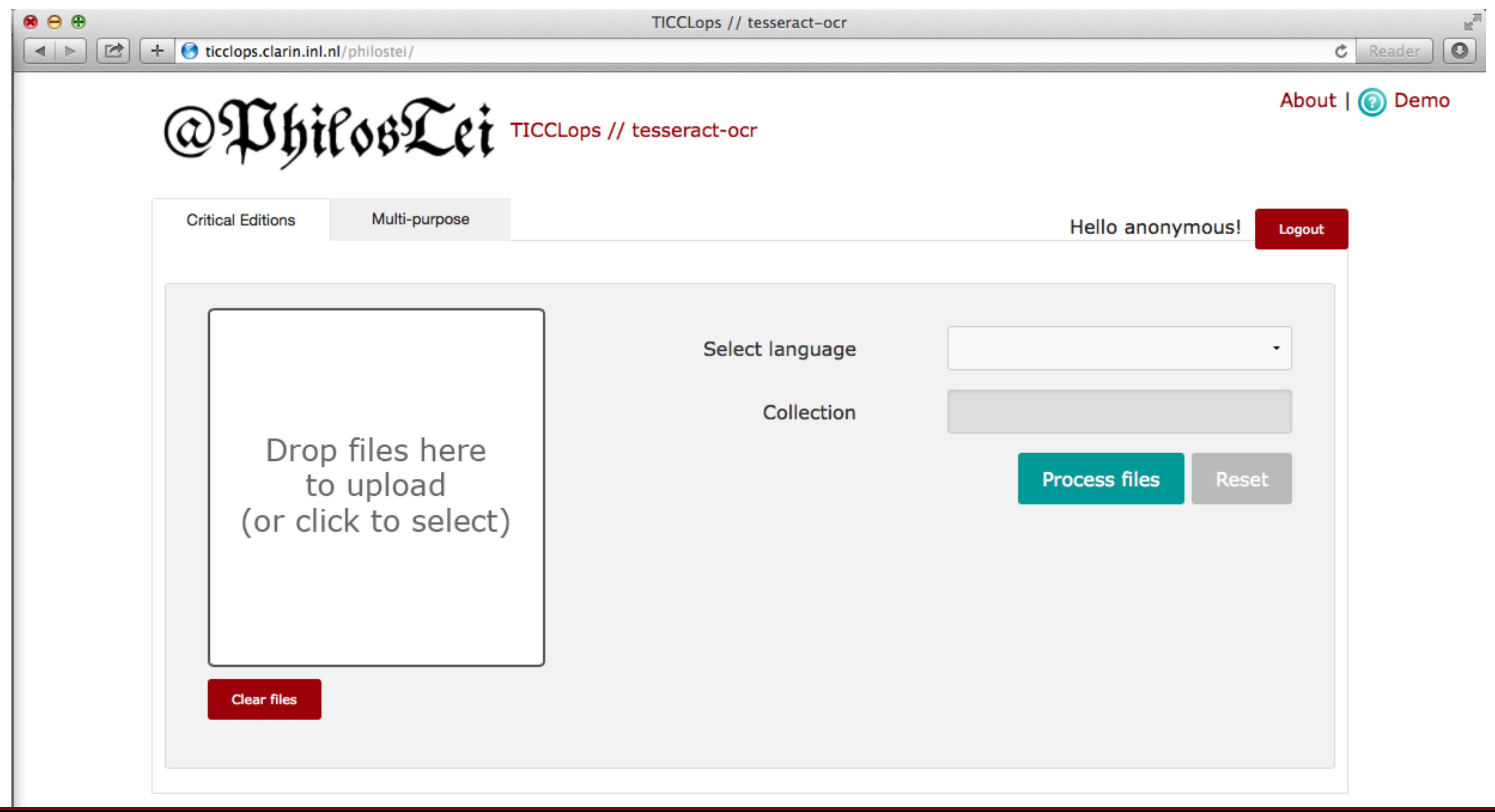
@PhilosTEI demonstrator

PICCL is to be available to all researchers in the CLARIN infrastructure and is hosted by certified CLARIN Centre INL in Leiden. PICCL is to have a highly intuitive user-friendly interface in order to allow even the most computer-weary user to obtain texts in a corpus-ready, annotated format. Already its predecessor, the @PhilosTEI system, in fact provides two distinct interfaces: the more traditional interface type that comes with CLAM [a] as well as a more modern interface that was custom made for the @PhilosTEI project, according to the specifications of the user[b].

[a] http://ticclops.clarin.inl.nl

[b] http://philostei.clarin.inl.nl

Interface of PICCL’s predecessor: @PhilosTEI. Available at http://philostei.clarin.inl.nl



TICCL: Multilingual Text-Induced Corpus Clean-up

- 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.inl.nl/`), the Dutch Institute for Lexicology. They were deliverables of the European project Impact and are available through the Impact Centre of Competence (`http://www.digitisation.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
  - exhaustive word variant look-up up to a given Levenshtein distance
  - a numerical 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 POS-tagger, is due to be integrated in PICCL and to be trained on 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. Dutch texts may optionally be annotated automatically by Frog, i.e. tokenized, lemmatized and classified for parts of speech, named entities and dependency relations. The FoLiA Linguistic Annotation Tool (FLAT) will provide for manual annotation of e.g. metadata elements within the text – for later extraction. FoLiA-stats delivers n-gram frequency lists for the texts’ word forms, lemmata, and parts of speech. Colibri Core allows for more efficient pattern extraction, on text only, and furthermore can index the text, allowing comparisons to be made between patterns in different (sub)corpora. BlackLab and front-end WhiteLab, developed in the OpenSoNaR project, allow for corpus indexing and in-depth online user querying. Convertors to other formats, e.g. TEI XML, will be at hand.

Radboud Research Team

LAnguage MAchines

Acknowledgements

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.

Radboud University Nijmegen

Radboud University

More information: http://ticclops.uvt.nl. TICCLops: http://ticclops.clarin.inl.nl and @PhilosTEI: http://philostei.clarin.inl.nl