

Dutch Parallel Corpus

Een multilinguaal &
multifunctioneel corpus

Accenta 2007

Donderdag 20 september

Dutch Parallel Corpus

- **Corpus:**
Gestructureerde verzameling van elektronische teksten
- **Parallel Corpus:**
Verzameling van vertaalde teksten
- **Dutch Parallel Corpus:**
Gestructureerde verzameling van vertaalde teksten uit het Nederlands

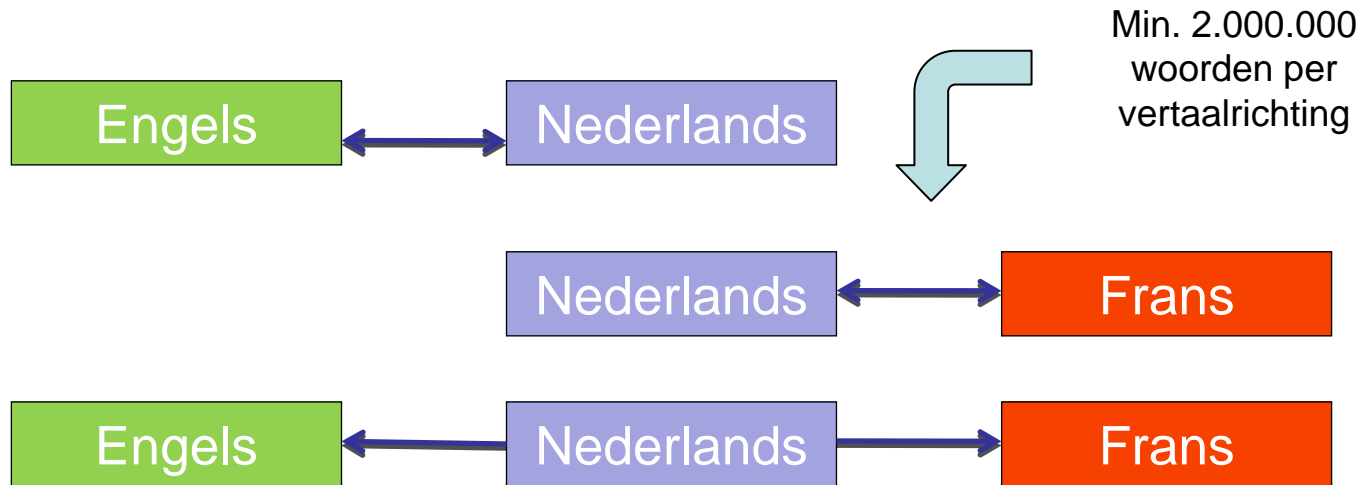
Dutch Parallel Corpus

- Corpus van 10 miljoen woorden
- Kwalitatief hoogstaand
- Ontwikkeling van een parallel corpus als prioriteit van het STEVIN-programma
- STEVIN: **S**praak- en **T**aaltechnologische **E**ssentiële **V**oorzieningen **I**n het **N**ederlands

Een multilinguaal corpus

- 2 taalparen:
 - Nederlands – Engels
 - Nederlands – Frans
- 4 vertaalrichtingen
- Gedeeltelijk drietalig:
Frans – Nederlands – Engels

