



DPC (Dutch Parallel Corpus) een multitalig multifunctioneel corpus

Hans Paulussen & Julia Trushkina

K.U.Leuven / K.U.Leuven Campus Kortrijk

ALT Research Center on CALL

Overzicht

- Situering
- DPC
- Tekststandaardisering
- Toepassingsvoorbeeld: corpusCALL

DPC: Dutch Parallel Corpus

- Parallel corpus
- Gealigneerd op zinsniveau
- 10 miljoen woorden NL-FR & NL-EN
- Kwaliteitscontrole
- Compatibel met D-COI





1

Situering

Voorgeschiedenis

- K.U.Leuven – Campus Kortrijk
 - parallel corpus als didactisch hulpmiddel
 - REBECA (samenwerking FUNDP, Namur)
- HoGent Departement Vertaalkunde
 - parallel corpus als vertaalhulpmiddel



Groeiende vraag naar corpora

- Engels / andere talen
- Geschreven / gesproken
- Referentie / monitor
- Monolinguaal / multilinguaal
- Parallel / comparable

NL in meertalige corpora

- OMC: 170.000 NL woorden
- Namur Corpus: 700.000
- MLCC: 7.100.000
- Europarl: > 29.000.000

2

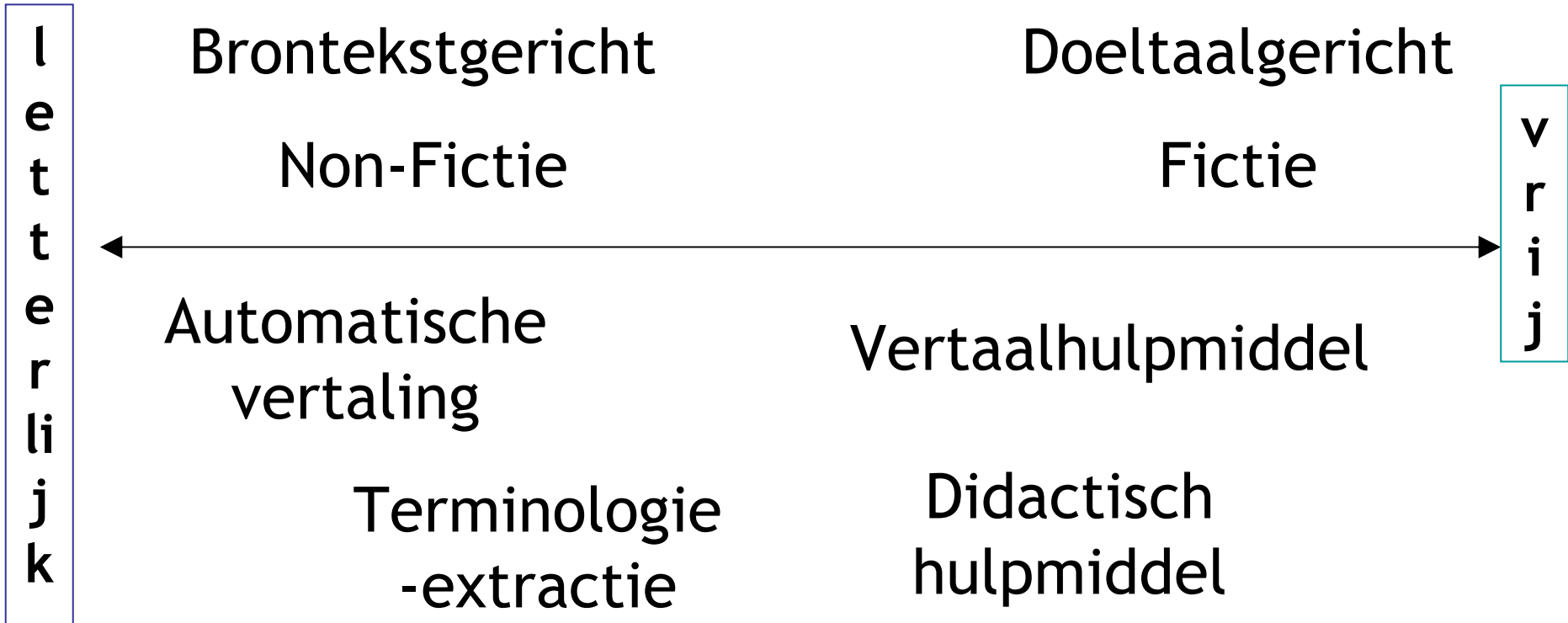
Dutch Parallel Corpus

DPC doelstellingen

- Evenwichtige compositie
 - verschillende teksttypes in fictie en non-fictie
- Annotatie
 - PoS, lemmata
- kwaliteitscontrole
 - verschillende niveau's
- beschikbaarheid
 - TST-centrale



Samenstelling corpus



Kwaliteitsvereisten

- Gedistribueerd
- Tekstkwaliteit & vertaalkwaliteit
- Annoteringskwaliteit
- Kwaliteitslabel



Kwalitatieve tekstleveranciers

Teksttypes (non fictie)	Sector
Journalistiek	Kranten en tijdschriften
Essayistisch	Uitgeverijen
Zakelijk	Bank & verzekering
Technisch	ICT & medisch
Ambtelijk	Government

Annotatie

- Zinnen
- Woorden
- Lemmata
- Woordsoorten
- Syntactisch

Metadata

- Vertaalrichting
- Vertaalmodaliteiten
- Directe vs. indirecte vertalingen
- Kwaliteitslabel

Kwaliteitslabel

