

# Dutch Parallel Corpus

## Multilinguaal & multifunctioneel

Lieve Macken

LT<sup>3</sup>

Hogeschool Gent

# Dutch Parallel Corpus

- Parallel corpus
  - Teksten + vertaling
  - Gealigneerd op zinsniveau
- 10 miljoen woorden
- Nederlands - Engels / Nederlands - Frans
- Kwalitatief
- Compatibel met Corpus Geschreven Nederlands
- Stevin-project
  - Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands
  - Gefinancierd door de Nederlandse Taalunie
- 2006-2009

# Voorgeschiedenis

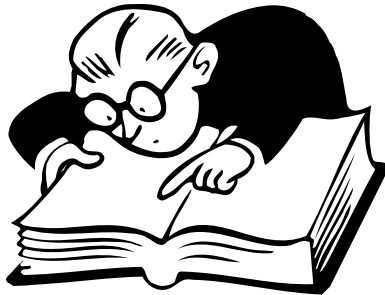
Departement Vertaalkunde

Hogeschool Gent

CALL-onderzoeksgroep

KU Leuven - Campus Kortrijk

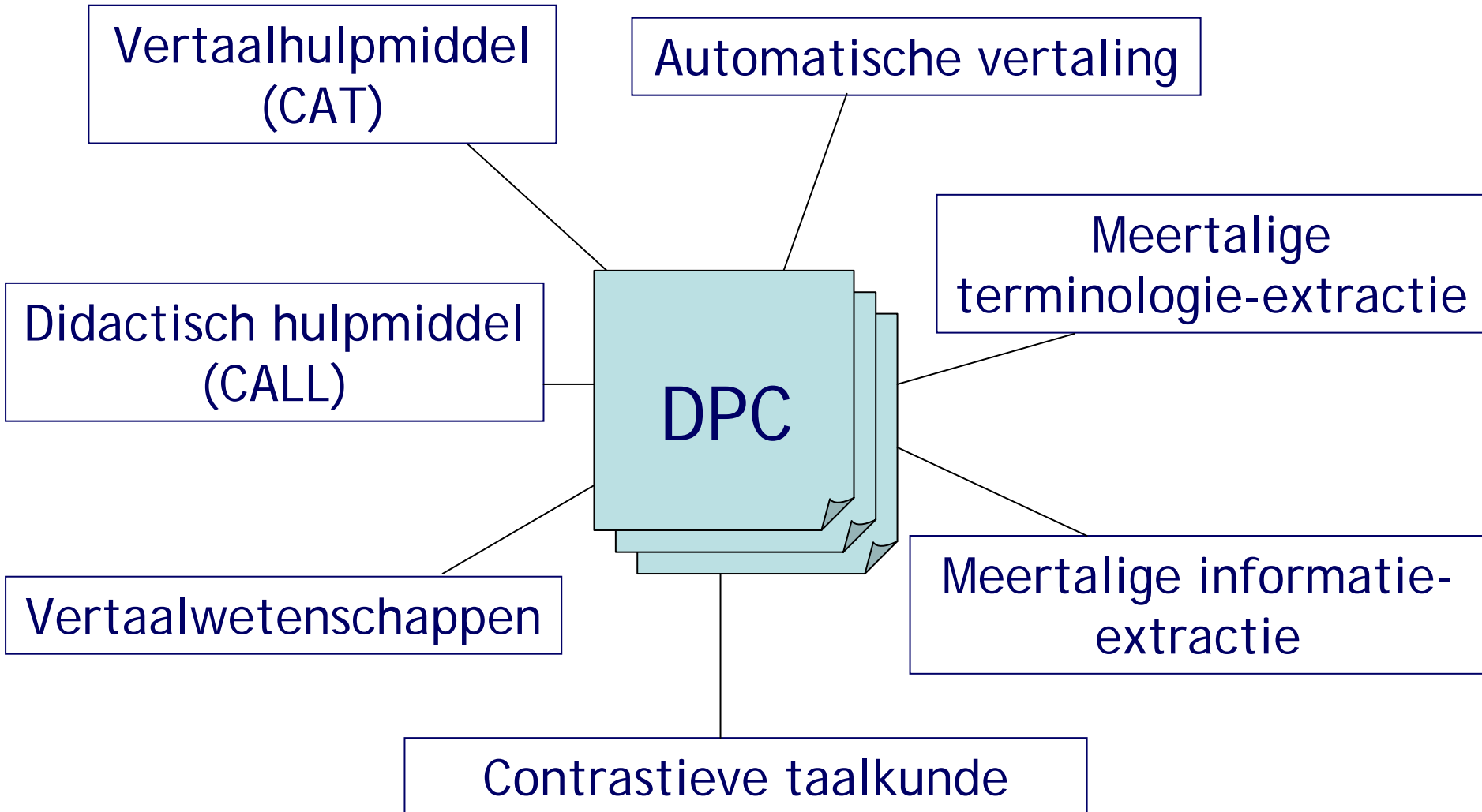
Parallel corpus als  
vertaalhulpmiddel



Parallel corpus als  
didactisch hulpmiddel



# Gebruikers en toepassingen



# Automatische vertaling

- Training- en testmateriaal voor corpus-gebaseerde MT
  - Example Based MT
  - Statistical MT
- P. Khoen 2005: 110 SMT-systemen getraind op Europarl-corpus
  - Voorbeeld uitvoer Fins-Engels:  
*we know very well that the current treaties are not enough and that in future , it is necessary to develop a better structure for the union and , therefore perustuslaillisempi structure , which also expressed more clearly what the member states and the union is concerned .*

# Terminologie-extractie

Term (720 terms)

Filter: <No filter>

Find:

Score	Domain	Dutch	EN English
92	<None>	<input checked="" type="checkbox"/> verandering	<input checked="" type="checkbox"/> change <input type="checkbox"/> changes <input type="checkbox"/> changing
92	<None>	<input checked="" type="checkbox"/> ontwerpresolutie	<input checked="" type="checkbox"/> draft resolution
92	<None>	<input checked="" type="checkbox"/> interne markt	<input type="checkbox"/> single market <input checked="" type="checkbox"/> internal market
92	<None>	<input checked="" type="checkbox"/> financiële controle	<input checked="" type="checkbox"/> financial control
92	<None>	<input type="checkbox"/> concurrentiebeleid	<input checked="" type="checkbox"/> competition policy
92	<None>	<input checked="" type="checkbox"/> Commissie-ambtenaren	<input checked="" type="checkbox"/> Commission officials
92	<None>	<input type="checkbox"/> Commissar	<input type="checkbox"/> Commissioner Monti
