

DPC

What?, Why?, For whom?, How?

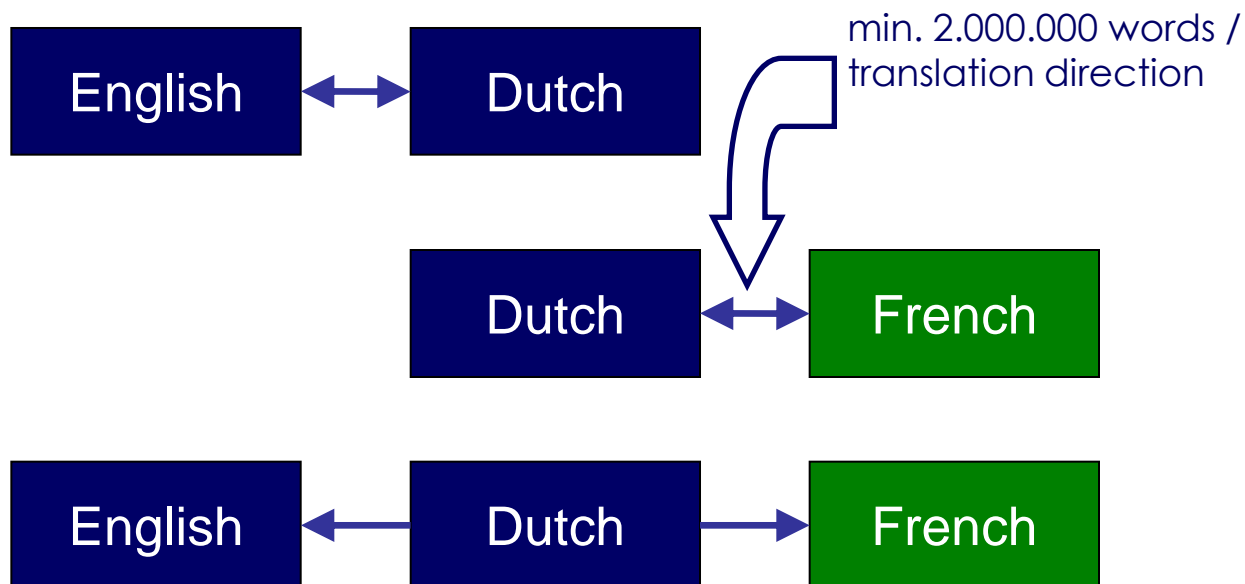
1. Objectives (What?)
2. Motivation (Why?)
3. Users (For whom?)
4. Team & workpackages (How?)

1. Objectives

- **High-quality corpus**
 - text providers (published materials, professional translators, etc.)
 - 10 million words (1M manually checked + 9M spot checking)
 - external validation (CST & Xplanation)
- **Annotated & sentence-aligned corpus**
 - Sentence alignment
 - Lemmatization
 - POS-tagging
- **Balanced composition**
 - Translation direction
 - Text types

1. Objectives

- **Parallel corpus**



- **Available corpus :**

Dutch Agency for Human Language Technologies (TST-centrale)

2. Motivation

Main parallel corpora with Dutch component (before DPC)

Corpus name	Size in words	Domains	Aligned ¹	Markup	PoS tagged
Namur	700,000	Fiction + Non Fiction (Unesco Courier + Debates of the European Parliament)	P	custom	No
ECI/MCI ²	25,000	EC Esprit program announcement text	X	TEI	No
MLCC	7,100,000	Debates of the European Parliament	X	TEI	No
Scania	216,424	Scania Truck manuals	S	TEI	No
OMC ³	170,000	Fiction	S	TEI	No
Europarl	29,188,340	Debates of the European Parliament	S	XML	No
OPUS ⁴	886,171	OS software manuals	S	XCES	Yes

Table 1: Main parallel corpora available with Dutch component

2. Motivation

Problems in existing corpora:

- Quantity > Quality
 - e.g. Europarl: alignment quality not verified
- Linguistic Annotation is lacking, only compilation (exc. OPUS)
- Unbalanced corpora
 - e.g. OPUS: documentation & manuals + Europarl
- Limited Availability
 - e.g. Scania: commercial corpus, Namur Corpus: PhD
- Minor position of Dutch: no or limited Dutch kernel
 - e.g. Oslo Multilingual Corpus (OMC) en OPUS: mainly translated texts for Dutch

- **Fundamental research**
 - Translation studies / contrastive linguistics
 - Corpus linguistics
- **Support for applications**
 - Translation support (CAT)
 - Didactic support (CALL)
- **HLT applications**
 - Machine Translation / Terminology Extraction
 - Training and test data

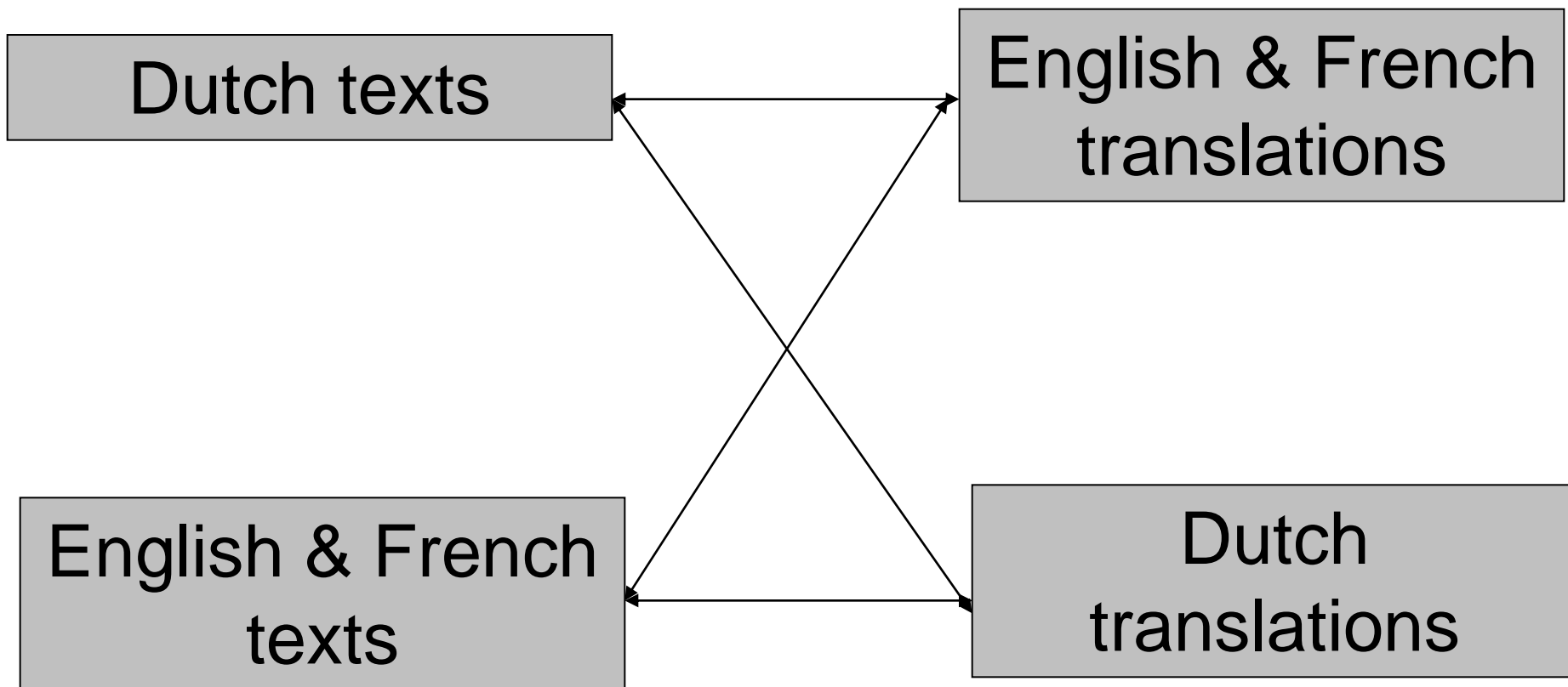
3.1. Fundamental Research

Translation product	Translation process
Language systems	Translation strategies

- High-quality data
- Balanced by translation direction

3.1. Fundamental Research

Parallel & comparable corpus



3.2. Support for Applications

Full text corpora as Translator's aid (=CAT)

- Computer assisted Translation
 - To identify more appropriate TL equivalent, idiomatic expressions
 - Extension to bilingual dictionaries
 - Words in context
- Example: TransSearch (Canadian Hansards)
 - Simard & Macklovitch 2005

<i>match</i>	<i>source</i>	<i>target</i>
1.	Members on that side of the House started ragging the puck .	Les députés d'en face ont commencé à tricoter avec la rondelle.
2.	Mr. Speaker, being a former hockey player I was used to ragging the puck 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 rag the puck.	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”

3.2. Support for Applications

Computer Assisted Language Learning: CorpusCALL

- Reference materials: learners dictionaries & grammars
- Support of learning activities
 - > **Nederlex** = Electronic reading platform for French students learning Dutch (FUNDP & K.U.Leuven Campus Kortrijk)
 - > **BLF & ALFALEX** = Lexical portal & learning environment (K.U.Leuven, ILT)

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.