- Manueel geverifiëerd
- Spot checking
- Automatische controleprocedures
 - e.g. automatische vergelijking van output van verschillende aligneringsprogramma's



Corpusontsluiting

- Databank
 - Bevraagbaar via webinterface
 - Bilinguale concordantie
- Volledige teksten
 - XML-bestanden (TEI)



Gebruikers en toepassingen

- Taaltechnologische toepassingen
 - Automatische vertaling / terminologie-extractie
 - Training- en testmateriaal
- Ondersteuningstoepassingen
 - Vertaalhulp
 - Dicactische ondersteuning (CALL)
 - Uitbreiding vertaalwoordenboeken
- Fundamenteel onderzoek
 - Vertaalwetenschap / contrastieve taalkunde
 - Multitalig corpusonderzoek

DPC kernteam

- K.U.Leuven Campus Kortrijk
 - Prof. Dr. Piet Desmet
 - Dr. Hans Paulussen
 - Dr. Julia Trushkina
 - Lic. Antoine Besnehard
- HoGent – Departement Vertaalkunde
 - Prof. Dr. Willy Vandeweghe
 - Dra. Lieve Macken
 - Lic. Lidia Rura



DPC Onderzoekspartners

- Rijksuniversiteit Groningen
- Radboud universiteit Nijmegen
- Universiteit van Tilburg
- K.U.Leuven
- Universiteit Antwerpen
- Universiteit Gent

DPC gebruikersgroep

- Geconsulteerd bij belangrijke ontwerpbeslissingen
- Industriële partners,
 - CALL, vertaaldiensten, Terminologie-extractie, Informatie-extractie
- Academische partners
 - Taaltechnologie
 - Vertaalwetenschappen
 - Contrastieve taalkunde

3

Tekststandaardisering

tekststandaardisering

- XML (eXtensible Markup Language)
 - well-formed
 - eenduidigheid formaat XML tags
 - valide
 - beantwoordt aan DTD (*documentgrammatica*)

XML well-formed

- (1) Het document bevat één root-element dat alle andere elementen omvat.
- (2) Elke begintag heeft een overeenkomstige eindtag.

```
<book>  
<chapter> ...</chapter>  
</book>
```

