

# Dutch Parallel Corpus

A Multifunctional & Multilingual Corpus

Dutch - English

Dutch - French

[www.kuleuven-kortrijk.be/dpc](http://www.kuleuven-kortrijk.be/dpc)

# Doelstelling

- Parallel corpus
- 10 miljoen woorden
- Nederlands - Engels
- Nederlands - Frans
- Kwalitatief hoogstaand
- Compatibiliteit met Corpus Geschreven Nederlands

# Corpusopbouw

- 10 miljoen woorden
- 2 taalparen, 4 vertaalrichtingen
- Min. 2 miljoen woorden / vertaalrichting
- Gedeeltelijk drietalig

Nederlandse teksten vertaald in Engels en Frans

# Corpussamenstelling

- Kwaliteitsvolle teksten
  - Voorkeur voor gepubliceerd materiaal
- Verschillende domeinen
  - Contacten met tekstleveranciers:  
Lannoo, Roularta, NLPVF, IBM, KBC,  
Belgische federale overheid
- IPR-overeenkomsten (copyright)

# Verwerking

- Tekstnormalisatie
  - XML-files in TEI-formaat
- Alignatie
  - Zinsalignatie voor hele corpus
  - Alignatie onder zinsniveau (proof-of-concept)
- Taalkundige annotatie
  - PoS, lemma (D-COI tools)
  - Shallow parsing

# Validatie

- Interne kwaliteitsbewaking
  - Volledig manuele verificatie (min. 10%)
  - Manuele steekproeven
  - Automatische controleprocedures
- Externe validatie
  - CST validatiecentrum
  - Xplanation: case study terminologie-extractie

# Corpusexploitatie

- Volledige teksten
  - XML-bestanden (TEI formaat)
- Databank
  - Bevraagbaar via web interface
  - Bilinguale concordantie (Keyword in Context)

# Kernteam

- KULeuven – Campus Kortrijk
  - Prof. Dr. Piet Desmet
  - Dr. Hans Paulussen
- HoGent – Departement Vertaalkunde
  - Prof. Dr. Willy Vandeweghe
  - Dra. Lieve Macken