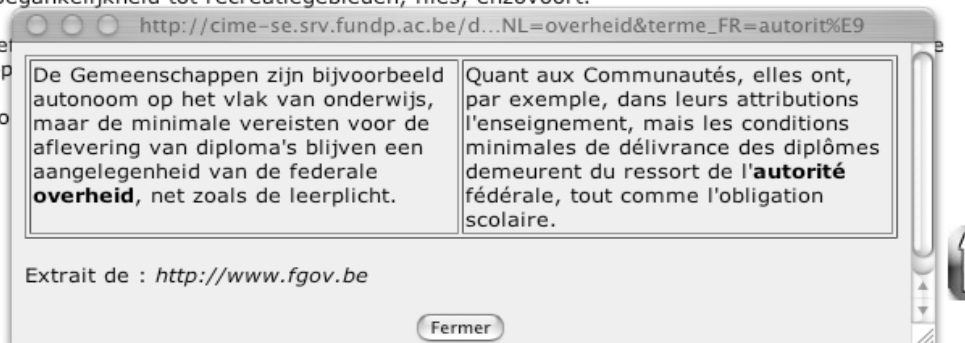
De opkomst van nieuwe chemische producten, nieuwe productieprocessen en technologieën en de vermenging van allerlei pollutiebronnen hebben hun weerslag op het klimaat, de kwaliteit van de lucht en de bodem, de biodiversiteit en de voedselketen. Vaak is het effect ervan pas na enkele jaren of zelfs decennia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

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 ...

Lexical Database for French (Base lexicale du français - BLF) - new site

(Almost) everything you always wanted to know about... French words


Get information on

- ⊕ a  [word](#)
Meaning, gender, use of prepositions, translation, ...
- ⊕ a  [word combination/expression](#)
... for *lors de, une ambiance règne, à tout prix, ...*

Verify

- ⊕ the [use](#) of a (sequence of) word(s)
what do you say: *apparaître {à or sur} l'écran? espérer {de or -} faire ...* - check on the web
- ⊕ if a [translation](#) is correct
salarié > salaried person?

Do (a lot of) exercises

- ⊕ on [all aspects of the vocabulary](#) of 
- On verb forms, gender, use of prepositions, word combinations, ...

Get the translation of

- ⊕ a      [word](#) to 
- ⊕ a      [word combination/expression](#) to 
*een vraag stellen, ask a question, hacer una pregunta > ?
une question*

Learn

- ⊕ how to [express an idea](#)
Something starts, something is intense, a large amount of something, words and word combinations about economics, ...
- ⊕ how to [combine words](#) (correctly) to form sentences
salary > to earn a ~ | employer, company, colleagues, ...
- ⊕ how to [avoid common errors](#)
Use of prepositions, position of the adjective, gender, ...

Help me

- ⊕ to [understand](#) a (short) text
"Beta" version: it's still a work in progress
- ⊕ while I'm writing
Work in progress

Interface

- ⊕ in English
Coming soon
- ⊕ en français
- ⊕ in het Nederlands

Help us to improve this tool

- ⊕ Did you find the information you needed?

yes no

- ⊕ Why not?

- ⊕ Suggestions?



DAFLES (Dictionnaire d'apprentissage du français langue étrangère ou seconde)

Dictionnaire de traduction

Suggestions de traduction pour *beheersing*:

Correspondance exacte:

Base lexicale du français

[gestion](#)

Opus, projet de corpus parallèles alignés de Jörg Tiedemann

Même catégorie grammaticale: -

Toutes les suggestions

pertinence de ...
la ... traduction voir des exemples dans ces corpus:

pertinence de ...
la ... traduction voir des exemples dans ces corpus:

**** contrôler	débats Parl. européen	Constitution européenne	soustitres	** contrôle	débats Parl. européen	Constitution européenne	soustitres
** contrôle	débats Parl. européen	Constitution européenne	soustitres	** gestion	débats Parl. européen	Constitution européenne	soustitres
** gestion	débats Parl. européen	Constitution européenne	soustitres	** maîtrise	débats Parl. européen	Constitution européenne	soustitres

Interglot, dictionnaire de traduction en ligne

[Voir](#)

Correspondance approximative:

zelfbeheersing [calme](#) (nom)

actieve leeromgeving Frans voor anderstaligen - lexicon

ALFALEX est un environnement d'apprentissage assisté par ordinateur qui porte sur de nombreux problèmes liés à l'utilisation des mots en français.

[Lire](#) une description détaillée de l'environnement ou [consulter](#) quelques statistiques.

Attention:

- ♦ accès sécurisé au site par Blackboard (étudiants et personnel de la K.U.Leuven uniquement): vous ne pouvez lancer les exercices que si vous avez passé le test;
- ♦ accès libre au site par Blackboard ou de l'extérieur: tous les liens sont activés. Vous n'avez toutefois pas accès à toutes les fonctionnalités.

Dernières informations	
6 octobre 2006	Pour le mode d'emploi de l'environnement, cliquez ici (en néerlandais, avec animations).

Vous pouvez obtenir des informations sur le contenu des exercices en cliquant sur le lien [\[info\]](#).

Les phrases des exercices sont tirées des journaux Le Monde et Le Soir. Les analyses morpho-syntaxiques ont été réalisées à l'aide du logiciel [Cordial Analyseur](#).

le test

[testez](#) rapidement votre niveau de français

les exercices

la formation des mots

- ♦ la [morphologie](#): les terminaisons irrégulières des noms et des adjectifs

[FAQ](#)

ALFALEX is een geautomatiseerde leeromgeving die oefeningen aanbiedt voor verschillende aspecten verbonden aan het gebruik van de woorden in het Frans.

Laatste berichten	
6 oktober 2006	Voor een handleiding van de leeromgeving, klik hier



Opgelet:

- ♦ beveiligde versie via Blackboard (enkel voor studenten en personeel K.U.Leuven): u kan de oefeningen enkel maken als u de test hebt afgelegd;
- ♦ niet beveiligde versie via Blackboard of andere: alle links zijn geactiveerd. Niet alle functionaliteiten zijn echter beschikbaar.

U kan bijkomende informatie verkrijgen over de inhoud van de oefeningen door op de link [\[info\]](#) te klikken.

De zinnen van de oefeningen komen uit de kranten Le Monde en Le Soir. De morfo-syntactische analyses werden uitgevoerd met [Cordial Analyseur](#).

de test

[test](#) snel uw kennis van het Frans

de oefeningen

de woordvorming

- ♦ de [morphologie](#): de onregelmatige uitgangen van de substantieven en de adjectieven

[accueil](#)
[mes scores](#)
[dico personnel](#)

exercices

[morphologie](#)
[collocations](#)

[conjugaison](#)
[synonymes](#)

[dérivation](#)
[schémas](#)
[actanciels](#)

[genre](#)
[traduction](#)
[F > NL](#)

[prépositions](#)
[traduction](#)
[NL > F](#)

Exercice sur les collocations

*** Complétez les cadres à l'aide des informations données entre parenthèses à la fin des phrases. ***

Si vous trouvez la phrase trop difficile, vous pouvez obtenir une nouvelle phrase en cliquant sur le lien [modifier] en fin de phrase.
Vous pouvez obtenir la première lettre de la forme du verbe à compléter en cliquant sur le lien [première lettre] en fin de phrase. Cette aide vous coûte 2 points sur 4.