XML-validatie

- DTD: document type definition
 - Documentgrammatica

DTD book

```
<!DOCTYPE book [  
  <!ELEMENT book      (title, chapter+)>  
  <!ELEMENT chapter   (heading, paragraph*)>  
  <!ELEMENT title     (#PCDATA)>  
  <!ELEMENT heading   (#PCDATA)>  
  <!ELEMENT paragraph (#PCDATA)>  
  <!ATTLIST book  
    language  CDATA #REQUIRED  
    author    CDATA #REQUIRED  
>  
>
```



book.xml

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE book SYSTEM "book.dtd">
<book language="en" author="Ernest Hemingway">
  <title>A Farewell to Arms</title>
  <chapter>
    <heading>Chapter One</heading>
    <paragraph> ... <paragraph>
    <paragraph> ... <paragraph>
  </chapter>
  <chapter>
    <heading>Chapter Two</heading>
    <paragraph> ... <paragraph>
    <paragraph> ... <paragraph>
  </chapter>
</book>
```

DTD poème

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE poème [
  <!ELEMENT poème      (préambule, corps)>
  <!ELEMENT préambule (titre, recueil?, date?, auteur)>
  <!ELEMENT titre      (#PCDATA)>
  <!ELEMENT recueil    (#PCDATA)>
  <!ELEMENT date       (#PCDATA)>
  <!ELEMENT auteur     (#PCDATA)>
  <!ELEMENT corps      (stance|ligne)+>
  <!ELEMENT stance     (ligne)+>
  <!ELEMENT ligne      (#PCDATA|r)*>
  <!ELEMENT r          EMPTY>
]>
```

TEI & CES

- TEI (Text Encoding Initiative) is een standardformaat om teksten te structureren in SGML of XML
- CES (Corpus Encoding Initiative) is een gelijkaardige standaard

TEI DTD

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main Document Type//EN" [
  <!ENTITY % TEI.XML      'INCLUDE'>
  <!--enable XML processing-->
  <!ENTITY % TEI.prose    'INCLUDE'>
  <!--base tag set for prose -->
  <!ENTITY % TEI.analysis 'INCLUDE'>
  <!--linguistic analysis-->
  <!ENTITY % TEI.linking  'INCLUDE'>
  <!--pointer mechanisms-->
]>
```



TEI unitary document

```
<TEI.2>
  <teiHeader>
    [ TEI Header information ]
  </teiHeader>
  <text>
    <front> [ front matter ... ] </front>
    <body> [ body of text ... ] </body>
    <back> [ back matter ... ] </back>
  </text>
</TEI.2>
```



TEI composite document

```
<TEI.2>
  <teiHeader>
    [ header information for the composite ]
  </teiHeader>
  <text>
    <front> [ front matter for the composite ] </front>
    <group>
      <text>
        <front> [ front matter of first text ] </front>
        <body> [ body of first text ] </body>
        <back> [ back matter of first text ] </back>
      </text>
      <text>
        <front> [ front matter of second text ] </front>
        <body> [ body of second text ] </body>
        <back> [ back matter of second text ] </back>
      </text>
      [ more texts or groups of texts here ]
    </group>
    <back> [ back matter for the composite ] </back>
  </text>
</TEI.2>
```

teiCorpus