92	<None>	<input checked="" type="checkbox"/> aansprakelijk	<input type="checkbox"/> liability <input checked="" type="checkbox"/> liable
91	<None>	<input checked="" type="checkbox"/> problemen	<input type="checkbox"/> problem <input checked="" type="checkbox"/> problems <input type="checkbox"/> difficulty <input checked="" type="checkbox"/> difficulties <input type="checkbox"/> operation

Concordance (5 sentences)

← ← → → Search

**NL**

**EN**

**NL** Wij zijn het eens met de zienswijze , die zeer duidelijk is geformuleerd door het Comité van onafhankelijke deskundigen en ook in de hier ter tafel liggende **ontwerpresolutie** , dat de tijd is aangebroken voor een grondige herziening van onze regels en procedures .

**EN** We share the view , forcefully expressed by the Committee of Independent Experts , and repeated in this **draft resolution** that the time has come for a thorough overhaul of our rules and procedures .

[Source file: tl-nl-00-01-16.tmx Sentence number: 659]

**NL** Ik ben enigszins verbaasd dat paragraaf 10 van de **ontwerpresolutie** niet lijkt in te gaan op de zware kritiek die in beide verslagen van het Comité van onafhankelijke deskundigen wordt geuit ten aanzien van het huidige gecentraliseerde systeem van financiële controle .

# Vertaalhulpmiddel

- Hulpmiddel tijdens vertaalproces
  - Bij zoektocht naar meest geschikte term, woord, stijl, idiomatisch taalgebruik, ...
  - Aanvulling op bilinguale woordenboeken
  - Uitbreiding op monolinguaal 'googelen'
  - Woorden in context
- Voorbeeld: TransSearch (Canadian Hansards)
  - Simard & Macklovitch 2005

<i>match</i>	<i>source</i>	<i>target</i>
1.	Members on that side of the House started <b>ragging the puck</b> .	Les députés d'en face ont commencé à tricoter avec la rondelle.
2.	Mr. Speaker, being a former hockey player I was used to <b>ragging the puck</b> whenever I was able to get it.	Monsieur le Président, en tant qu'ancien joueur de hockey, j'ai l'habitude de taquiner la rondelle chaque fois que j'en ai la chance.
3.	They are trying to rag the puck just as the Detroit Red Wings tried to <b>rag the puck</b> .	Nos vis-à-vis tricotent avec la rondelle en quelque sorte à l'instar des Red Wings de Détroit.
4.	...	...