Een multilinguaal corpus



→ Nederlands als scharniertaal

Een multifunctioneel corpus

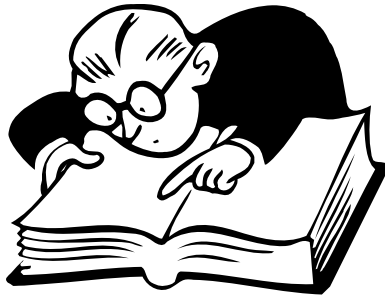
Departement Vertaalkunde

Hogeschool Gent

CALL-onderzoeksgroep

KU Leuven - campus Kortrijk

Parallel corpus als
vertaalhulpmiddel

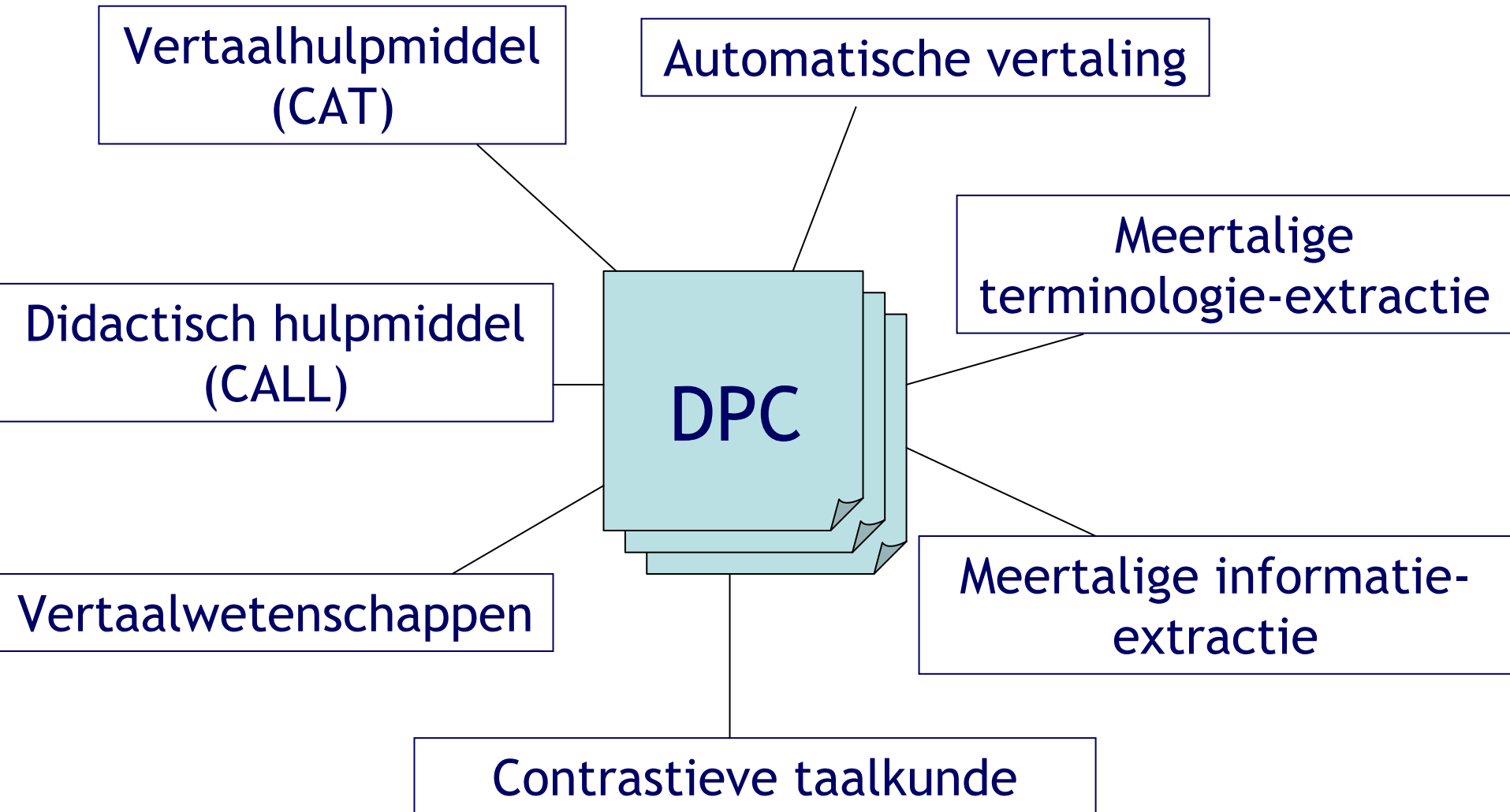


Parallel corpus als
didactisch hulpmiddel

alt research center on CALL
acquiring language through technology



Een multifunctioneel corpus



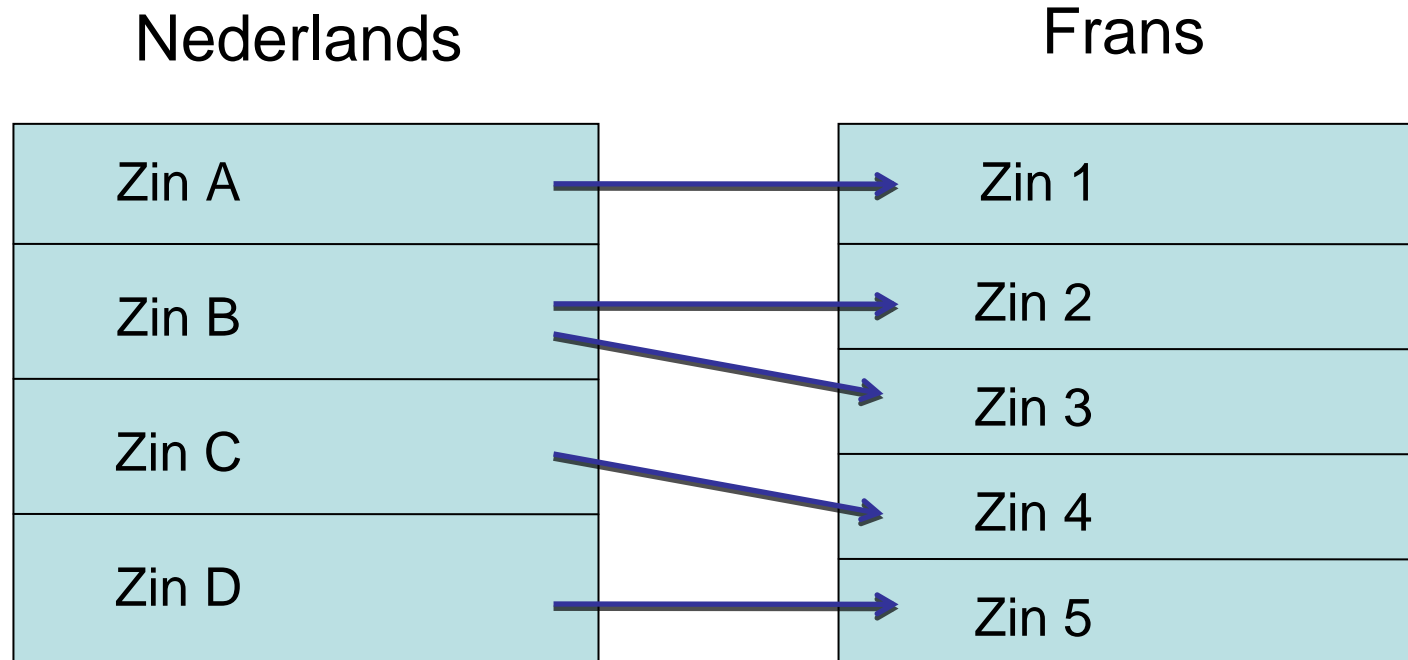
Samenstelling corpus

- Kwalitatieve tekstleveranciers
- Voorkeur voor gepubliceerd tekstmateriaal
- Variatie aan teksttypes:
Fictie/ Non-fictie
- IPR-overeenkomst (copyright)
noodzakelijk voor openbaar maken van teksten.

Corpusstructuur

- Alignering op zinsniveau
- Taalkundige annotatie
- Metadata

Alignering op zinsniveau



→ Zin A in brontaal komt overeen met zin 1 in doeltaal

Taalkundige annotatie

- Markeren van woorden, woordsoort, lemma, zinnen, syntactische structuren

Metadata

- Metadata maken het opsporen van parallele zinnen makkelijker
 - Datum publicatie
 - Vertaalrichting
 - Vertaalmodaliteiten
 - Directe vs. indirecte vertalingen
 - Kwaliteitslabel

Corpusontsluiting

