

Stevin Nederlandstalig Referentiecorpus *or* SoNaR project

Consortium:
RU, UvT, HoGent, UTwente
(KU Leuven, RUU)

SoNaR project: aim

Aim:

to construct a 500 MW corpus of written Dutch for use in different types of linguistic and HLT research and the development of applications

Corpus is to include

- contemporary (post-1954), standard written Dutch texts
- texts originating from the Dutch speaking language area in Flanders and the Netherlands as well as Dutch translations published in and targeted at this area
- native speaker language and the language of (professional) translators
- conventional genres and texts from new media

SoNaR project: tasks

WP A: Corpus building

- Acquisition and IPR
- Conversion to common XML format
- Tokenization
- Orthographic annotation
- Morphological annotation, lemmatization and POS tagging
- Consolidation and storage of corpus data

WP B: Semantic annotation

WP C: Quality control

SoNaR project: tasks

WP A: Corpus building

WP B: Semantic annotation (1 MW Lassy data)

- Named entity identification and classification
- Annotation of co-reference relations (cf. COREA)
- Annotation of spatial and temporal relations (cf. D-Coi)
- Annotation of semantic roles (cf. D-Coi)

WP C: Quality control

- Pre-validation
- Safeguarding the quality of the end product
- Monitoring the external validation

SoNaR project: phase 1

January 2008 until December 2008

Budget € 99,500

Dedicated entirely to WP A, excl. morph. annotation, lemmatization and POS tagging

Targets:

- Acquisition of and IPR clearance for min. 50 MW of data from Flanders and 25 MW from the Netherlands; another 30 MW from TWNC
- Development of procedures for acquiring texts from the Internet
- Conversion to XML format
- Correction of run-on and split words