Figure 2: Results for the *TransSearch* query "rag+ . . puck"

# CorpusCall

- Computerondersteund talenonderwijs
  - Leeractiviteiten
  - Referentiemateriaal
- Woorden in context
  - Authentiek materiaal in leertaal
  - Ondersteuning in moedertaal
- Voorbeeld Nederlex
  - Leesomgeving voor Franstalige studenten
  - Ontwikkeling leesomgeving: FUNDP, Namur
  - Compilatie parallel corpus: REBECA project (K.U.Leuven Campus Kortrijk)

# Nederlex

## 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.

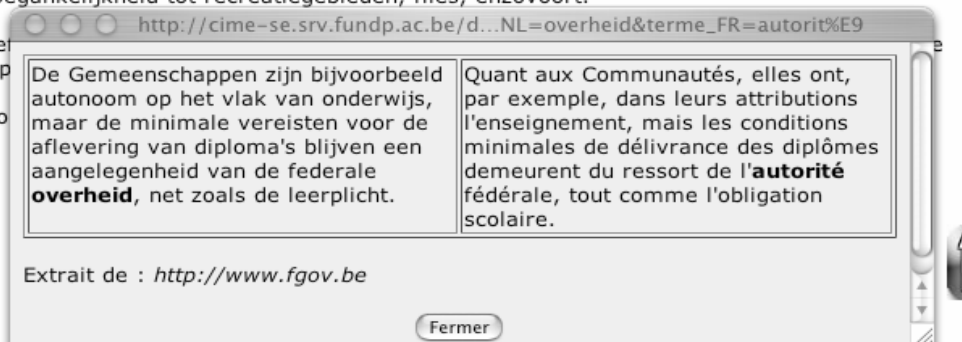
De opkomst van nieuwe chemische producten, nieuwe productieprocessen en technologieën en de vermenging van allerlei pollutiebronnen hebben hun weerslag op het klimaat, de kwaliteit van de lucht en de bodem, de biodiversiteit en de voedselketen. Vaak is het effect ervan pas na enkele jaren of zelfs decennia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

De strijd tegen ziekte en vervuiling kan dus maar succesvol zijn als we de welzijn van de hele bevolking verbeteren.

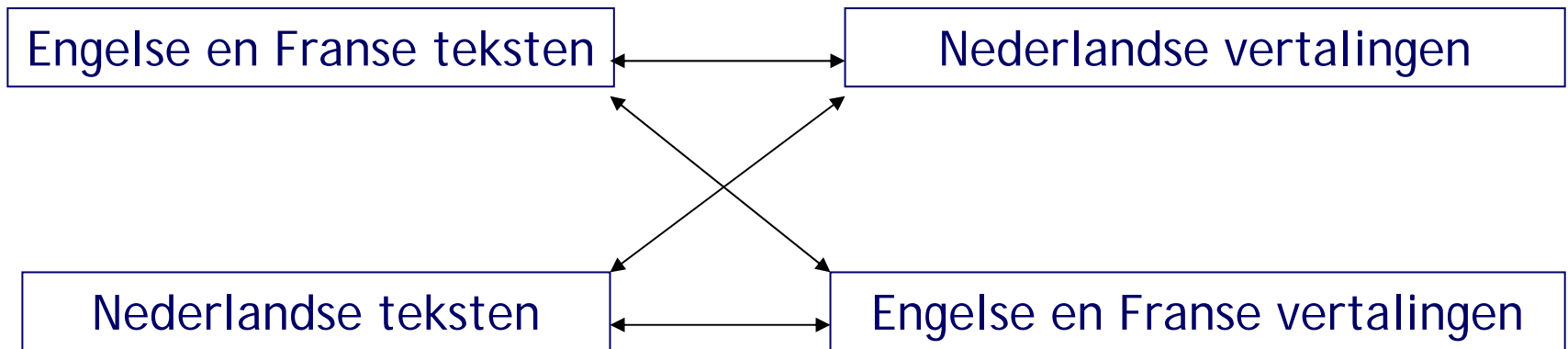
bron: <http://www.belgium.be> - 19.09.2003



