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 **D**iachronic **N**ewspaper **C**orpus

- 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

de Volkskrant

KATHOLIEK DAGBLAD VOOR NEDERLAND

Staatkundig Hoofdredacteur Prof. mr C. P. M. Romme • Algemeen Hoofdredacteur J. M. Lücker • Directeur J. Kolkman

WOENSDAG 29 AUGUSTUS 1951

29ste Jaargang No. 8138 ZES PAGINA'S

Nieuwe Zijds Voorburgwal 345, Amsterdam — Telefoon directie en redactie 64633 — Administratie (Dam 4) 40739 en 42793 — Giro 46006 — Abonnement f 5,45 per kwartaal, 42 cent per week — 8 cent per nummer

PLUK DE DAG
als het fruit rijp is
Pagina 4

PLAN voor ENORME VOLKSVERHUIZING | Orkaan verwoestte stad op Jamaica | Overleg-Soepomo

PLAN voor ENORME VOLKSVERHUIZING

Drie millioen Aziaten naar Oost-Europa

TRANSPORT UITGEVOERD PER TREIN EN PER BOOT

MUNCHEN, 28 Aug. (K.N.P.) — Onder leiding van Rusland en met nauwe medewerking van China is een plan ontworpen om in de jaren 1952 en 1953 rond drie millioen Aziaten in Midden- en Oost-Europa te vestigen. In de jaren 1949 en 1950 zijn reeds 650.000 arbeiders en boeren uit Azië naar Europa overgebracht. Uit de zojuist gepubliceerde statistieken van Hongarije en Roemenië blijkt, dat een kwart millioen van hen in genoemde landen te werk zijn gesteld. Ook zijn vele Mongolen overgebracht naar het door Rusland geannexeerde

Ridgway verwerpt rood voorstel

TOKIO, 29 Aug. (A.P.) — Generaal Ridgway heeft vierkant geweigerd in te gaan op het communistische voorstel om een nieuw onderzoek te doen instellen naar het incident van Kaesong. De geallieerde bevelhebber verklaarde, dat een verder onderzoek, zoals de communisten dat eisen, „alleen maar een niet te rechtvaardigen uitstel van de wapenstilstandsonderhandelingen ten doel kan hebben". Dit was Ridgway's antwoord op de boodschap van Kim Il soeng en Peng te hoogi-

Kamerverbouwing kostte f 691.700

(Van onze correspondent)

DEN HAAG, 28 Aug. — De verbouwing van het Tweede-Kamergebouw heeft tot dusverre een bedrag gevergd van 691.700 gulden. Aanvankelijk was een som begroot van 300.000 gulden, die reeds eerder moest worden verhoogd met 275.000 gulden. Bij een wetsontwerp tot wijziging van de begroting 1950 van de Hoge Colleges van Staat worden nu de resterende 116.700 gulden aangevraagd. Ook dit jaar zullen echter nog enige credieten moeten worden verstrekt, aldus de toelichting.

Donderdag uitspraak in zaak Bertha Hertogh

SINGAPORE, 28 Aug. (A.P.) — Het Hoge Hof van Appel in Singapore zal Donderdag uitspraak doen in de zaak Bertha Hertogh. De hele politiemacht van Singapore is voor Donderdag gealarmeerd om een herhaling van de relletjes van December te voorkomen. Bovendien worden Britse troepen in gereedheid gehouden, om de politie te assisteren in geval van ordeverstoringen.

EISENHOWER:

„Maak van Europa federale unie"

WASHINGTON, 28 Aug. (A.P.) — „Ik ben mij bewust, dat een aantal van mijn medewerkers denken, dat ik volslagen gek ben, maar ik zeg u, dat een „Verenigd Europa" de sleutel tot de gehele zaak is. En als om dit te bereiken een Europees leger van belang is, ben ik bereid daar veel werk aan te besteden. Persoonlijk hoop ik, dat tal van problemen in West-Europa zullen verdwijnen, als Europa in een federale unie ondergebracht is. Ik ben hiervan zo diep overtuigd, dat ik niet geloof, dat men in Amerika, in Engeland en andere landen een werkelijk veilig gevoel zal hebben, zo lang dit niet gebeurd is."

Dit verklaarde generaal Eisenhower aan een groep Amerikaanse senatoren, die een bezoek hebben gebracht aan Europa en vandaag verslag uitbrachten in Washington. Generaal Eisenhower en diens stafchef, generaal Gruenther hebben de senatoren er met nadruk op gewezen, dat de Amerikaanse steun aan West-Europa in de vorm van wapens en materieel niet groot genoeg kan zijn. Over het militaire beeld in Nederland zei generaal Gruenther: „Nederland gaat methodisch te werk bij het opbouwen van zijn leger. Op het ogenblik zijn belangrijke legeronderdelen in wording en onmiddellijk mobiliseerbaar."

Lefschopper veroordeeld

(Van onze correspondent)

HAARLEM, 27 Aug. — De 19-jarige P. H. S. had onlangs binnen een half uur met vrienden 9 borrels gedronken en wilde toen met een andere knaap achterop door Haarlem fietsen. Dat ging natuurlijk niet al te best. Hij zigzagde over de weg, totdat de politie een eind aan deze rit maakte. De Officier van Justitie eiste — ondanks het blanco strafregister van de knaap — 50 gulden of een maand gevangenis. De rechter hield rekening met huiselijke omstandigheden en halveerde deze boete.

Thank you for your attention!

Toeschouwers dachten aan grap

KASTELEIN DOOR ROVER DOODGESCHOTEN

Daders vluchtten zonder buit

(Van onze correspondent)

ROTTERDAM, 28 Aug. — In

moeilijkheden riep de man met het pistool: „Let op, ik schiet" en nadat zijn makker zich met enkele vlugge stappen uit de vuurlinie had teruggetrokken, viel inderdaad een schot. De kogel trof het slachtoffer van een afstand van ongeveer acht meter in het gelaat en ging hem dwars door het hoofd. De overvallers sloegen onmiddellijk op de vlucht en verdwenen op de fiets.