```
<teiCorpus>
  <teiHeader> [header information for the corpus]</teiHeader>
  <TEI.2>
    <teiHeader>[header information for first text]</teiHeader>
    <text> [first text in corpus] </text>
  </TEI.2>
  <TEI.2>
    <teiHeader>[header information for second text]</teiHeader>
    <text> [second text in corpus] </text>
  </TEI.2>
</teiCorpus>
```

4

Toepassingsvoorbeeld : CorpusCALL



Corpora voor CALL /1

- Corpora voor **leeractiviteiten**
- Corpora als **referentiemateriaal**



Corpora voor CALL /2

- Corpora voor **leeractiviteiten**
 - voor: *voorbereiding* van oefeningen
 - tijdens: corpusmateriaal als onderdeel van *leeractiviteit*
 - na: corpusmateriaal als *feedback*

Corpora voor CALL /3

- Corpora als **referentiemateriaal**
 - leerwoordenboeken
 - DAFLES: dictionnaire d'apprentissage du français langue étrangère ou seconde
 - leergrammatica's
 - ALFAGRAM

REBECA

- *Resources électroniques bilingues extraites de corpus alignés*
- Corpusondersteuning NEDERLEX-project (FUNDP Namur)
- Webapplicatie leesoefeningen Nederlands met ondersteuning van parallel corpusmateriaal
- Automatische corpuselectie



5. Gezondheid en leefmilieu in België

Index

Nauwelijks een eeuw geleden leden duizenden mensen in ons land nog aan ziekten veroorzaakt door de slechte kwaliteit van het leef- en werkmilieu.

Die tijd is intussen voorbij. Tal van ziekten zijn onder controle. Maar vandaag heeft de overheid totaal andere gezondheidsproblemen, die zijn veroorzaakt door vervuiling door de industrie, het verkeer en door de menselijke activiteit in het algemeen.

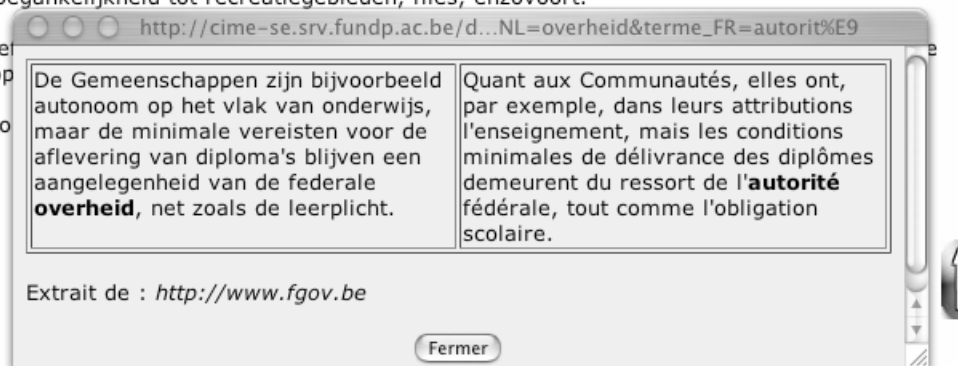
De opkomst van nieuwe chemische producten, nieuwe productieprocessen en technologieën en de vermenging van allerlei pollutiebronnen hebben hun weerslag op het klimaat, de kwaliteit van de lucht en de bodem, de biodiversiteit en de voedselketen. Vaak is het effect ervan pas na enkele jaren of zelfs decennia later zichtbaar.

Bovendien is de verstedelijking sterk toegenomen. In 2000 leefde ongeveer 80% van de bevolking in stedelijke gebieden. Dat heeft gevolgen. In heel wat steden duiken hoe langer hoe meer stressverschijnselen op die te maken hebben met het leefmilieu: ozonpieken, zware luchtvervuiling, toenemend lawaai, stijgende afvalproductie, moeilijkere toegankelijkheid tot recreatiegebieden, files, enzovoort.

En dan is er nog de maatschappelijke ongelijkheid. Die heeft verschillende factoren. Die factoren hebben een rechtstreekse invloed op de gezondheid.

De strijd tegen ziekte en vervuiling kan dus maar succesvol zijn als de overheid erin slaagt de levensomstandigheden te verbeteren voor het welzijn van de hele bevolking.

bron: <http://www.belgium.be> - 19.09.2003



overheid (nom, de, overheden)	
... nder toezicht van alle hogere overheden , in het kader van de fe nales en étant subordonnées à toutes les autorités supérieures.
... angelegenheid van de federale overheid , net zoals de leerplicht mes demeurent du ressort de l' autorité fédérale, tout comme l'o ...
De overheid heeft een nieuw reglement uitgevaardigd.	Les autorités ont promulgué un nouveau règlement.
... erd in welke administratie en/of overheid daarbij betrokken is.	... r quelle administration et/ou pouvoir public est impliqué dans ...
... erd in welke administratie en/of overheid daarbij betrokken is.	... s intéressés de savoir quelle administration et/ou pouvoir publ ...
... hillende administraties en/of overheden die hierbij betrokken z r quelle administration et/ou pouvoir public est impliqué dans ...
... hillende administraties en/of overheden die hierbij betrokken z s intéressés de savoir quelle administration et/ou pouvoir publ ...
Administraties en overheden zullen elkaars gegevens zoveel mog ...	Les administrations et les autorités doivent partager et utili ...

Informatie

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