PICCL & its Pilots

Martin Reynaert[31] **Maarten van Gompel**[2] **Ko van der Sloot**[2] **Antal van den Bosch**[12]
KNAW Meertens Institute, Amsterdam[1] CLS / Radboud University Nijmegen[2]
TSHD / Tilburg University[3]
The Netherlands
martin.reynaert|antal.van.den.bosch@meertens.knaw.nl
m.vangompel|k.vandersloot@let.ru.nl

Abstract

The corpus building work flow Philosophical Integrator of Computational and Corpus Libraries or PICCL, previewed in Reynaert et al. (2015), has now grown into a mature production system, part of the CLARIN research infrastructure. As such it is available to anyone with an academic logon as a single sign-on system offering the user her own private work space. The user may now freely upload her own texts in a wide range of formats and have them digitized by means of Optical Character Recognition through Tesseract[1] in case these are just images of printed pages. The electronic text, by then in PICCL's pivot format FoLiA XML (van Gompel et al., 2017), may at will be corrected for OCR-errors, normalized or modernized into today's Dutch by the state-of-the-art OCR post-correction system TICCL. The text can then be tokenized by Ucto[2] or, if Dutch, further linguistically enriched by Frog[3].

Available as a web application or service running on dedicated hardware at CLARIN Center INT for less data-intensive purposes, the system is freely available as part of the LaMachine distribution[4] designed to allow any research team to set up its own local installation in a hassle-free manner. Moreover, due to being implemented in the Nextflow[5] work flow environment, PICCL is now fully distributable, ready for any parallelization framework and capable of seamlessly running on your server, a grid platform, or in the cloud.

PICCL is currently being tested in three CLARIAH pilot projects we are directly involved in.

In pilot 'Names' (PI Gerrit Bloothooft, UU)[6] TICCL will be set to cluster both first and family names derived from the Dutch vital statistics. On the basis of character confusion and frequency statistics derived from large sets of manually linked name pairs, TICCL is expected to be able to assign as yet unpaired names to reasonable sets of base name clusters.

In both pilot 'DB:CCC' (PI Karin Hofmeester, IISH)[7] and pilot 'OpenGazam' (PI

[1] https://github.com/tesseract-ocr/tesseract
[2] https://languagemachines.github.io/ucto/
[3] https://languagemachines.github.io/frog/
[4] https://proycon.github.io/LaMachine/
[5] https://www.nextflow.io/
[6] https://www.clariah.nl/projecten/research-pilots/names
[7] https://www.clariah.nl/projecten/research-pilots/db-ccc/db-ccc

Rombert Stapel, IISH)[8] PICCL is called upon to digitize an encyclopedia, the first about diamond industry terminology, the second of American place names. Both call on PICCL to preserve the entries as separate entities, which calls for page segmentation capabilities. Both further present their own particular challenges: the first is in Dutch, but also multilingual in that it gives English, French and German translations for each term. The second is in English but printed in double column format, presenting a high incidence of split words. In the second project, PICCL is also set to recover geographical names from a 1625 Dutch book describing the Americas by De Laet, printed in Fraktur, the second edition of which we also have a Latin, an English and a Spanish digitized version for.

All demos given will be based on the actual pilot project data. Pilot project results are to be made available as linked open data.

References

[Reynaert et al.2015] Martin Reynaert, Maarten van Gompel, Ko van der Sloot, and Antal van den Bosch. 2015. PICCL: Philosophical Integrator of Computational and Corpus Libraries. In ***Proceedings of CLARIN Annual Conference 2015 – Book of Abstracts***, pages 75–79, Wrocław, Poland. CLARIN ERIC.

[van Gompel et al.2017] Maarten van Gompel, Ko van der Sloot, Martin Reynaert, and Antal van Den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In J. Odijk and A. van Hessen, editors, ***CLARIN-NL in the Low Countries***, chapter 6, pages 71–81. Ubiquity.

[8] `https://www.clariah.nl/projecten/research-pilots/opengazam`