1 tort ou à raison, ils s'estiment mal-aimés

(verbe support de **tort**, présent)

[\[modifier\]](#) [\[première lettre\]](#)

2 UNION EUROPÉENNE : le programme INFO2000 vient de deux appels à propositions dont la date limite de soumission est fixée au 17 avril.

(verbe support de **appel**, infinitif)

[\[modifier\]](#) [\[première lettre\]](#)

3 Aujourd'hui, certains des gestes théâtraux qui produisent l'effet inverse.

(verbe support de **geste**, présent)

[\[modifier\]](#) [\[première lettre\]](#)

4 On aurait pu croire que les joueurs de deux équipes, dont le sort était déjà scellé depuis la semaine dernière, ne pas la peine d'assurer le spectacle.

(verbe support de **peine**, conditionnel)

[\[modifier\]](#) [\[première lettre\]](#)

5 Il a indiqué qu'il ne pas appel de la décision prise par le préfet Bernard Bonnet de désarmer les policiers municipaux de la ville (Le Monde du 5 août).

(verbe support de **appel**, conditionnel)

[\[modifier\]](#) [\[première lettre\]](#)

6 Certains grands groupes recours à la croissance externe pour compenser l'insuffisance de leur croissance interne.

(verbe support de **recours**, présent)

[\[modifier\]](#) [\[première lettre\]](#)

7 N' pas peur, on est des frères, lui a lancé l'homme aux cheveux longs, très sale, " halouf, on aurait dit un sanglier qui pillait les réserves de semoule ".

(verbe support de **peur**, impératif)

[\[modifier\]](#) [\[première lettre\]](#)

8 Mais la régie, précise-t-on, ne pas un intérêt que pour les enfants.

(verbe support de **intérêt**, conditionnel)

[accueil](#)

[mes scores](#)

[dico personnel](#)

exercices

[morphologie](#)

[collocations](#)

[conjugaison](#)

[synonymes](#)

[dérivation](#)

[schémas](#)
[actanciels](#)

[genre](#)

[traduction](#)
[F > NL](#)

[prépositions](#)

[traduction](#)
[NL > F](#)

Exercice sur les collocations

••• Complétez les cadres à l'aide des informations données entre parenthèses à la fin des phrases. •••

Si vous trouvez la phrase trop difficile, vous pouvez obtenir une nouvelle phrase en cliquant sur le lien [\[modifier\]](#) en fin de phrase.
Vous pouvez obtenir la première lettre de la forme du verbe à compléter en cliquant sur le lien [\[première lettre\]](#) en fin de phrase. Cette aide vous coûte 2 points sur 4.

1 Mais il a tort de croire que les Américains ne partagent pas les graves réserves qu'exprime leur gouvernement à propos de la politique et du comportement de l'Iran.
(verbe support de *tort*, participe passé)
[\[modifier\]](#) [\[première lettre\]](#)

Exercice sur les collocations

••• Complétez les cadres à l'aide des informations données entre parenthèses à la fin des phrases. •••

Si vous trouvez la phrase trop difficile, vous pouvez obtenir une nouvelle phrase en cliquant sur le lien [\[modifier\]](#) en fin de phrase.
Vous pouvez obtenir la première lettre de la forme du verbe à compléter en cliquant sur le lien [\[première lettre\]](#) en fin de phrase. Cette aide vous coûte 2 points sur 4.

1 Il n' pas tort lorsqu'il laisse entendre, sur un ton plaintif, que des dizaines de personnages officiels pourraient être assis à sa place s'ils n'avaient eu la chance de mourir plus tôt.
(verbe support de *tort*, présent)
[\[modifier\]](#) [\[première lettre\]](#)

3.3. HLT Applications

- Machine Translation (statistical & example-based MT)

P. Koehn 2005: 110 SMT-systems trained on Europarl-corpus

Example output Finnish-English:

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 .

- Multilingual Terminology Extraction
- Cross-lingual Information Retrieval (CLIR)
- Training and test data for HLT tools

4. Team & WP's

- DPC Core Research Team

K.U.Leuven Campus Kortrijk:

Piet Desmet, Hans Paulussen & Yulia Trushkina,
Antoine Besnehard & Maribel Montero Perez

HoGent – School of Translation Studies

Willy Vandeweghe, Lieve Macken,
Lidia Rura & Orphée De Clercq

- DPC Research partners

University of Groningen, Radboud University Nijmegen, Tilburg University,
K.U.Leuven, University of Antwerp, University of Ghent

- DPC User group: academic & industrial partners

4. Team & WP's

WP1 - Corpus design

WP2 - Text Normalization

WP3 - Alignment

WP 4 - Linguistic annotation

WP5 - Corpus exploitation

WP 6 - Validation

Programme

10u00 – 10u15

DPC – What?, Why?, For whom?, How?

Piet Desmet

10u15 – 10u30:

Corpus Design

Willy Vandeweghe

10u30 – 11u10:

Acquisition & IPR

Orphée De Clercq & Maribel Montero Perez

11u10 – 11u30:

Coffee break

11u30 – 12u30:

Processing stages – preprocessing,
alignment, annotation

Lieve Macken & Hans Paulussen

12u30 – 13u00:

DEMO Web Interface

Serge Verlinde & Geert Peeters

Designing a Multifunctional Parallel Corpus

Typical Difficulties

- availability of translated data
- quality of the translated material
- proportional availability of translated material for all targeted languages and translation directions

Unidirectional or bidirectional?

- Unidirectional: from language A to language B
- Bidirectional: both from language A to language B and vice versa
→ “additional difficulties” (Olohan 2004:25).

Trilingual corpus

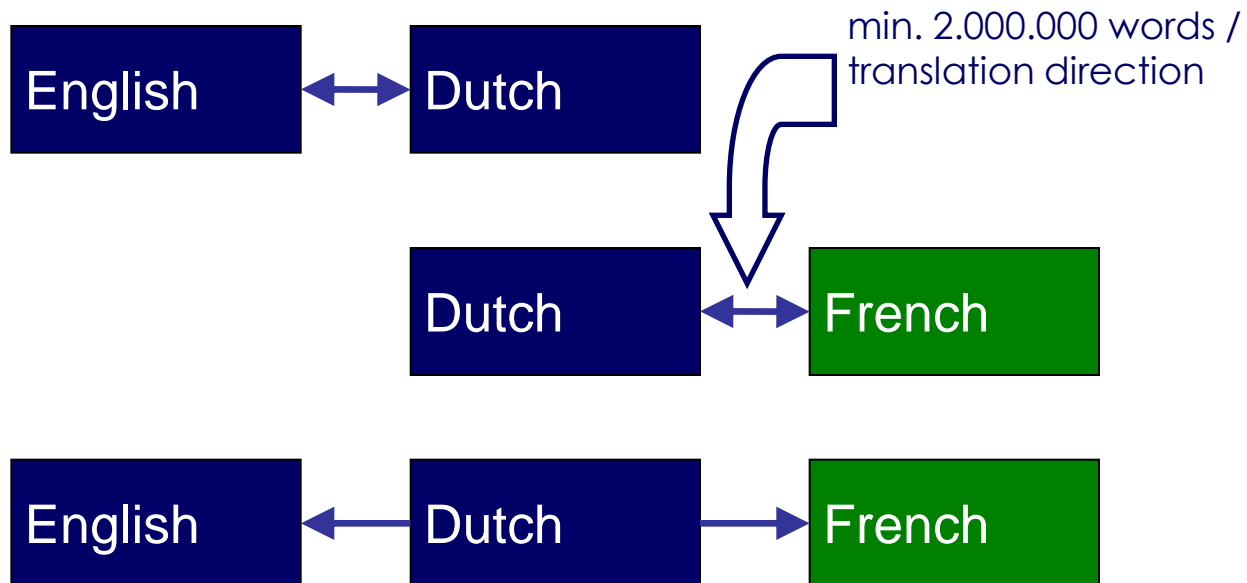
Difficulty of finding parallel data with a version in every chosen language.