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

# Vertaalwetenschappen

- Studie van het vertaalproduct
  - Vertaaluniversalia en translationese
  - Vertaalproces
- Parallele en vergelijkbare corpora



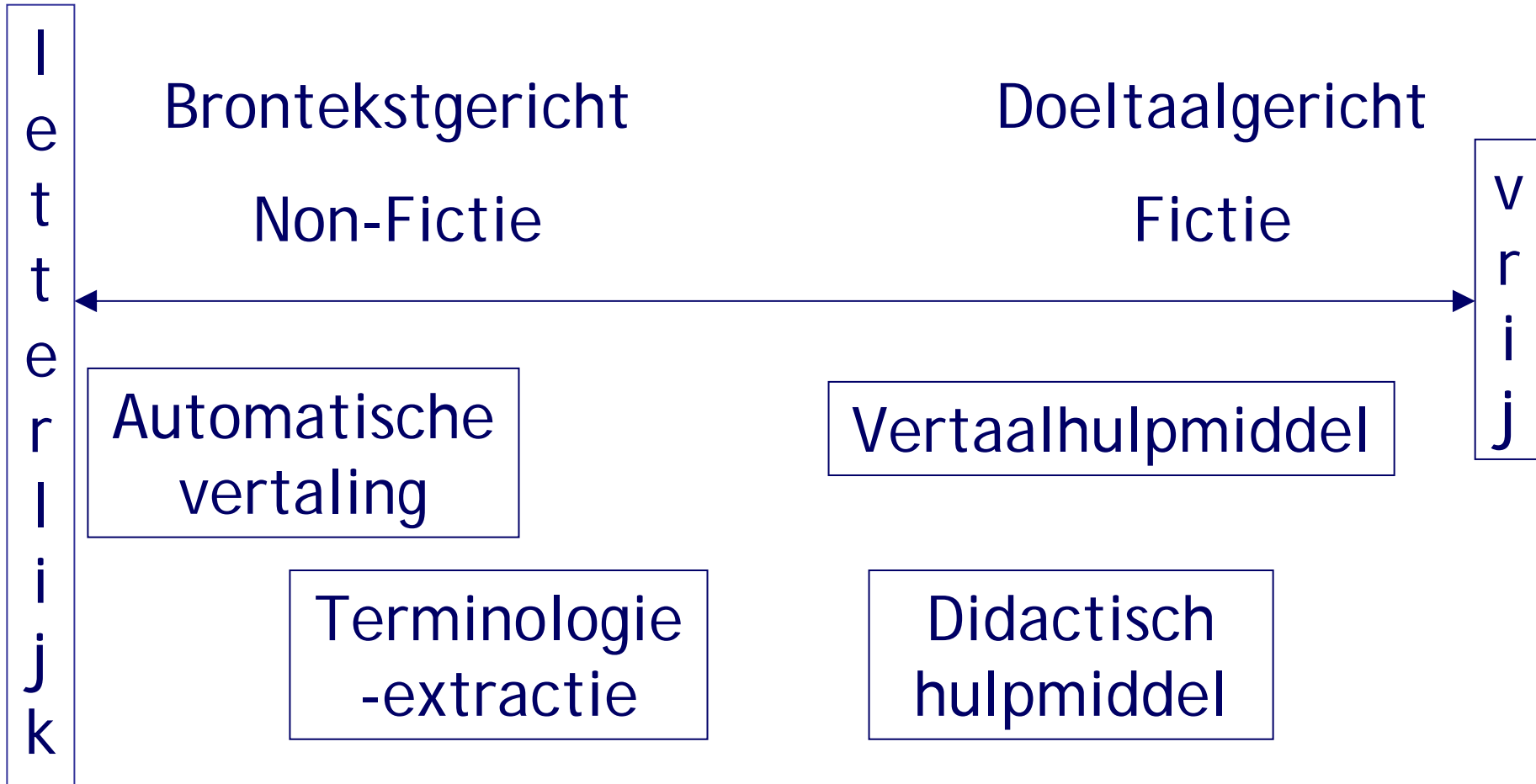
# Verschillende gebruikers ...

