Synergy of @PhilosTEI and OpenSoNaR: Towards PICCL or Goodbye CLARIN-NL, Welcome CLARIAH!!

Martin Reynaert

Tilburg center for Cognition and Communication - Tilburg University
Centre for Language and Speech Technology - Radboud Universiteit Nijmegen

Radboud Universiteit Nijmegen. March 11, 2015

@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.

@PhilosTEI Team

- @PhilosTEI User Team:
  - Arianna Betti (Project Leader)
  - Hein van den Berg
  - Pam Rossel
- @PhilosTEI Technical Team:
  - Martin Reynaert
  - Ko van der Sloot
- De Taalmonsters:
  - Matje van de Camp (@PhilosTEI interface developer)

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 (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 (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.

@PhilosTEI in the CLARIN-NL Infrastructure

http://philostei.clarin.inl.nl

SoNaR

- The Dutch Language Union (Nederlandse Taalunie) funded STEVIN programme delivered SoNaR: a major new reference corpus for contemporary written Dutch from both the Netherlands and Flanders
- Covers old and new media, contains over 540M tokens
- Over 2.1M text files and 2.1M metadata files
- Situation: SoNaR corpus too big for less technically inclined researchers to handle

CLARIN-NL Call 4 project OpenSoNaR

- Teamed up with INL who had been developing a corpus exploration back-end system: BlackLab
- Proposed the OpenSoNaR project to CLARIN-NL
- Developed WhiteLab: user-tailored interface to the online SoNaR-500 and SoNaR New Media corpora

WhiteLab development: Tilburg University

- WhiteLab Team:
  - Martin Reynaert
  - Ko van der Sloot
  - Menno van Zaanen
- De Taalmonsters:
  - Matje van de Camp (WhiteLab developer)
- Project Coordination:
  - Max Louwerse

OpenSoNaR User Groups

- User Groups
  - Linguistics: TiCC, Tilburg University: Ad Backus, Jan Renkema
  - Media and Communications: TiCC: Emiel Krahmer, Fons Maes
  - Literary Sciences: Huygens-ING: Karina van Dalen-Oskamp
  - Cultural Sciences: Meertens Institute: Nicoline van der Sijs

OpenSoNaR in the CLARIN-NL Infrastructure

http://opensonar.clarin.inl.nl

ENJOY!!

Thanks for your attention!

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

Synergy of @PhilosTEI and OpenSoNaR: Towards PICCL
or
Goodbye CLARIN-NL, Welcome CLARIAH!!

Martin Reynaert