

Resultaten en Beschikbaarheid 1e Ronde projecten, SPRAAK en CORNETTO

Jan Odijk i.s.m. TST-Centrale



STEVIN Programmabijeenkomst, Hoeven, 11 september 2008

Projecten

- Autonomata
- COREA
- D-Coi
- IRME
- JASMIN-CGN
- SPRAAK
- CORNETTO

Projectafronding

- Voor algemene procedure, zie Hans Kruithof later vandaag
- Portefeuillehouders beoordelen resultaten en stellen advies op
- PC adviseert op basis hiervan het bestuur
- Target officiële afronding 1e ronde projecten+SPRAAK: 16 oktober 2008

Projectresultaten

- De meeste projecten
 - Onderzoeksresultaten
 - Vastgelegd in rapporten en artikelen
 - Taalbronresultaten
 - Data
 - Tools
 - Protocollen
 - Documentatie
- Focus hier op taalbronresultaten

Licenties

- 'Niet-commercieel'
 - d.w.z. niet-commercieel gebruik in instellingen zonder winstoogmerk
 - licenties gratis
- Commercieel gebruik:
 - licenties marktconform, meestal niet gratis
 - meestal 'lumpsum' model
 - soms 'royalty' model
- Voor allen: soms distributiekosten

Automata

- Automata for deriving phoneme transcriptions of Dutch and Flemish names
- Resultaten op hoofdlijnen
 - G2p, p2p's en (training)tools
 - Namencorpus (manueel geverifieerd)
 - Rapporten en documentatie

Autonomata - beschikbaarheid

- Resultaten overhandigd aan de TST-Centrale in augustus 2007
- Overdracht nog te bekrachtigen
- Software
 - Niet-commercieel: geen licentiekosten
 - Commercieel: via royaltymodel
 - Vaak licentie op Nuance G2P nodig (voor iedereen)
- Beschikbaar via TST-Centrale na formele projectafroding

COREA

- Coreference Resolution for Extracting Answers
- Resultaten op hoofdlijnen
 - Annotatieprotocol en -software
 - Corpus
 - 50% meer dan gepland
 - visualisatietools
 - Software voor coreferentieresolutie
 - Rapporten en documentatie

COREA - beschikbaarheid

- Resultaten opgeleverd eind 2007
- Coreferentiesoftware online
 - Als demo
 - File-upload service (nu nog alleen intern)
- Overleg over overdracht in finale fase
- Niet-commercieel: gratis licentie
- Commercieel: Bescheiden kost voor het corpus
 - Bevat Medische Winkler Prins
 - Spectrum vraagt hier een vergoeding voor
- Beschikbaar via TST-Centrale na formele projectafroding

D-Coi

- Dutch Language Corpus Initiative
- Resultaten op hoofdlijnen
 - Rapporten over
 - corpus design, sampling en metadata;
 - Basisformaten en validatietools
 - Corpus-induced Corpus Clean-up
 - Protocollen + tools voor
 - POS-tagging en lemmatisering
 - Syntactische annotatie
 - Semantische annotatie (verkenkend; incl. data)
 - Pilot corpus (50+MW) incl. documentatie
 - Automatische POS-tagging, lemmatisering, parsing
 - Gedeeltelijk handmatig gevalideerd (500k/500k/200k)
 - COREX: extensie van exploitatiesoftware en documentatie

D-Coi - beschikbaarheid

- Resultaten begin 2007 in STEVIN-wiki
- Corpus en COREX in juni 2008 naar TST-Centrale
- Laatste activiteiten voor overdracht lopen (IPR, data/software)
- Beschikbaar via TST-Centrale na formele projectafroonding

IRME

- Identification and Representation of Multiword Expressions
- Resultaten op hoofdlijnen (eindrapport)
 - Meerwoordlexicon (DuELME)
 - Met meer informatie dan in oorspronkelijke plan
 - Gevalideerd door CST
 - DuELME GUI (extra)
 - Rapporten, documentatie en tools

IRME - beschikbaarheid

- Resultaten eind 2007 in STEVIN-wiki
- Overleg over overdracht afgerond
- Beschikbaar via TST-Centrale na formele projectafroding

JASMIN-CGN

- Jongeren, Anderstaligen, Senioren en Machine Interactie voor het Nederlands
- Resultaten op hoofdlijnen
 - Spraakcorpus
 - Opnames, orthografische transcriptie, automatische fonetische transcriptie, uitspraaklexicon, POS-tagging en lemmatisering
 - Gevalideerd door BAS
 - Opnameplatform
 - Protocollen en documentatie

JASMIN-CGN - beschikbaarheid

- Resultaten in augustus 2008 naar TST-Centrale en STEVIN-wiki
- Overleg over overdracht in finale fase
- Beschikbaar via TST-Centrale na formele projectafroeding

SPRAAK

- Speech Processing, Recognition & Automatic Annotation Kit (SPRAAK)
- Resultaten op hoofdlijnen
 - Modulair opgebouwde state-of-the-art spraakherkenner voor het Nederlands
 - Demoherkenners (voor 'proof-of-concept')
 - Documentatie en rapporten

SPRAAK - beschikbaarheid

- Stand van zaken
 - SPRAAK(-broncode) via K.U. Leuven
 - SPRAAK-fund waarborgt beheer en verdere (door)ontwikkeling
 - Overleg over overdracht afgerond
 - Na formele afronding beschikbaar
- Niet-commercieel: gratis
- Commercieel: “bescheiden kost”

CORNETTO

- Combinatorial and Relational Network as Toolkit for Dutch Language Technology
- Resultaten op hoofdlijnen
 - Lexicaal-semantische databank voor het Nederlands
 - Rapporten, documentatie en tools

CORNETTO - beschikbaarheid

- Tussenresultaten beschikbaar gesteld
- Licenties hiervoor reeds verkrijgbaar
- Na formele afronding volledige CORNETTO beschikbaar
- Niet-commercieel: gratis licentie
- Commercieel: bescheiden kost
 - Bevat Van Dale data (uit Nederlandse WordNet)
 - Van Dale vraagt hier een vergoeding voor

Conclusies

- De genoemde projecten hebben veel interessante resultaten opgeleverd
 - → bieden vele nieuwe mogelijkheden voor onderzoek en ontwikkeling
- De meeste resultaten zijn af
- De overdracht is in een vergevorderd stadium en in veel gevallen afgerond
- Binnenkort zijn alle resultaten verkrijgbaar via de TST-Centrale