- Taaltechnologische toepassingen
  - Automatische vertaling / terminologie-extractie
  - Andere NLP-toepassingen (bijv. WSD)
  - Training- en testmateriaal
- Menselijke gebruikers
  - Vertaalhulpmiddel / didactisch hulpmiddel
  - Concordantieprogramma's
  - Aanvulling bilinguale woordenboeken
- Fundamenteel Onderzoek
  - Vertaalwetenschap / contrastieve taalkunde
  - Parallel en vergelijkbaar corpus

# ... stellen verschillende eisen

- 1) Samenstelling Corpus
- 2) Metadata
- 3) Taalkundige annotatie
- 4) Kwaliteitsvereisten
- 5) Corpusontsluiting

# Samenstelling Corpus



# Samenstelling corpus /2

- Fictie
- Non-fictie
  - Essayistische teksten
  - Journalistieke teksten
  - Zakelijke teksten
  - Technische teksten
  - Ambtelijke teksten

# Metadata

- Vertaalrichting
  - Engels → Nederlands vs. Nederlands → Engels
- Vertaalmodaliteiten
  - Menselijke vertaling, CAT, MT
- Directe vs. indirecte vertalingen
  - Indirect via Engels (vb. Europarl)

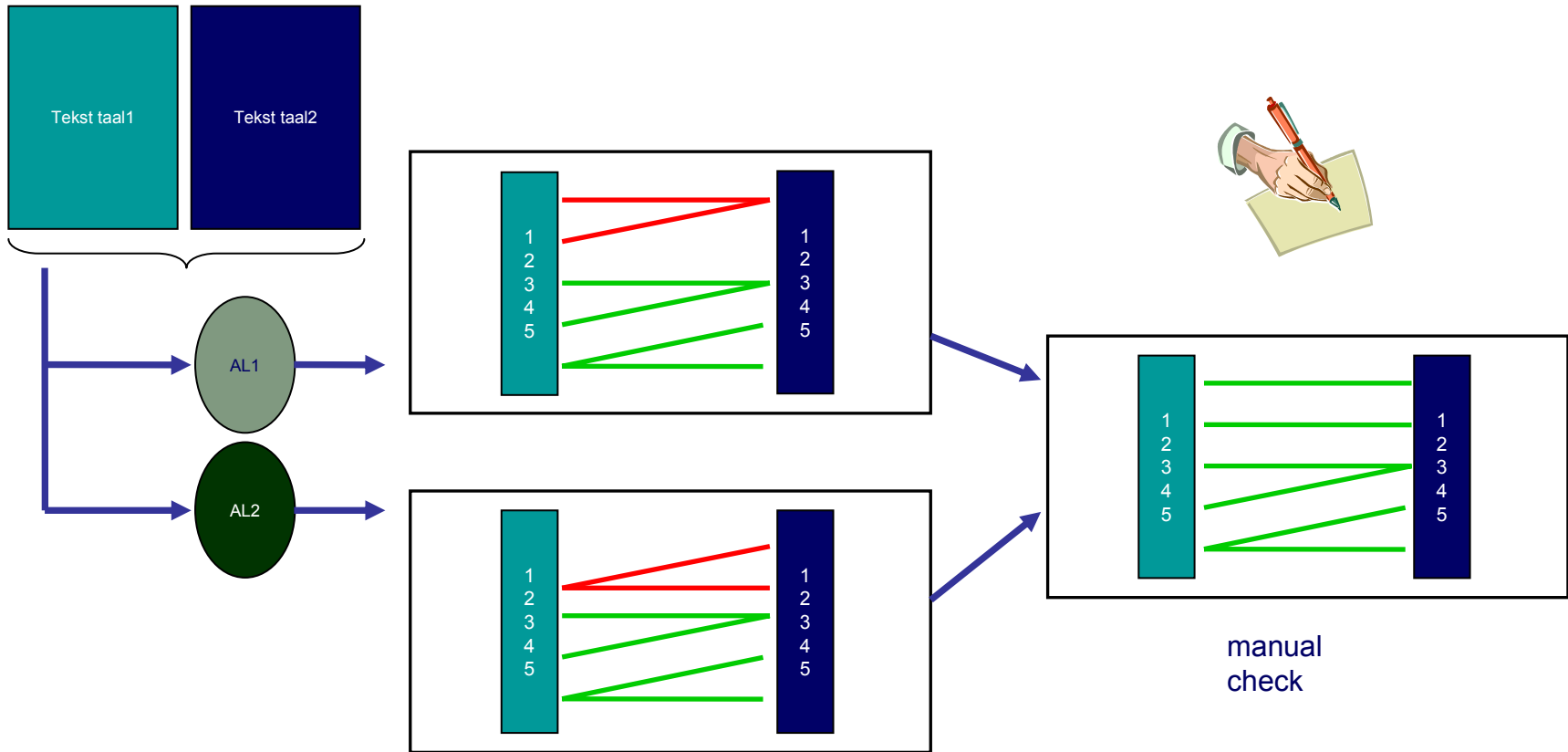
# Taalkundige annotatie

- Basiselementen
  - Paragrafen, zinnen, woorden
- Alignatie
  - zinsniveau
- Taalkundige verrijking
  - Lemma
  - Woordsoort
  - Syntactische structuren

