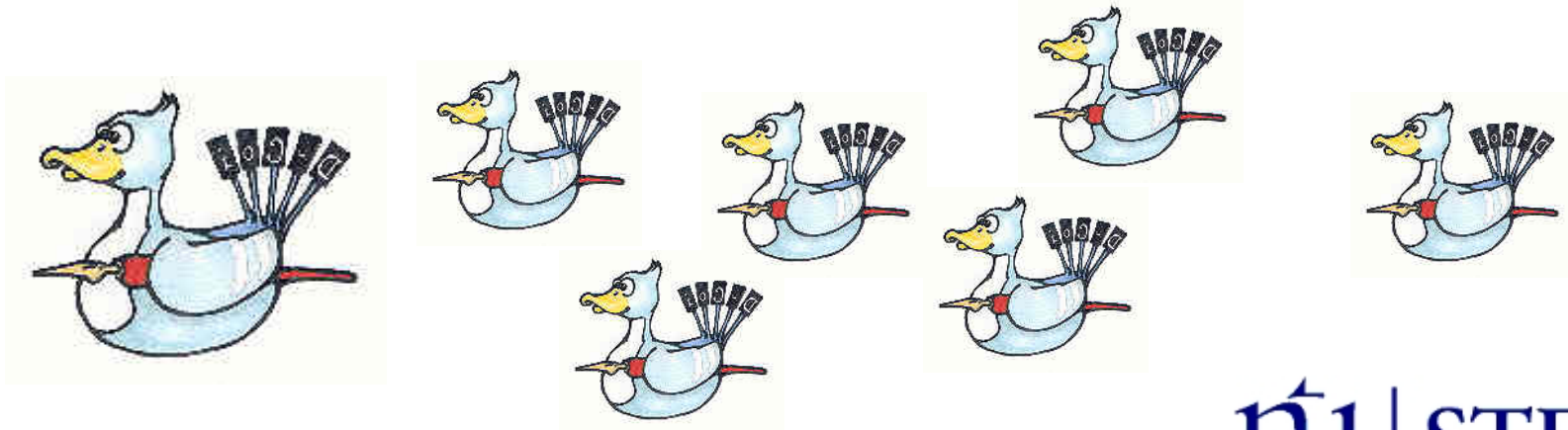


Dutch Language Corpus Initiative (D-Coi)

STEVIN-programmadag
11 september 2006



Overzicht

- Doelstellingen
- Stand van zaken
- Evaluatie
- Disseminatie
- Opvolging

Doelstellingen

- Design van een 500 Mwoord corpus geschreven Nederlands en ontwikkeling van protocollen, procedures en tools benodigd voor de aanleg en annotatie van de data
- Aanleg 50 Mwoord pilot corpus

Stand van zaken [1]

Corpus design (incl. metadata) en sampling

<ul style="list-style-type: none">• 50 Mwoord pilot corpus<ul style="list-style-type: none">- design- sampling- documentatie	<p>√ ! ±</p>	<p>» IPR problematiek</p>
<ul style="list-style-type: none">• 500 Mwoord corpus<ul style="list-style-type: none">- design- user requirements study- consultatie focusgroepen- finalisering design; docum.	<p>√ √</p>	<p>» zie D-Coi website</p>

Stand van zaken [2]

Conversie en markup

• definitie basisformaat (XML)	√	
• conversie bronbestanden	!	» doc/html/pdf arbeidsintensiever dan voorzien
• validatietool	√	
• documentatie	±	» zie D-Coi website

Stand van zaken [3]

Regularisatie

• tokenisatie en opsplitsing in zinnen	√	
• autom. spellingcorrectie	√	
• correctie confusibles	±	
• documentatie	±	» zie D-Coi website

Stand van zaken [4]

POS tagging en lemmatisering

• adaptatie CGN-tagset + man.	√	» zie D-Coi website
• hertrainen tagger/lemmatiser	√	» nog eens op basis van geverifieerd materiaal
• verificatie 500.000 woorden	√	
• tagging en lemmatisering volledige pilot corpus		
• evaluatie en documentatie	±	» tagger accuracy: 97,87%

Stand van zaken [5]

Syntactische annotatie

• adaptatie Alpino parser	√	» zie D-Coi website
• verificatie 200.000 woorden	√	
• autom. syntactische ann. volledige pilot corpus		
• evaluatie en documentatie	±	» zie D-Coi website

Stand van zaken [6]

Semantische annotatie

- | | | |
|--|---|---------------------|
| • pilot study 1: temporele en spatiale semantiek | | |
| - literatuurstudie | ✓ | |
| - ontwikkeling protocol | ✓ | |
| - annotatie 3000 woorden | ✓ | |
| • pilot study 2: sem. rollen | | » zie D-Coi website |
| - literatuurstudie | ✓ | |
| - ontwikkeling protocol | ✓ | |
| - annotatie 3000 woorden | ✓ | |

Stand van zaken [7]

Corpus exploitatie

- adaptatie COREX software
- pilot corpus toegankelijk m.b.v. COREX
- documentatie

±

!

±

» extra inspanning nodig
t.b.v. integratie manueel
geverifieerde bestanden

Evaluatie

- **Intern** (deels in samenspraak met gebruikersgroep):
 - testen van protocollen, procedures en tools
 - in kaart brengen van do's en don'ts
 - haalbaarheid vaststellen (nieuwe) annotaties
 - kwaliteitsbewaking
- **Extern (CST, Kopenhagen)**
 - evaluatie van de projectresultaten als zodanig
 - advisering t.a.v. gewenste aanpassingen/bijstellingen t.b.v. toekomstige projecten

Disseminatie

- Uiteindelijk projectresultaten worden overgedragen aan NTU/TST-Centrale
- Informatie over project, incl. reeds beschikbare (tussen-)resultaten:

<http://lands.let.ru.nl/projects/d-coi>

Opvolging

- Gebruik van D-Coi resultaten
 - in andere STEVIN-projecten, o.a. Corea, DPC, Lassy
 - in diverse ander projecten, waaronder
 - CONLL shared task <<<http://nextens.uvt.nl/~conll>>>
 - METIS II
 - NWO VICI project 'Implicit Linguistics'
- Aanleg van een Corpus Geschreven Nederlands
 - Call for tender