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]

Workshop: Morfosyntactisch verrijken van historische teksten
Utrecht 16 November 2015

Faculty of Arts
Centre for Language and Speech Technology
Radboud University Nijmegen

PICCL: Introduction

- We present a new corpus building tool called PICCL. It constitutes a complete workflow for corpus building.
- PICCL is to be the integrated result of recent developments in the CLARIN-NL project @PhilosTEI, which ended November 2014, and further work in NWO ‘Groot’ project Nederlab, which continues till end 2018 and in CLARIAH, which will run till 2020.
- PICCL wants to move beyond demonstrator status and be an actual production system.

PICCL: An integrated pipeline

The integrated PICCL 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 indexing for corpus exploration and exploitation environments

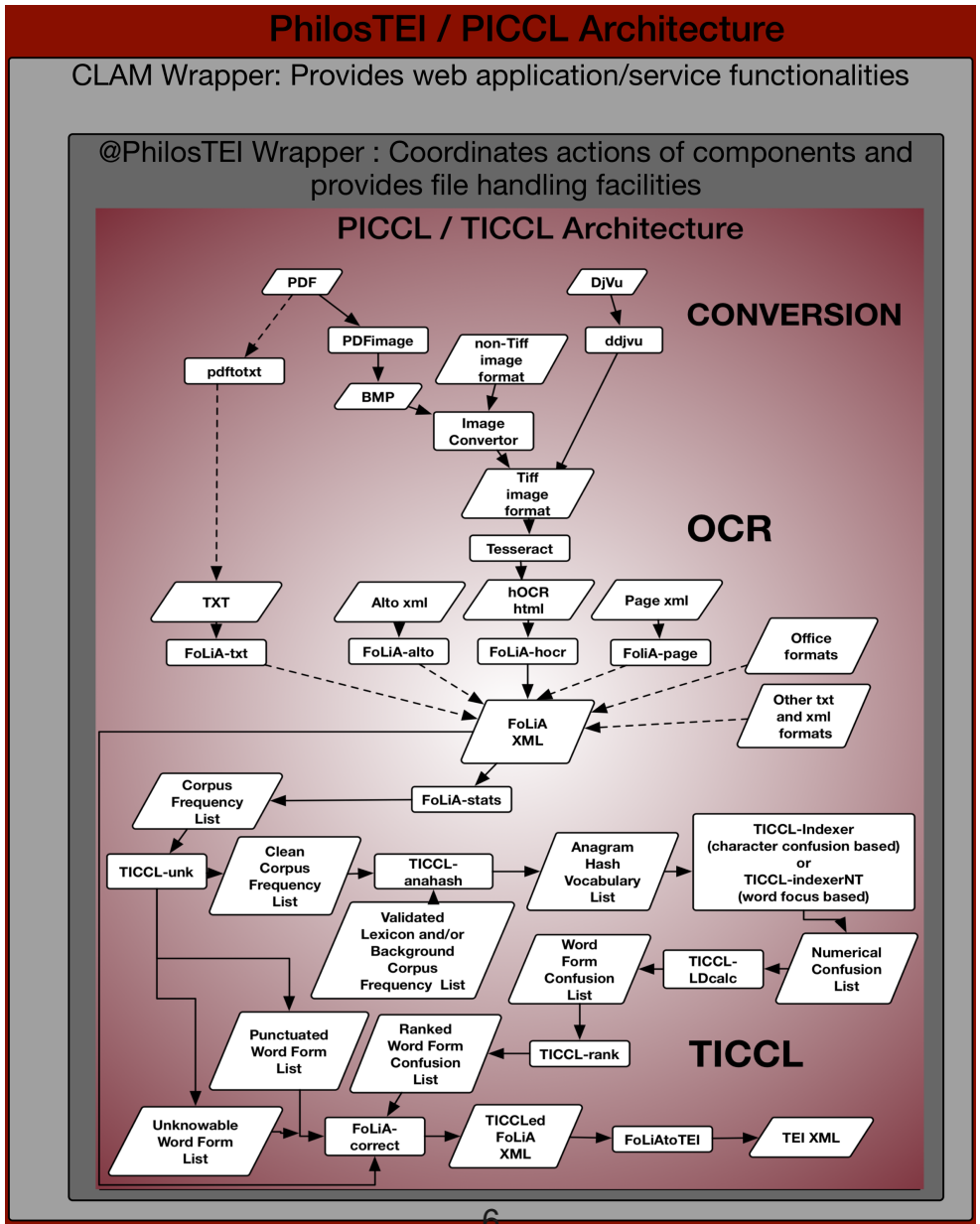
Prior CLARIN-NL project @PhilosTEI: Contribution

- The main idea behind this 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. The preferred format for this is TEI XML. Due to @PhilosTEI we have gained experience with integrating an OCR-engine, i.e. Tesseract, into our corpus building work flow.
- The philosopher-users study works in a broad range of European languages. The OCR post-correction tool TICCL has therefore been made multilingual within this project. In the @PhilosTEI system, we currently have dedicated TICCL web applications and services for 18 European languages/language varieties.