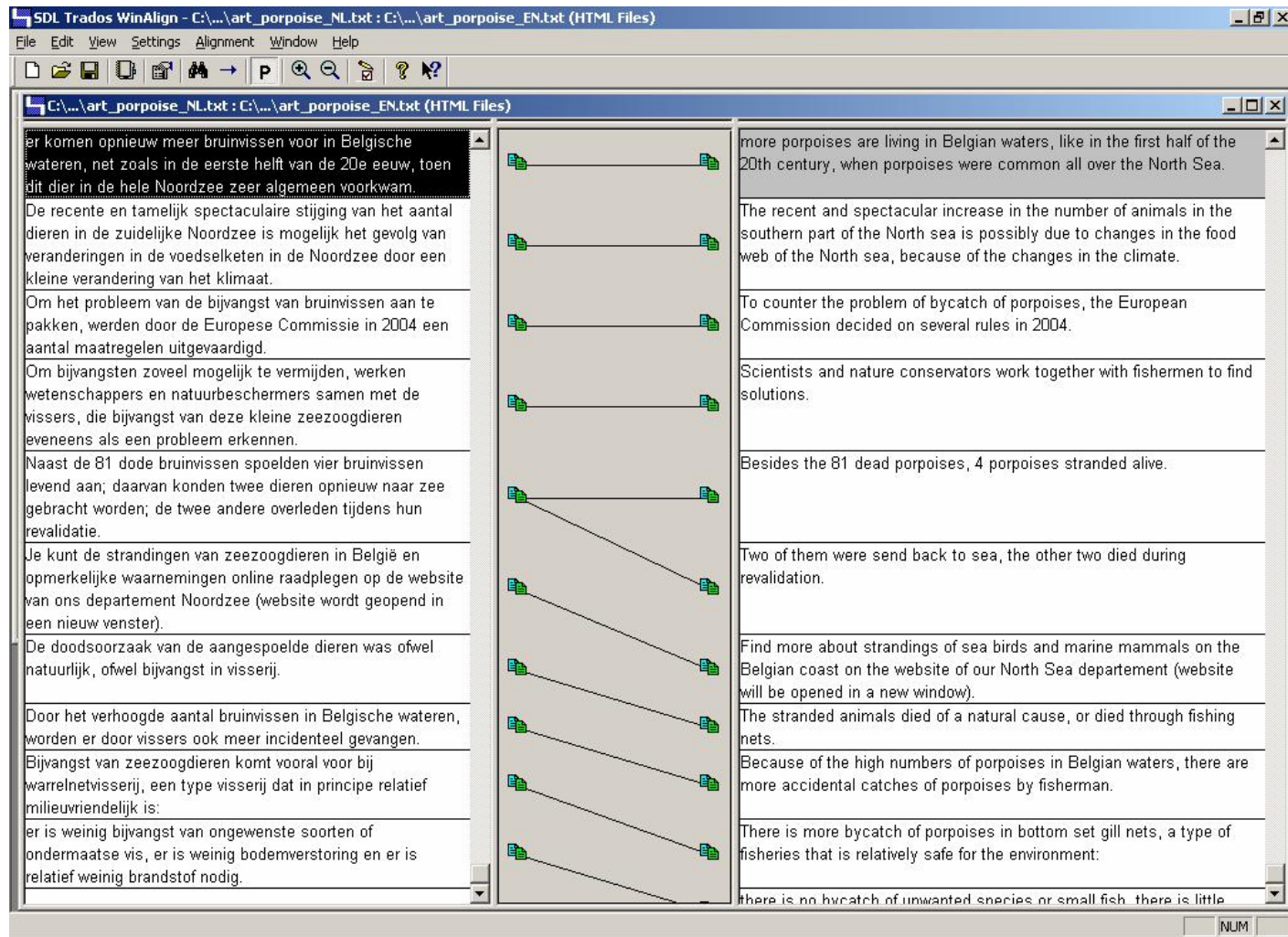
# Kwaliteitsvereisten

- Verschillende niveaus
  - Volledig manuele verificatie
  - Manuele steekproeven
  - Automatische controleprocedures
    - Bijv. Automatische vergelijking van uitvoer van verschillende alignatieprogramma's
- Kwaliteitslabel

# Samenvoegen van alignaties



# Manuele verificatie



# Corpusontsluiting

- Webinterface
  - Gebruiksvriendelijk
  - Beperkte technische know-how bij taaldocenten & vertalers
  - Eenvoudige & complexe zoekopdrachten
- Volledige teksten
  - Lerende systemen (data-driven automatic learning)
  - Statistische MT

# Eenvoudige zoekopdracht

## Eenvoudige zoekopdracht : [\*spoel\*]

### Parallel Concordance - [\*spoel\*]

Aantal [aangespoelde](#) bruinvissen verdubbeld

In 2005 [spoelden](#) aan de Belgische kust 81 bruinvissen aan, wat meer is dan ooit tevoren.

Naast de 81 dode bruinvissen [spoelden](#) vier bruinvissen levend aan; daarvan konden twee dieren opnieuw r  
Wetenschappers van ons departement Noordzee nemen een [aangespoelde](#) bruinvis mee voor onderzoek.  
De doodsoorzaak van de [aangespoelde](#) dieren was ofwel natuurlijk, ofwel bijvangst in visserij.

Number of [stranded](#) porpoises doubled

In 2005, 81 porpoises [stranded](#) on the Belgian coast. This is more than any other year.

Besides the 81 dead porpoises, 4 porpoises [stranded](#) alive. Two of them were send back to sea, the other t  
Scientists of our North Sea department take a [stranded](#) porpoise with them for further research.  
