PICCL : Philosophical Integrator of Computational and Corpus Libraries

Martin Reynaert

Center for Language Studies - Radboud University Nijmegen / TiCC - Tilburg University

CLIN-25 2015 Universiteit Antwerpen, Belgium February 5-6, 2015

CLARIN-NL Call 4 project @PhilosTEI: logo

@PhilosTei

@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. We currently have dedicated TICCL web applications and services for 18 European languages / language varieties.

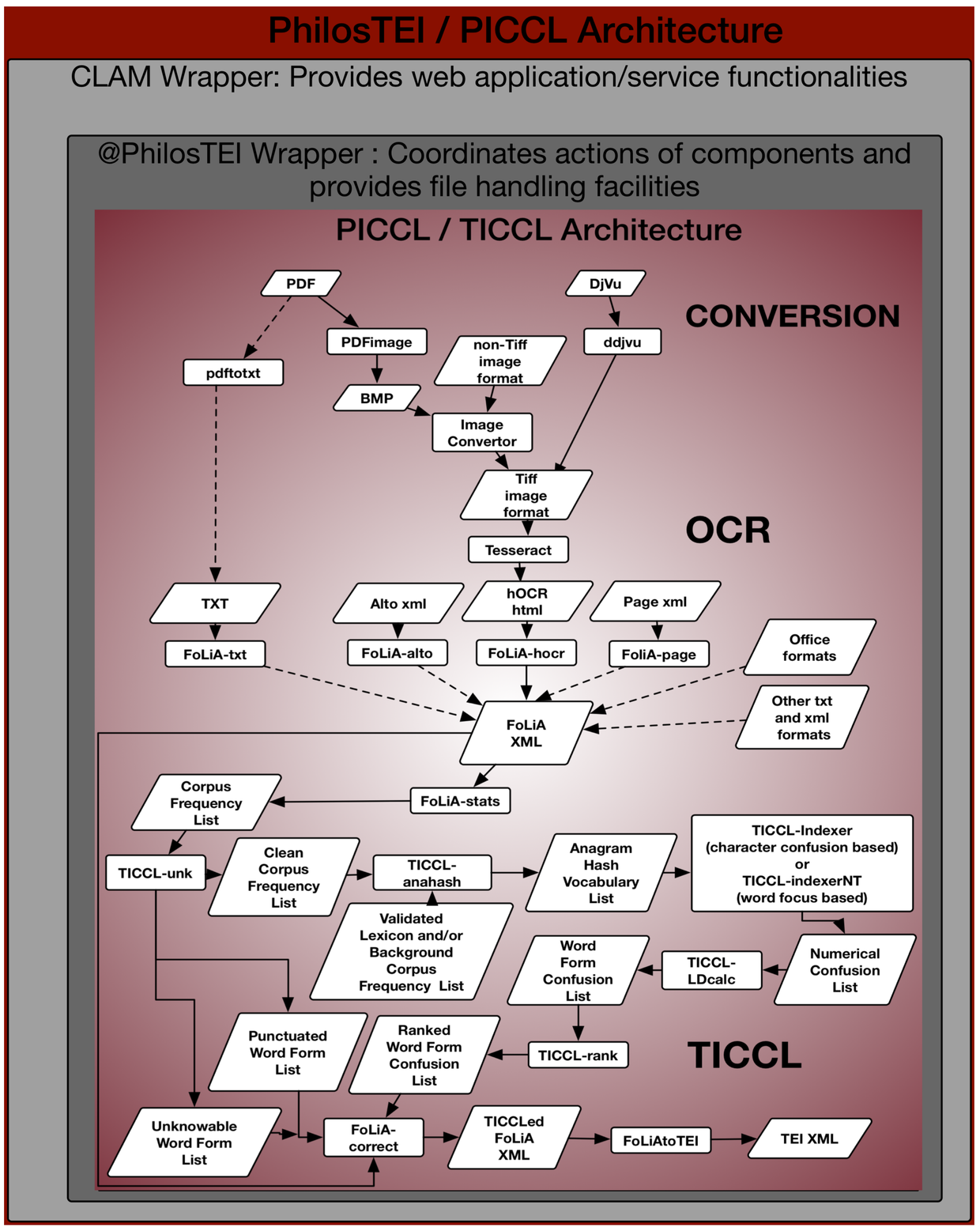
Main Work Flow Components for corpus building

- Conversion: a choice selection of available open-source image and text convertors have been incorporated in the work flow. The term 'philosophical' in the system's title should be understood to denote: 'well-considered'.
- Optical Character Recognition: Tesseract (of Google Books' fame) 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: 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.