“This is one of the reasons why European Union texts are often used” (Koskinen 2000: 55)

Optimising corpus design

- Balancing
- Text type diversity
- Representative corpus samples

Balancing



Text type diversity

Parallel Concordance - [gelet]

... rd in de uitbating van het restaurant. Gelet op de aanwezigheid van de keuken, en ge ...
 ... et op de aanwezigheid van de keuken, en gelet op de verplichting van de overheidsdien ...
 ... ups, de bijwerkingen van applicaties). Gelet op de projecten communit~e, moeten de g ...
 ... ag ontbijt ik (h)eerlijk en biologisch Gelet op het succes van deze actie verleden) ...
 ... de ter beschikking gestelde werknemer. Gelet op de moeilijkheden ondervonden in de s ...
 ... en beschouwd als voltijdse prestaties. Gelet op de gelijkstelling van daglonen met h ...
 ... ck@minsoc.fed.be 4. Informatiesessie. Gelet op de complexiteit van de opdracht, hee ...
 ... e stellen op het vlak van coördinatie. Gelet op de beperkte onderzoeksperiode en het ...
 ... g en andere belangrijke verstrekkingen Gelet op paragraaf 3, punt (2), van artikel 1 ...

... qu'ils étaient pas intéressés par l'exploitation du restaurant. Compte tenu de la présence de la cuisine et de ...
 ... qu'ils étaient pas intéressés par l'exploitation du restaurant. Compte tenu de la présence de la cuisine et de ...
 ... les mises à jours des applications par exemple). Face aux projets communit~e, les communes et plus ...
 ... bio et de l'équitable, je passe à table Forte du succès qu'a rencontré cette action l'année dernière, ...
 ... rémunération réelle du travailleur mis à disposition. Au vu des difficultés rencontrées par le secteur, l'article 5 ...
 ... considérées comme étant des prestations à temps plein. Vu l'alignement des rémunérations journalières sur le RMMMG et ...
 ... dries.gellynck@minsoc.fed.be 4. Séance d'information Vu la complexité du marché, le pouvoir adjudicateur a ...
 ... proposer des améliorations concrètes en matière de coordination. Vu la petitesse de la période de recherche et ...

9 matches French (Standard) - Search word, 1st right Strings matching: gelet

Figure 1. Example of an Administrative Text (DPC)

Representative corpus samples

- Ideally full texts
- If only fragments: variation
 - (text beginnings, central part, end of text)

Corpus Structure

- delimiting the target population of texts
- determining text categories (e.g. text types, genres, topics, etc.)
- finding a way of organising the typology, i.e. designing a corpus taxonomy.

Determining text categories

Definition criteria text categories

- External (situational) vs internal (linguistic)
- Possible categories based on external criteria:
 - topic-based categories
 - established categories
 - basic-level (prototype) categories

Basic level categories

SUPERORDINATE	Mammal
BASIC-LEVEL	Dog/Cat [GENRE]
SUBORDINATE	Labrador/ Siamese [SUBGENRE]

Taxonomy

SUPERORDINATE	Mammal	Literature	Advertising
BASIC-LEVEL	Dog/Cat [GENRE]	Novel/ Poem/ Drama [GENRE]	Advertisement [GENRE]
SUBORDINATE	Labrador/ Siamese [SUBGENRE]	Western/ Romance/ Adventure [SUBGENRE]	Print ad, Radio ad, TV ad, T-shirt ad [SUBGENRE]

DPC typology & structure

1. Fictional literature	1.1 Novels 1.2 Short stories
2. Non-fictional literature	2.1 Essayistic texts 2.2 (Auto)biographies 2.3 Expository non-fictional literature
3. Journalistic texts	3.1 News reporting articles 3.2 Comment articles (background articles, columns, editorials)
4. Instructive texts	4.1 Manuals 4.2 Internal legal documents 4.3 Procedure descriptions
5. Administrative texts	5.1 Legislation 5.2 Proceedings of parliamentary debates 5.3 Minutes of meetings 5.4 Yearly reports 5.5 Correspondence 5.6 Official speeches
6. External communication	6.1 (Self-)presentations of organisations, projects, events 6.2 Informative documents of a general nature 6.3 Promotion and advertising material 6.4 Press releases and newsletters 6.5 Scientific texts

DPC typology & structure

1a. Literature Fictional	1.1 Novels 1.2 Short stories
1b. Literature non-fictional	1.3 Essayistic texts 1.4 (Auto)biographies 1.5 Expository non-fictional literature
2. Journalistic texts	2.1 News reporting articles 2.2 Comment articles (background articles, columns, editorials)
3. Instructive texts	3.1 Manuals 3.2 Internal legal documents 3.3 Procedure descriptions
4. Administrative texts	4.1 Legislation 4.2 Proceedings of parliamentary debates 4.3 Minutes of meetings 4.4 Yearly reports 4.5 Correspondence 4.6 Official speeches
5. External communication	5.1 (Self-)presentations of organisations, projects, events 5.2 Informative documents of a general nature 5.3 Promotion and advertising material 5.4 Press releases and newsletters 5.5 Scientific texts

To conclude

- Bidirectional (2, sometimes 3 languages)
- Mostly full texts
- Guaranteed quality, reliable source
- Representative text diversity
- Taxonomy guaranteeing corpus transparency, navigability and efficient data retrieval

Acquisition & IPR

DPC Workshop
18 September 2009

10 MW

- 5 text types
- 4 translation directions

Ideally

TEXT → RESEARCHER → AUTHOR → AGREEMENT

TEXT TYPES

- Administrative texts
- External communication
- Instructive texts
- Journalistic texts
- Literature

→ *Published*
→ *Translation division*

MATRIX

- Words included in DPC = → *10 MW*

Legislation, proceedings, reports, minutes, speeches

- Public institutions
- Large Belgian companies
- Europarl

→ No real difficulties

Self-presentation, document general nature, promotion, newsletters and press releases.

- Large Belgian companies
- Institutions

→ No real difficulties

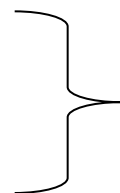
Manuals, internal legal documents, procedure descriptions

- Multinationals
- Institutions

→ Source language?

- **Administrative** = public institutions
- **External** = large Belgian companies
- **Instructive** = multinationals

- Journalistic texts
- Literature



text = income



News articles, background articles, columns

- Newspapers
- Publications of institutions

→ ≠ players

TEXT → RESEARCHER → AUTHOR → AGREEMENT

ORIGINAL TEXT → JOURNALIST → LAWYER → NEWSPAPER
 → LEGAL DEPARTMENT → ADAPTATIONS → DLU →
 → NEGOTIATIONS → ADAPTED AGREEMENT

TRANSLATION → TRANSLATOR → NEWSPAPER → ≠
 TRANSLATORS → LAWYER → ADAPTATIONS → DLU →
 NEGOTIATIONS → ADAPTED AGREEMENT

Fiction (novels), non-fiction (essayistic texts, expository works)

- Public institutions
- Publishing houses

➔ Many difficulties especially with fiction

6 → 5 text types

- negotiations started = beginning of project
- ≠ contact persons ≠ terms
- commercial aspect
- Dutch Language Union
- success ±

