VU-DNC

CLARIN-NL data curation project

Wilbert Spooren, Martin Reynaert, Kirsten Vis

CLARIN-NL Kickoff meeting Call 2

February 9, 2011
Overview

• VU-DNC resource
• VU-DNC project
Resource

• VU-DNC = VU University Diachronic Newspaper Corpus

• Newspaper texts
  – 1950/1: 3615 texts; 931,574 words
  – 2002: 3003 texts; 971,059 words
  – 5 major Dutch national newspapers
  – 8 sections

• Origin
  – 1950/1: OCR from paper copies
    Scanning, OCR, post-OCR-corrections (semi-automatic/manual), labelling
  – 2002: from digital news database (LexisNexis)
Resource (2)

• Annotation
  – Part of speech (Tadpole; ILK)
  – Lemma (ibid.)
  – Subjectivity: modality, 1st and 2nd p pronouns, intensifiers, etc
  – Direct quotations

• Format
  – XML

• Metadata
  – Newspaper
  – Page
  – Section
  – Publication date
Project

• Data curation project

• Participants
  VU University – Amsterdam
    Wilbert Spooren, Kirsten Vis
  Tilburg centre for Creative Computing (TiCC) – Tilburg
    Martin Reynaert, Maarten van Gompel
  Institute for Dutch Lexicology (INL) – Leiden
    Remco van Veenendaal, Laura van Eerten

• 1 April – 1 December 2011
Project (2)

• Goals

1. Make diachronic corpus available
2. Make the linguistic annotation of subjectivity and quotation available
3. Create a gold standard benchmark for testing and training of OCR-postcorrection tools
4. Extend ISOcat categories of CLARIN
Goal 1: Make diachronic corpus available

• in line with STEVIN Nederlandstalig Referentiecorpus (SoNaR)
  – extend SoNaR format
    • FoLiA (Format for Linguistic Annotation)
  – conversion

• make data and documentation available at CLARIN Centre INL (TST Centrale)

• IPR arranged: license agreements with newspapers
Goal 2: Make linguistic annotation available

- linguistic annotation
  - subjectivity (modality, etc)
  - direct quotations (marked by quotation marks)

- will be made available with corpus
  - including documentation
Goal 3: Create gold standard benchmark

• 1950/1 component obtained by OCR, followed by semi-automatic corrections

• Pre- and post-correction files
  – benchmark set for OCR-postcorrection tools
  – older texts (spelling)

• Procedure
  – align pre- and postcorrection versions at word level
Goal 4: Extend ISOcat categories of CLARIN

• Create CMDI Metadata for the curated resource

• Map data categories (part of speech) to ISOcat data categories

• Extend ISOcat for encoding of subjectivity and quotations

• Report on requirements and desiderata for CLARIN infrastructure
Thank you for your attention!