Multilingual and diachronic 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, partner in Nederlab. They were deliverables of the European project Impact and are available through the Impact Centre of Competence (`http://www.digitisation.eu/`). We measured their effect on OCR post-correction and normalisation of the Nederlab corpora.
- In Nederlab the main challenge for TICCL is to be able to distinguish between historical and OCR spelling variants.
- That is, apart from the enormous sizes of the text collections to be fully automatically post-corrected. The Dutch National Library (KB) collection Early Dutch Books Online has over 10K Dutch books, about 1.7M pages of digitised text, representing about 435M word tokens. The Dutch Digitized Daily Newspapers for the years 1618 to 1899 alone represent about 35M articles.
- TICCL uses:
  - a large lexicon
  - exhaustive word variant look-up up to a given Levenshtein distance
  - a numerical list of Known Historical Character Confusions
  - a combination of corpus-induced ranking features to determine the most likely correction candidate

The PICCL Corpus Building Work Flow



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 is equal or more than x) image files to be sent to 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 1 single output FoLiA and final TEI xml output 'book'. By contrast, TTNWW passes a single file on from one web service to the next.
- The user-friendly system may 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 may 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

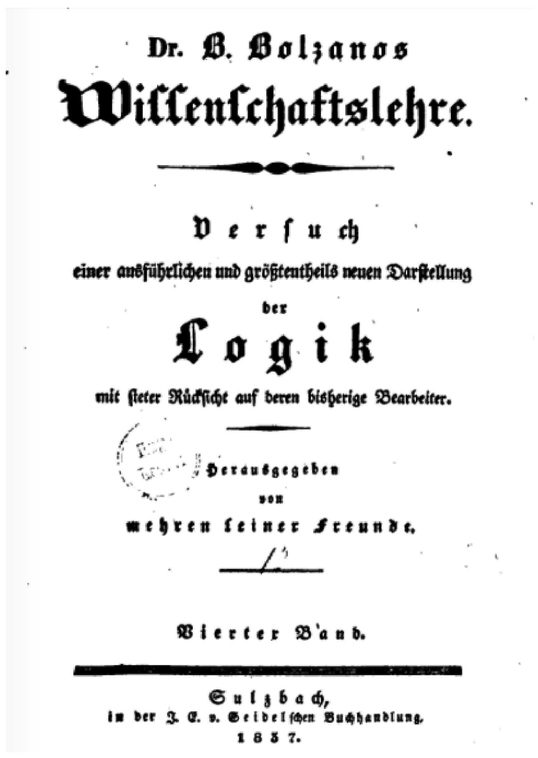

Dr. B. Bolzanos
Wissenschaftslehre.
Versuch
einer ausführlichen und größtentheils neuen Darstellung
der
Logik
mit steter Rücksicht auf deren bisherige Bearbeiter.
Herausgegeben
von
mehren seiner Freunde.
Vierter Band.
Sulzbach,
in der J. E. v. Seidelschen Buchhandlung.
1837.

Figure 1: Title page of 1837 book by Bolzano, the 4 tome demonstrator data for 'German Fraktur' in @PhilosTEI

Radboud Research Team

NWO Groot project Nederlab: logo

nederlab

Nederlab: contribution

- The Nederlab project aims to bring together all digitized texts relevant to the Dutch national heritage (c. A.D. 800 – present) consisting of terabytes of data in one user-friendly and tool-enriched web interface, allowing scholars to simultaneously search and analyze textual data in a virtual research environment.
- The focus in Nederlab is currently 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 Early Dutch Books Online or EDBO (`http://www.delpher.nl/`).
- These were digitized by means of OCR. If there is one thing all results of large digitization programmes have, it is that they are riddled with OCR misrecognition errors.
- These texts spanning four centuries present a wealth of diachronic spelling variation.

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 on the diachronic texts. The Dutch lemmatiser and POS-tagger FROG is due to be integrated in PICCL.

TICCL evaluation on Early Dutch Books Online (EDBO)

L	C	acc	prec	recall	f-score
10 best-first ranked					
1	A	91.92	99.77	61.01	75.71
1	B	93.23	99.48	66.92	80.02
2	B	93.42	99.39	69.22	81.60
3	B	94.50	99.46	72.77	84.05
4	B	95.97	99.81	77.27	87.11
best-first ranked					
4	B	94.51	99.79	70.98	82.96

Evaluation results on the task of fully automatically normalizing and OCR post-correcting as measured on the full DPO35 Gold Standard (Dutch book by Martinet, 1789). The results clearly show the effect of using different lexicons (L): the contemporary TICCL lexicon (1), the INL historical Dutch lexicon (2), the previous two combined (3), the combined lexicons further enhanced with the INL Historical names list (4).
We have measured on two corpora (C), i.e. the book DPO35 only (A), or the book as part of the 10,000 books EDBO collection (B). We first list results as measured on the 10 best first ranked CCs, then list the highest best-first ranked combination.

Acknowledgements

Martin Reynaert's work is supported by the Netherlands Organisation for Scientific Research (NWO), the Royal Netherlands Academy of Arts and Sciences (KNAW), and Common Language Resources and Technology Infrastructure (CLARIN-NL-12-006 project @PhilosTEI and CLARIN-NL-12-013 project OpenSoNaR).