→ Future = reverse this process

Matrix

- > 50%
- > 99%
- > 100%
- > 110%

Text type	SRC -> TGT	NL	EN	FR	Total	%
Administrative texts	EN -> NL	255,455	246,137	0	501,592	100.26
	FR -> NL	320,383	0	333,997	654,380	130.88
	NL -> EN	247,729	255,372	0	503,101	100.62
	NL -> FR	278,745	0	300,231	578,976	115.80
SubTotal		1,102,012	501,509	634,228	2,237,749	111.89
External Communication	EN -> NL	278,515	272,460	0	550,975	110.19
	FR -> NL	233,277	0	250,604	483,881	96.78
	NL -> EN	246,448	255,634	0	502,082	100.42
	NL -> FR	237,116	0	264,801	501,917	100.38
SubTotal		1,031,227	563,165	531,148	2,125,540	106.28
Instructive texts	EN -> NL	330,511	327,773	0	658,284	131.66
	FR -> NL	40,487	0	42,017	82,504	16.50
	NL -> EN	19,011	20,696	0	39,707	7.94
	NL -> FR	110,278	0	115,034	225,312	45.06
	XN-F	64,840	0	79,780	144,620	28.92
	XNE-	279,176	274,926	0	554,102	110.82
	XNEF	159,493	166,875	188,867	515,235	103.05
SubTotal		1,003,796	790,270	425,698	2,219,764	110.99
Journalistic texts	EN -> NL	262,768	264,900	0	527,668	105.53
	FR -> NL	225,833	0	195,314	421,147	84.23
	NL -> EN	250,580	259,764	0	510,344	102.07
	NL -> FR	322,420	0	245,438	567,858	113.57
SubTotal		1,065,796	524,664	426,088	2,027,017	101.35
Literature	EN -> NL	148,488	143,185	0	291,673	58.33
	FR -> NL	184,262	0	188,045	372,307	74.46
	NL -> EN	346,802	361,140	0	707,942	141.59
	NL -> FR	321,183	0	342,323	663,506	132.70
SubTotal		1,000,735	504,325	530,368	2,035,428	101.77

MIN. 3 providers/cell



170,000 words

Matrix

- > 50%
- > 99%
- > 100%
- > 110%

Text type	SRC -> TGT	NL	EN	FR	Total	%
Administrative texts	EN -> NL	255,155	246,137	0	501,292	100.26
	FR -> NL	320,383	0	333,997	654,380	130.88
	NL -> EN	247,729	255,372	0	503,101	100.62
	NL -> FR	278,745	0	300,231	578,976	115.80
SubTotal		1,102,012	501,509	634,228	2,237,749	111.89
External Communication	EN -> NL	278,515	272,460	0	550,975	110.19
	FR -> NL	233,277	0	250,604	483,881	96.78
	NL -> EN	246,448	255,634	0	502,082	100.42
	NL -> FR	237,116	0	264,801	501,917	100.38
SubTotal		1,031,227	563,165	531,143	2,125,540	105.28
Instructive texts	EN -> NL	330,511	327,773	0	658,284	131.66
	FR -> NL	40,487	0	42,017	82,504	16.50
	NL -> EN	19,011	20,696	0	39,707	7.94
	NL -> FR	110,278	0	115,034	225,312	45.06
	XN-F	64,840	0	79,780	144,620	28.92
	XNE-	279,176	274,926	0	554,102	110.82
	XNEF	159,493	166,875	188,867	515,235	103.05
SubTotal		1,003,796	790,270	425,693	2,219,764	111.99
Journalistic texts	EN -> NL	262,768	264,900	0	527,668	105.53
	FR -> NL	225,833	0	195,314	421,147	84.23
	NL -> EN	250,580	259,764	0	510,344	102.07
	NL -> FR	322,420	0	245,438	567,858	113.57
SubTotal		1,065,796	524,664	426,083	2,027,017	101.35
Literature	EN -> NL	148,488	143,185	0	291,673	58.33
	FR -> NL	184,262	0	188,045	372,307	74.46
	NL -> EN	346,802	361,140	0	707,942	141.59
	NL -> FR	321,183	0	342,323	663,506	132.70
SubTotal		1,000,735	504,325	530,368	2,035,428	101.77

10 MW

IPR

DPC Workshop
18 September 2009

IPR agreements: why?

- Copyright clearance
 - All DPC text samples
 - Use of texts
 - Accessibility and availability of data
 - Protection of intellectual/economic property rights

- HLT-Agency (TST-centrale):
 - Validation of agreements
 - Contact
 - Archive
 - Distribution
 - ...

TST-CENTRALE)))
voor taal- en spraaktechnologie

→ Different players → different types

- IPR for commercial use
- IPR for publishers
- IPR short version
- E-mail or letter with permission

- Right to use, process, modify texts
- Right to store texts as part of DPC
- Right to convert/integrate DPC in other systems
- Right to grant sub-licences for research, education, product development
- Right to grant sub-licences for commercial purposes, texts are not recognizable

= IPR for **commercial** use

+

- For non-commercial purposes: texts partially recognizable in the new products to be developed
- no competition

||

IPR for **publishers**

Types

	Recognizable	Not recognizable
Commercial purposes	O	X Standard C2 /for publishers/ Short version
Non-commercial purposes	X Standard C2/ Short version	X C2 for publishers

- IPR for commercial use
- IPR for publishers
- IPR short version
 - Facilitate IPR procedure
 - Data free available on website – not public domain

- IPR for commercial use
- IPR for publishers
- IPR short version: use
- E-mail with permission

Only in exceptional cases:

- Texts are already publicly accessible
- Texts are not a substantial part of the corpus

- The Netherlands – Belgium
- Refusals
- No reaction

DPC & IPR: numbers

IPR for commercial use 15

IPR for commercial use – short version 12

No contract necessary 6

E-mail or letter with permission 16

IPR for publishers 7

Total: 53

IPR: examples

Data provider	IPR agreement
Campuskrant	<i>IPR for Commercial use</i>
FTPN Namur	
De Post	
Vlaamse Overheid	
RIZIV	
FOD Justitie	
Quarterly Fortis	
Roularta	<i>IPR for publishers</i>
Ons Erfdeel	
Transmed	
Bosch	<i>Short version of IPR</i>
Melexis	
Electrolux	<i>E-mail permission</i>
Eli Lilly	
European Parliament Speeches	<i>No contract necessary</i>
Speeches from the throne Beatrix	

Questions?

- DPC: what?, why?, for whom?, how?
- Corpus Design
- Acquisition & IPR

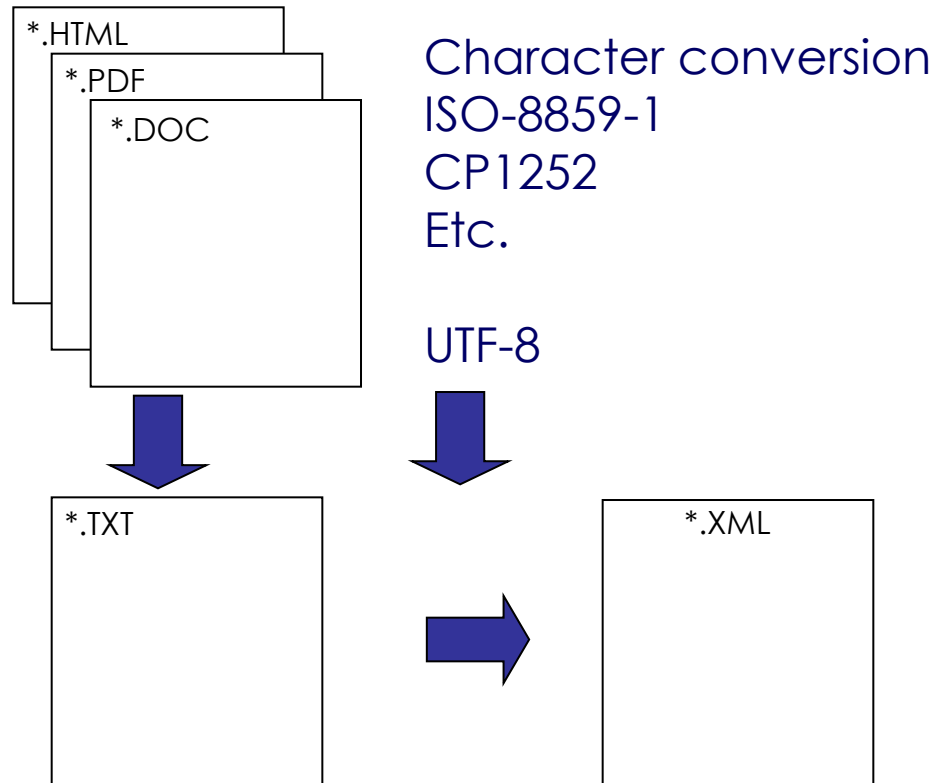
Coffee break

- 11u10 – 11u30

Processing stages

Preprocessing, alignment and
annotation

- **Preprocessing**
- Metadata & matrix
- Processing principles
- Alignment
- Linguistic annotation
- Terminology