- Consulteren van corpus via webinterface
- Elementen die het ontsluiten van corpusdata vergemakkelijken:
 - Zinsalignering
 - Taalkundige annotatie
 - Metadata

DPC consortium

- Kernteam
- Onderzoekspartners
- Gebruikersgroep

Kernteam

- K.U. Leuven campus Kortrijk

Prof. Dr. Piet Desmet

Dr. Hans Paulussen

Dr. Julia Trushkina

Lic. Maribel Montero Perez

- Hogeschool Gent

Departement Vertaalkunde

Prof. Dr. Willy Vandeweghe

Dra. Lieve Macken

Lic. Lidia Rura

Onderzoekspartners

- Rijksuniversiteit Groningen
- Radboud Universiteit Nijmegen
- Universiteit van Tilburg
- Katholieke Universiteit Leuven
- Universiteit Antwerpen
- Universiteit Gent

Gebruikersgroep

- Potentiële gebruikers van een vertaalcorpus, geselecteerd uit de academische wereld en de bedrijfswereld
 - Academische partners
 - Universiteit Antwerpen
 - Vlekho Hogeschool, Brussel
 - Lessius Hogeschool, Antwerpen
 - Hogeschool Gent
 - Katholieke Universiteit Leuven
 - Universiteit Gent
 - Industriële partners
 - Indiegroupp, Kortrijk
 - Xplanation, Leuven

Distributie

- Via TST-centrale
(Taal- en Spraaktechnologie)

TST-CENTRALE)))
voor taal- en spraaktechnologie

Informatie

<http://www.kuleuven-kortrijk.be/dpc>