CLARIN-NL: OpenSoNaR-CGN

Martin Reynaert (Tilburg University / Radboud University), Matje van de Camp (De Taalmonsters), Henk van den Heuvel and Nelleke Oostdijk

TiCC - Tilburg University
CLST - Radboud Universiteit Nijmegen

Workshop: Morfosyntactisch verrijken van historische teksten, Utrecht. November 16, 2015

Corpus Geschreven Nederlands: de TST-blurb

- Het SoNaR-corpus is een tekstcorpus dat bestaat uit twee delen, nl. SoNaR-500 en SoNaR-1.
- SONAR-500 bevat meer dan 500 miljoen woorden tekst afkomstig uit uiteenlopende domeinen en genres. Alle teksten werden getokeniseerd, ge-POS-tagd en gelemmatiseerd. Ook de named entities werden gelabeld. Alle annotaties van SoNaR-500 werden automatisch geproduceerd.
- De nieuwemediateksten (tweets, chats en sms'en), die ook verzameld werden in het kader van het STEVIN-project SoNaR maken geen deel uit van het SoNaR-corpus 1.0. en zijn apart als SoNaR Nieuwe Media Corpus beschikbaar.


openSoNaR

OpenSoNaR: Consortium

- TiCC & Huygens ING & Meertens Institute: Users
- TiCC & INL: Technology Providers
  - Martin Reynaert & Menno van Zaanen & Katrien Depuydt & Jesse de Does
  - Developers: De Taalmonsters (Matje van de Camp and partners) and Jan Niestadt
  - Scientific programmer: Ko van der Sloot
- INL (Institute for Dutch Lexicology) - Leiden: Infrastructure Specialist and CLARIN Centre
  - Katrien Depuydt

Intentions of OpenSoNaR

- Address the corpus exploration and exploitation problem
- Develop and make available an open corpus exploration and exploitation environment
- Tailor the front-ends to the desiderata of 4 CLARIN-NL priority groups of users

Means employed by OpenSoNaR

- Intended originally to build on the well-known Corpus Workbench back-end
- Turned out INL had been developing a new alternative
- Java system BlackLab, based on Apache Lucene, open source via GitHub
- OpenSoNaR built Whitelab front-ends according to the user groups specifications, open source via GitHub

OpenSoNaR User Groups in more detail

- Literary Sciences: Huygens ING, Karina van Dalen-Oskam
- Cultural Sciences: Meertens Institute, Nicoline van der Sijs, assisted by Ewoud Sanders
- Linguistics: TiCC: Jan Renkema & Ad Backus & colleagues
- Communication and Media Studies: Fons Maes & Emiel Krahmer & colleagues

OpenSoNaR online

- URL: http://opensonar.inl.nl

OpenSoNaR-CGN

- New CLARIN-NL project run by Radboud University (Nelleke Oostdijk and Henk van den Heuvel)
- Allows for complete overhaul of BlackLab and WhiteLab tandem by De Taalmonsters
- Available spring 2016

Corpus Gesproken Nederlands: de TST-blurb

- Het Corpus Gesproken Nederlands (CGN) is een verzameling van 900 uur (bijna 9 miljoen woorden) hedendaags Nederlandse spraak, afkomstig van Vlamingen en Nederlanders. De spraakfragmenten (spontaan en voorbereid) zijn opgelijnd met diverse transcripties (o.a. orthografisch, fonetisch) en annotaties (syntactisch, POS-tags). Metadata, lexica, frequentielijsten en de corpusexploratiesoftware Corex behoren ook tot het CGN.

OpenSoNaR-CGN online

- WhiteLab versie 2 brengt SoNaR-500 samen met CGN online
- Doorgedreven integratie van de gezamelijke doorzoekbaarheid van beide corpora
- Biedt ook toegang tot de spraakfragmenten

Thanks!!

Thanks for your attention!

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

CLARIN-NL: OpenSoNaR-CGN

Martin Reynaert (Tilburg University / Radboud University), Matje van de Camp (De Taalmonsters), Henk van den Heuvel and Nelleke Oostdijk

TiCC - Tilburg University
CLST - Radboud Universiteit Nijmegen