

DPC Dutch

Parallel

Corpus

Corpus Design

Lidia Rura

Geschiedenis corpora

- Engelstalige corpora (Brown 1961, LOB 1961)
- Nederlandstalige corpora (Eindhoven Corpus of Corpus Uit den Boogaart 1960-1973)
- EN referentiec corpora voor een taalgebied: BNC (1991-1994), ANC (vanaf 1999 (2003 eerste versie beschikbaar))
- NL referentiec corpora voor een taalgebied: CGN (1998-2004), Corpus Geschreven Nederlands (D-coi: haalbaarheidsstudie afgerond, oproep uitgeschreven)

Verschillende types corpora

Eéntalige corpora

- Referentiecorpora voor een taalgebied (CGN, BNC, ANC, ARTFL/FRANTEXT)
- Eéntalige (geen referentie-) corpora (INL-corpora)
- Corpora bestaande uit vertaalde teksten (TEC)
- Gespecialiseerde corpora (bepaald teksttype, vakgebied)

• Meertalige corpora

- Parallele corpora (vertalingen)
- Vergelijkbare corpora (gelijkaardige teksten)

Globalisering

- Dominerende positie van het Engels als wereldtaal wordt sterker
- Kleinere talen bouwen steeds meer achterstand op
- Nederlandse Taalunie \Rightarrow STEVIN + TST-centrale (taaltechnologische voorzieningen voor NL)

Geschiedenis van parallelle corpora

- **Grote corpora:**
 - Ontstaan binnen internationale instellingen
 - EN meestal automatisch kerntaal
 - Omvatten veel talencombinaties
- **Kleinere corpora:**
 - Ontstaan binnen een project als inhaalbeweging van andere talen dan EN
 - Omvatten meestal een paar talencombinaties

Grote internationale corpora

- EUROPARL
 - JRC-Acquis
 - OPUS
 - UN-Parallel Text
 - Hansard corpus
 - BAF corpus
- (NL geen brontaal)
 - (zonder NL)

Kleinere lokale corpora

- Zweeds-Turks
- MULINCO (Deens, Engels, Frans, Duits, Italiaans, Spaans)
- COMPARA (Engels-Portugees)
- Engels-Noors
- Engels-Zweeds
- Engels-Sloveens (ACQUIS)
- Spaans-Engels-Arabisch (UN)

Beoogde doelstellingen grote vs kleine corpora

- (Ver)taalkundig onderzoek, onderwijs, contrastieve linguïstiek
- Taaltechnologie: terminologie-extractie, automatische vertaling, CAT etc

Dutch Parallel Corpus

- Talen NL-EN-FR, NL kerntaal
- Nog geen corpora beschikbaar (vrij consulteerbaar) met NL als kerntaal (bv. Namur Corpus (1999))
- STEVIN: DPC moet beschikbaar zijn voor commerciële en niet-commerciële doeleinden

Design van andere parallelle corpora

- fictie vs non-fictie
- beoogd gebruik : (ver)taalkunde vs taaltechnologische toepassingen en terminologie
- design criteria: beschikbaarheid vs relevantie
- parallel vs vergelijkbaar

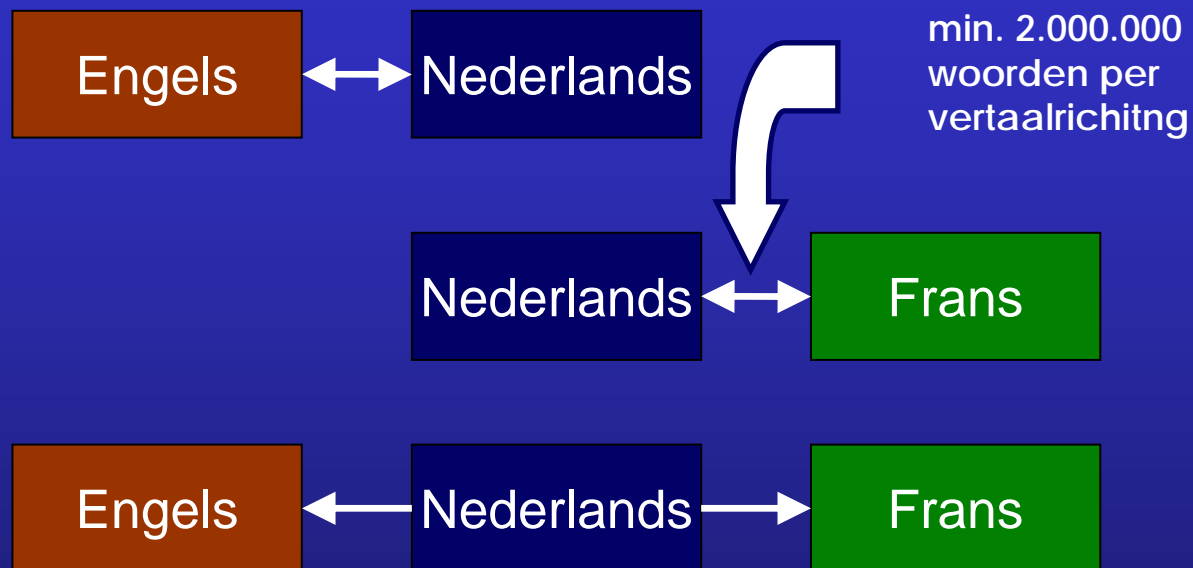
DPC – een ambitieus project met een uitgedacht design

- Kwalitatief hoogstaand
- Groot qua omvang
- Gevarieerd qua teksttypes
- NL als kerntaal met een evenwichtige verhouding tussen de talen
- Multifunctioneel: (ver)taalkundig onderzoek, CALL, CAT, taaltechnologie
- Beschikbaar voor gebruik in onderwijs, onderzoek en voor de ontwikkeling van toepassingen

Belangrijke richtlijnen

- Relevantie materiaal belangrijker dan beschikbaarheid
- 10 miljoen woorden
- NL als bron- en doeltaal
- Evenwichtig: min. 2 miljoen woorden per vertaalrichting

Samenstelling van het corpus



Mogelijke problemen met de samenstelling van het corpus

- Beschikbaarheid van tekstmateriaal van toereikende kwaliteit in alle vertaalrichtingen
- Brontaal en vertaalwijze achterhalen (gevallen van indirecte vertaling)
- IPR om het corpus openbaar te mogen maken

Design criteria DPC

- Taal (vertaalrichting)
- Teksttype: fictie en non-fictie: essayistisch, journalistiek, zakelijk, technisch, ambtelijk
- gepubliceerd bij voorkeur
- Tekstgrootte

Gebruikersenquête als toetsmiddel voor het design

- Achtergrond gebruiker: gevolgen voor designvereisten
- Belangstelling gebruiker voor (parallele) corpora
- Door de informant gebruikte corpora als indicatie voor een gewenst design
- Beoogd gebruik als indicatie voor een gewenst design
- Gebruikersverwachtingen ten opzichte van het DPC

Dank u voor uw aandacht!

- Vragen en suggesties
- U kunt de gebruikersenquête DPC invullen op het adres: <http://www.kuleuven-kortrijk.be/DPC/data/enquete-v2/enquete-dpc/>
- Bijkomende info over bestaande en beschikbare corpora op hetzelfde adres