	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	<u>NUL</u> 0000	<u>STX</u> 0001	<u>SOT</u> 0002	<u>ETX</u> 0003	<u>EOT</u> 0004	<u>ENQ</u> 0005	<u>ACK</u> 0006	<u>BEL</u> 0007	<u>BS</u> 0008	<u>HT</u> 0009	<u>LF</u> 000A	<u>VT</u> 000B	<u>FF</u> 000C	<u>CR</u> 000D	<u>SO</u> 000E	<u>SI</u> 000F
10	<u>DLE</u> 0010	<u>DC1</u> 0011	<u>DC2</u> 0012	<u>DC3</u> 0013	<u>DC4</u> 0014	<u>NAK</u> 0015	<u>SYN</u> 0016	<u>ETB</u> 0017	<u>CAN</u> 0018	<u>EM</u> 0019	<u>SUB</u> 001A	<u>ESC</u> 001B	<u>FS</u> 001C	<u>GS</u> 001D	<u>RS</u> 001E	<u>US</u> 001F
20	<u>SP</u> 0020	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{		}	~	<u>DEL</u> 007F
80																
90																
A0	<u>NBSP</u> 00A0	ı	ç	£	*	¥		§	¨	@	ª	«	¬	-	®	—
B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ø	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ISO-8859-1
latin1

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	<u>NUL</u> 0000	<u>STX</u> 0001	<u>SOT</u> 0002	<u>ETX</u> 0003	<u>EOT</u> 0004	<u>ENQ</u> 0005	<u>ACK</u> 0006	<u>BEL</u> 0007	<u>BS</u> 0008	<u>HT</u> 0009	<u>LF</u> 000A	<u>VT</u> 000B	<u>FF</u> 000C	<u>CR</u> 000D	<u>SO</u> 000E	<u>SI</u> 000F
10	<u>DLE</u> 0010	<u>DC1</u> 0011	<u>DC2</u> 0012	<u>DC3</u> 0013	<u>DC4</u> 0014	<u>NAK</u> 0015	<u>SYN</u> 0016	<u>ETB</u> 0017	<u>CAN</u> 0018	<u>EM</u> 0019	<u>SUB</u> 001A	<u>ESC</u> 001B	<u>FS</u> 001C	<u>GS</u> 001D	<u>RS</u> 001E	<u>US</u> 001F
20	<u>SP</u> 0020	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{		}	~	<u>DEL</u> 007F
80	€ 20AC	⋮	/ 201A	f 0192	# 201E	… 2026	† 2020	‡ 2021	ˆ 02C6	‰ 2030	Š 0160	< 2039	Œ 0152	⋮	Ž 017D	⋮
90	⋮	\ 2018	/ 2019	" 201C	" 201D	• 2022	– 2013	— 2014	˜ 02DC	™ 2122	š 0161	> 203A	œ 0153	⋮	ž 017E	ÿ 0178
A0	<u>NBSP</u> 00A0	ı 00A1	ı̇ 00A2	£ 00A3	* 00A4	¥ 00A5	ı̈ 00A6	§ 00A7	¨ 00A8	@ 00A9	ª 00AA	« 00AB	¬ 00AC	– 00AD	® 00AE	— 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C0	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D0	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E0	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F0	ø 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

CP1252

Alphabet soup: ≤

Unicode	Unicode name	UTF-8	CP1252	Mac Roman	Latin1	Latin9
U0153	LATIN SMALL LIGATURE OE	œ	9C	CF	-	BD
U20AC	EURO SIGN	€	80	DB	-	A4
U2030	PER MILLE SIGN	‰	89	E4	-	-
UFB02	LATIN SMALL LIGATURE FL	fl	-	DF	-	-
U2264	LESS-THAN OR EQUAL TO	≤	-	B2	-	-
U00BC	VULGAR FRACTION ONE QUARTER	¼	BC	-	BC	BC
U2026	HORIZONTAL ELLIPSIS	...	85	C9	-	-

- All processing: *encoded* CP1252

```
2264 LESS-THAN OR EQUAL TO \u2264
```

```
In clinical trials with olanzapine in over  
5000 patients with baseline non-fasting  
glucose levels \u2264 7.8 mmol/l, the  
incidence of non-fasting plasma glucose  
levels \u2265 11 mmol/l (suggestive of  
diabetes) was 1.0%, compared to 0.9% with  
placebo.
```

- Distribution: UTF8

- Preprocessing
- **Metadata & matrix**
- Processing principles
- Alignment
- Linguistic annotation
- Terminology

- Text-related data
 - Language, author, text type, domain
- Translation-related data
 - Translation direction, translation modality
 - Alignment tool & alignment quality
- Annotation-related data
 - Annotation tools & annotation quality

Metadata vertical

	A	B
1	Text-related data	Values
2	1. Language	<i>EN (UK)</i>
3	2. Author/translator	Emile Wennekes
4	3. Text unit title ¹	Loud Chords and Calm Moments.Louis Andriessen
5	4. Publishing info	
6	magazine/journal title	The Low Countries (14)
7	publisher	vzw Ons Erfdeel
8	ISBN/ISSN	
9	date of publication	2006
10	original date of publication ²	
11	place of publication	Rekkem: Ons Erfdeel
12	original place of publication ²	
13	info on previous editions	
14	editor	
15	article number	
16	page of the article in the magazine	260-210
17	keywords	
18	class of the article	
19	5. Intended outcome	<i>written to be read</i>
20	6. Text type	<i>Non-fictional literature</i>
21	7. Text subtype	<i>Expository works of a general nature</i>
22	8. Domain	<i>Culture</i>
23	9. Keywords	<i>---Culture</i>
24	10. Copyright/IPR-agreement	<i>light version</i>
25	11. Type of institution	<i>profit</i>
26	12. Intended audience	<i>broad external audience</i>
27	Translation-related data	
28	14. Original text & language	<i>NL</i>
29	15. Translated text & language	<i>EN</i>
30	16 Intermediate text & language	<i>choose</i>
31	17. Translation mode	<i>human</i>
32		
33	Statistics	
34	18. Number of words	2.155
35	19. Number of sentences	
36		
37	Extra	
38	20. Sub-documents	

Metadata horizontal

	A	B	C	D
1	dpc-ons-000402-en	EN (UK)	Emile Wennekes	Loud Chords and Calm Moments.Louis Andriessen,Composer
2	dpc-ons-000402-nl	NL (BE)	Emile Wennekes	Louis Andriessen, gelauwerd componist van luide slagakkoorden en verst
3	dpc-ons-000403-en	EN (UK)	Raf De Bruyn / Translated by Sheila M. Dale	Arrival & Departure Travelling to and from the Low Countries
4	dpc-ons-000403-nl	NL (BE)	Raf De Bruyn	Arrival & Departure Travelling to and from the Low Countries -
5	dpc-ons-000404-en	EN (UK)	Hans Aarsman / Translated by Gregory Ball	Mystery on the March On Dirk Braeckman's Photos
6	dpc-ons-000404-nl	NL (BE)	Hans Aarsman	Het mysterie rukt op. Over de foto's van Dirk Braeckman
7	dpc-ons-000405-en	EN (UK)	Manfred Sellink / Translated by Laura Watkinson	Out & about with Bruegel
8	dpc-ons-000405-nl	NL (BE)	Manfred Sellink	Met Bruegel op stap
9	dpc-ons-000476-en	EN (UK)	Wim Daniëls	Talking Dutch
10	dpc-ons-000476-fr	FR (BE)	Wim Daniëls	Vous avez dit néerlandais?
11	dpc-ons-000476-nl	NL (BE)	Wim Daniëls	Spraakmakend Nederlands
12	dpc-ons-000477-fr	FR (BE)	Saskia de Bodt – Frank Hellemans	Taverne du passage – Peintres et écrivains néerlandais en Belgique
13	dpc-ons-000477-nl	NL (BE)	Saskia de Bodt – Frank Hellemans	Taverne du passage – Nederlandse schilders en schrijvers in België
14	dpc-ons-001027-fr	FR (FR)	Adriaan van Dis / D. Cunin	Langue libertine
15	dpc-ons-001027-nl	NL (NL)	Adriaan Van Dis	Vrijtaal
16	dpc-ons-001028-fr	FR (FR)	Abdelkader Benali / J.M. Jacquet	un enterrement qui ressemblera à une noce
17	dpc-ons-001028-nl	NL (NL)	Abdelkader Benali	De begrafenis die zal lijken op een bruiloft
18	dpc-ons-001029-fr	FR (FR)	Reinier Salverda	Les langues dans notre vie
19	dpc-ons-001029-nl	NL (NL)	Reinier Salverda	De talen in ons leven