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.
- Optical Character Recognition: Tesseract 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 (online processing system) or TICCL(ops) provides diachronic and multilingual normalisation and transcription facilities.
- Book collation: The digitised and post-corrected book is finally delivered as a single tome in TEI XML format whatever the number of input files, whatever their original format.

PICCL Overview



PICCL Overview

- PICCL is wrapped in a single efficient CLAM-based web service/application. The Computational Linguistic Application Mediator is also one of our early CLARIN-NL achievements.
- 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.

PICCL: further functionalities

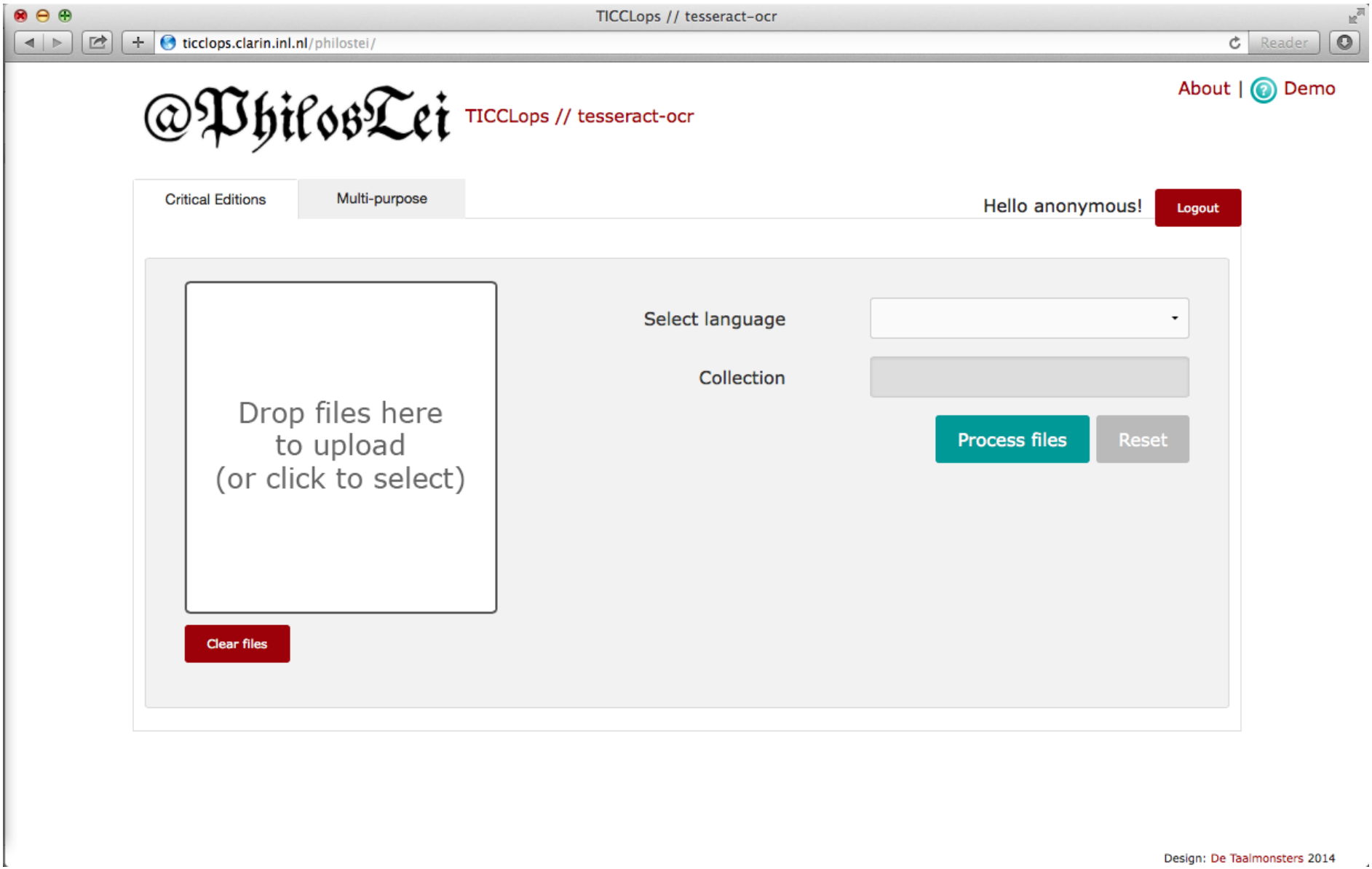
Output text is in FoLiA XML. The pipeline will therefore offer the various software tools that support FoLiA.

- 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.

PICCL: further functionalities II

- 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.

PhilosTEI screenshot



PICCL: future availability

- Will have its own website soon.
  Currently at http://philostei.clarin.inl.nl
- Soon to be available in LaMachine.
  Cf. https://github.com/proycon/LaMachine
  - As a Virtual Machine - easiest, allows you to run our software on any host OS.
  - As a Docker application
  - As a compilation/installation script in a virtual environment

SoNaR Spaces

- A new web service that offers 9 flavours of semantic spaces built on the basis of SoNaR-500.
- Available from http://ticclops.uvt.nl/vector/

ENJOY!!

Thanks for your attention!

`http://philostei.clarin.inl.nl/`

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]