The [stranded](#) animals died of a natural cause, or died through fishing nets.

# Complexe zoekopdracht

PoS query: [ PoS = "ADJ" | PoS = "WW.vd" ] "bruinvis\*"

<-- 01 -->

NL: In 2005 spoelden aan de Belgische kust 81 bruinvissen aan, wat meer is dan ooit tevoren. Het aantal << **aangespoelde bruinvissen** >> betekent een verdubbeling tegenover 2004, toen 41 dieren aanspoelden.

FR: En 2005, 81 marsouins se sont échoués sur la côte belge, un record en la matière ! Ce nombre de marsouins échoués est le double de celui enregistré en 2004, où 41 échouages de ces mammifères marins avaient été recensés.

EM: In 2005, 81 porpoises stranded on the Belgian coast. This is more than any other year, and this is the double of the 41 stranded animals in 2004.

<-- 02 -->

NL: Naast de 81 << **dode bruinvissen** >> spoelden vier bruinvissen levend aan; daarvan konden twee dieren opnieuw naar zee gebracht worden; de twee andere overleden tijdens hun revalidatie.

FR: À côté de ces 81 marsouins morts, 4 se sont échoués vivants.

EM: Besides the 81 dead porpoises, 4 porpoises stranded alive.

# Gebruikerscommissie

- Geconsulteerd bij belangrijke ontwerpbeslissingen
- Industriële partners
  - Computer-assisted language learning
  - Vertaaldiensten
  - Terminologie-extractie
  - Informatie-extractie
- Academische partners
  - Taaltechnologie
  - Vertaalwetenschappen
  - Contrastieve taalkunde

# Kernteam

- KULeuven - 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

Bedankt voor uw aandacht !

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