Consistency checks

```

1 dpc-ons-000402-en|EN (UK)|Emile Wennekes |Loud Chords and Calm Moments.Louis Andriessen,Co
2 dpc-ons-000402-nl|NL (BE)|Emile Wennekes |Louis Andriessen, gelauwerd componist van luide
3 dpc-ons-000403-en|EN (UK)|Raf De Bruyn / Translated by Sheila M. Dale|Arrival & Departure
4 dpc-ons-000403-nl|NL (BE)|Raf De Bruyn|Arrival & Departure Travelling to and from the Low
5 dpc-ons-000404-en|EN (UK)|Hans Aarsman / Translated by Gregory Ball|Mystery on the March
6 dpc-ons-000404-nl|NL (BE)|Hans Aarsman|Het mysterie rukt op. Over de foto's van Dirk Braec
7 dpc-ons-000405-en|EN (UK)|Manfred Sellink / Translated by Laura Watkinson|Out & about wit
8 dpc-ons-000405-nl|NL (BE)|Manfred Sellink|Met Bruegel op stap|The Low Countries (14)|vzw O
9 dpc-ons-000476-en|EN (UK)|Wim Daniëls|Talking Dutch||vzw Ons Erfdeel|90-75862-74-1|2005||
10 dpc-ons-000476-fr|FR (BE)|Wim Daniëls|Vous avez dit néerlandais? ||vzw Ons Erfdeel|90-75
11 dpc-ons-000476-nl|NL (BE)|Wim Daniëls|Spraaakmakend Nederlands||vzw Ons Erfdeel|90-75862-7
12 dpc-ons-000477-fr|FR (BE)|Saskia de Bodt - Frank Hellemans|Taverne du passage - Peintr
13 dpc-ons-000477-nl|NL (BE)|Saskia de Bodt - Frank Hellemans|Taverne du passage - Nederl

```

```

14 dpc-ons-001027-
15 dpc-ons-001027-
16 dpc-ons-001028-
17 dpc-ons-001028-
18 dpc-ons-001029-
19 dpc-ons-001029-

```

```
## 26 Original text & language
```

```

awk -F'|' '{ print $26 }' META/group2/meta-???.csv |
sort | uniq -c |
awk '{ tot+=$1; print } END { print "Total:", tot }'

```

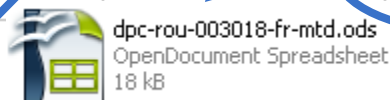
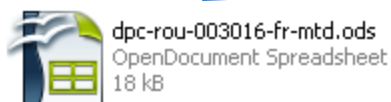
```

7
1383 EN
690 FR
2006 NL
732 unknown
Total: 4818

```

Metadata compilation

	A	B	C	D	E
1	magazine/journal title				
2	publisher	ROU			
3	date of publication	april			
4					
5	14. Original Text&Lang	File Name (Du)	File Name (Fr)	2. Author/translator	dpc name
89					
90	Dutch	BN0804-0851	BF0804-0791	Guido Eekhaut	dpc-rou-003016
91	Dutch	BN0804-0852	BF0804-0792		
92	Dutch	BN0804-0853	BF0804-0793		
93					
94	Dutch	BN0804-0881	BF0804-0821	Hans Hermans	dpc-rou-003017
95	Dutch	BN0804-0882	BF0804-0822		
96	Dutch	BN0804-0883	BF0804-8823		
97	Dutch	BN0804-0884	Geen overeenkomst		
98	Dutch	BN0804-0885	BF0804-8824		
99					
100	French	BN0804-0911	BF0804-0851	Catherine Pleeck	dpc-rou-003018
101	French	BN0804-0912	BF0804-0852		
102					
103	Dutch	BN0804-0931	BF0804-0871	Thierry Debels	dpc-rou-003019
104	Dutch	BN0804-0932	BF0804-0872		
105					



Matrix /1

Text type	SRC -> TGT	NL	EN	FR	Total	%
Administrative texts						
	EN -> NL	255,155	246,137	0	501,292	100.26
	FR -> NL	320,383	0	333,997	654,380	130.88
	NL -> EN	247,729	255,372	0	503,101	100.62
	NL -> FR	278,745	0	300,231	578,976	115.80
<i>SubTotal Administrative texts:</i>		1,102,012	501,509	634,228	2,237,749	111.89
External Communication						
	EN -> NL	278,515	272,460	0	550,975	110.19
	FR -> NL	233,277	0	250,604	483,881	96.78
	NL -> EN	246,448	255,634	0	502,082	100.42
	NL -> FR	237,116	0	264,801	501,917	100.38
	XNE-	21,679	20,118	0	41,797	8.36
	XNEF	14,192	14,953	15,743	44,888	8.98
<i>SubTotal External Communication:</i>		1,031,227	563,165	531,148	2,125,540	106.28

DPC Statistics

Group	NL	EN	FR	Total
ABY	24603	23327	0	47930
ARC	68041	69730	0	137771
BAL	60548	60688	0	121236
BCO	114820	113052	0	227872
BEK	6193	6341	0	12534
BEV	73938	70533	0	144471
BMM	35757	34957	15623	86337
BOS	179498	185817	203985	569300
CAM	19425	19658	0	39083
DNS	3395	3478	3679	10552

BOS

Group	NL	EN	FR	Total
External Communication	13689	14724	15373	43786
Instructive texts	165809	171093	188612	525514
Total	179498	185817	203985	569300

BOS: External Communication

File	wordcount	Vertaalrichting	Interm. lg	In	Cl
dpc-bos-001379-en	14724	NL -> EN, FR	Unknown	*	
dpc-bos-001379-fr	15373	NL -> EN, FR	Unknown	*	
dpc-bos-001379-nl	13689	NL -> EN, FR	Unknown	*	
Total	43786				

- Flexible
 - G1: accepted data
 - G2: fridge
 - G3: unknown or unresolved
- Dynamic
 - crontab
 - hyperlinks
- Regulating factors
 - metadata required
 - file naming constraints

- Phase 1
 - follow-up acquisition & cleaning
 - provider oriented focus
- Phase 2
 - follow-up data processing
 - process oriented focus

dpc-named files

tekstleverancier

dpc-nummer

taal

dataverwerking

dpc-kok-001321-en-in.txt

dpc-kok-001321-en-cl.txt

dpc-kok-001321-en-sen.txt

dpc-kok-001321-en-pps.txt

dpc-kok-001321-en-pps-ok.txt

dpc-kok-001321-en-al-vanH.txt

dpc-kok-001321-en-an-ok.txt

dpc-kok-001321-en-an.txt

dpc-kok-001321-en-tok-ans-ok.txt

dpc-kok-001321-en-tok-ans.txt

Processing stages

cl	clean
pps	sentence splitting
pps-ok	sentence splitting verified
seg	paragraph alignment
sen-align	sentence alignment
tok	tokenisation
tok-ok	tokenisation verified
ann	linguistic annotation
ann-ok	linguistic annotation verified

Focus on provider

vla

```

--dpc_named
  |--ann
  |--seg
  |--sen-align
  |   |--nl-en
  |   `--nl-fr
  `--tok
  
```

pos

```

--dpc_named
  |--ann
  |--seg
  |--sen-align
  |   |--nl-en
  |   `--nl-fr
  `--tok
  
```

kok

```

--dpc_named
  |--ann
  |--seg
  |--sen-align
  |   |--nl-en
  |   `--nl-fr
  `--tok
  
```

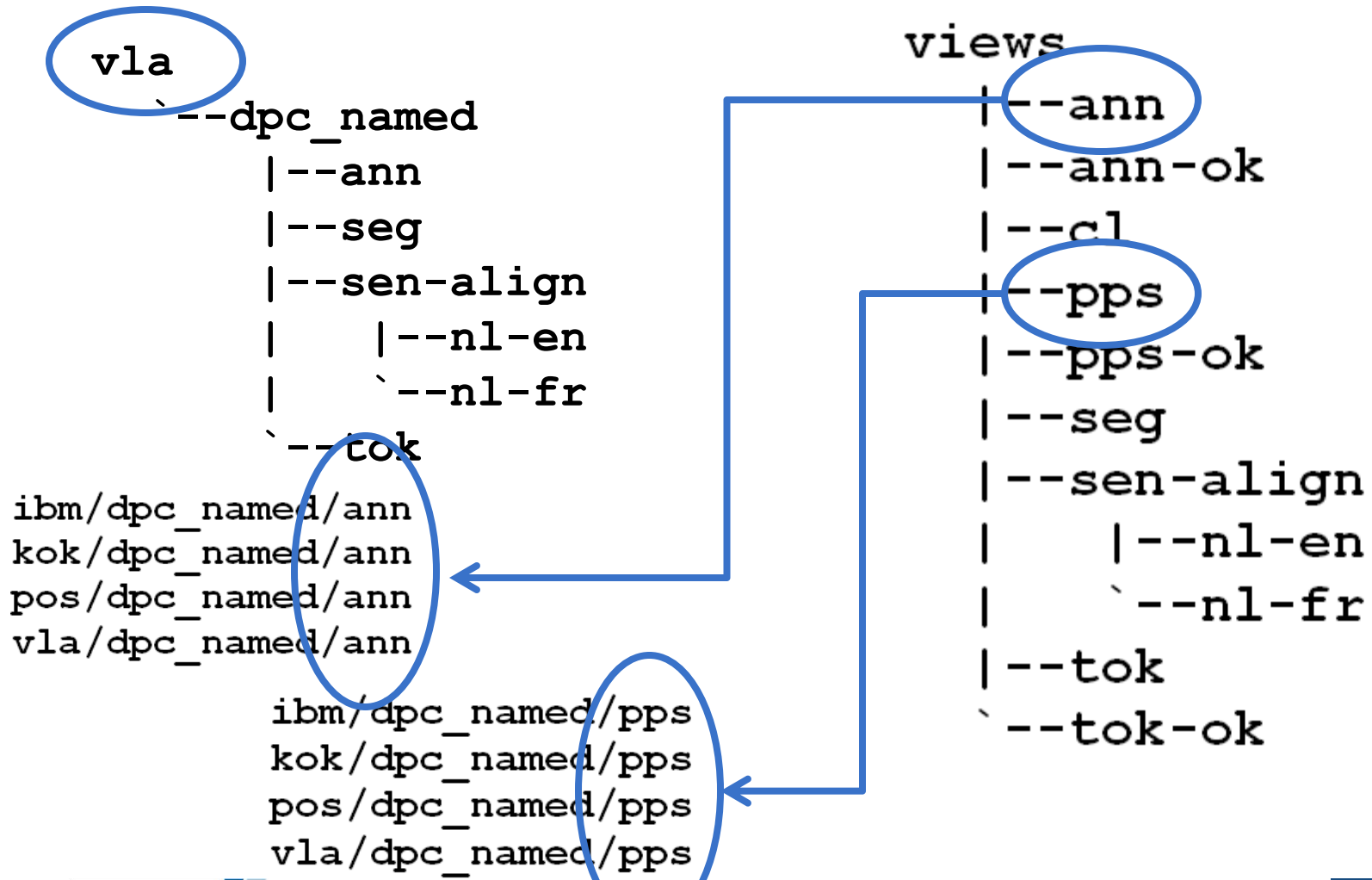
ibm

```

--dpc_named
  |--ann
  |--seg
  |--sen-align
  |   |--nl-en
  |   `--nl-fr
  `--tok
  
```

- Shift from provider to process follow-up
- Direct access to processed data necessary
- Adapted structure facilitates
 - Data processing follow-up
 - Production process
 - Detecting processing errors

Process access /1



Process access /2

views

```
| --ann  
| --ann-ok  
| --cl  
| --pps  
| --pps-ok  
| --seg  
| --sen-align  
|   |--nl-en  
|   `--nl-fr  
|--tok  
`--tok-ok
```

dpc-bal-001236-en-an.txt
dpc-bal-001237-en-an.txt
dpc-bal-001238-en-an.txt
dpc-bal-001239-en-an.txt
dpc-bos-000800-en-an.txt

dpc-bal-001236-en-pps.txt
dpc-bal-001236-nl-pps.txt
dpc-bal-001237-en-pps.txt
dpc-bal-001237-nl-pps.txt
dpc-bal-001238-en-pps.txt

- Preprocessing
- Metadata & matrix
- **Processing principles**
- Alignment
- Linguistic annotation
- Terminology

Maximal quality, minimal effort

1M corpus = 10% of whole corpus

- all processing steps manually checked
- error analysis on manually verified data
- (improve tools)
- development of spot-checking heuristics

9M corpus

- spot-checking or automatic control procedures

Tools used:

- Adapted version of CPAN Sen.pm module

Improvement tool:

- Updating abbreviations list
- Rules for abbreviations + digit/roman number
 - No. , art., ca.
- Rule if capitalized frequent word occurs after full stop
 - The, This, These, Any, ...
 - 20 to 140° C. The table shows when

Spot-checking heuristic

- False paragraph breaks: paragraphs starting with lowercase letter or digit (preceded by brackets)
- Missing spaces between full stop and capital letter
- Estimated correctness: 99%



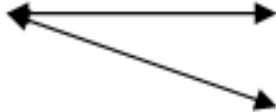


- Preprocessing
- Metadata & matrix
- Processing principles
- **Alignment**
- Linguistic annotation
- Terminology

Sentence Alignment /1

What?

- Align corresponding sentences
- Accepted alignments
 - 1:1 1:many
 - many:1 many:many
 - 1:0 0:1
- No crossing alignments are allowed
 - Crossing alignments are grouped into many:many alignments

Sentence Alignment /2

<i>English text</i>	<i>Alignment Links</i>	<i>Dutch text</i>
What do we see in a face?		Wat staat er in de trekken van een gelaat allemaal te lezen?
Not only its colour, shape and expression, but also the character, experience, hopes and fears which have moulded it.		Niet alleen vorm, kleur en expressie, maar ook karakter, ervaring, hoop en vrees.
The fiercer its battles, the deeper its sorrows, the more hectic its joys, the greater are the traces of life it bears.		Hoe heviger het strijd heeft geleverd, hoe meer leed het heeft gedragen. Hoe uitbundiger het heeft gejuicht, des te dieper staan de sporen van het leven erin gegroefd.
Moods, changing from one moment to the next, show also, while first impressions, acquired at the moment of meeting are never lost.		Ook de voortdurend wisselende gemoedsgesteldheid kun je ervan aflezen, terwijl je de indruk van een eerste ontmoeting nooit meer vergeet.
Such is the face of Flanders.		Zo is het gelaat van Vlaanderen.

Tools used:

– **Vanilla aligner**

- Danielsson & Ridings 1997
- Sentence length-based statistical approach (Gale & Church 1997)
- Requires paragraph alignment

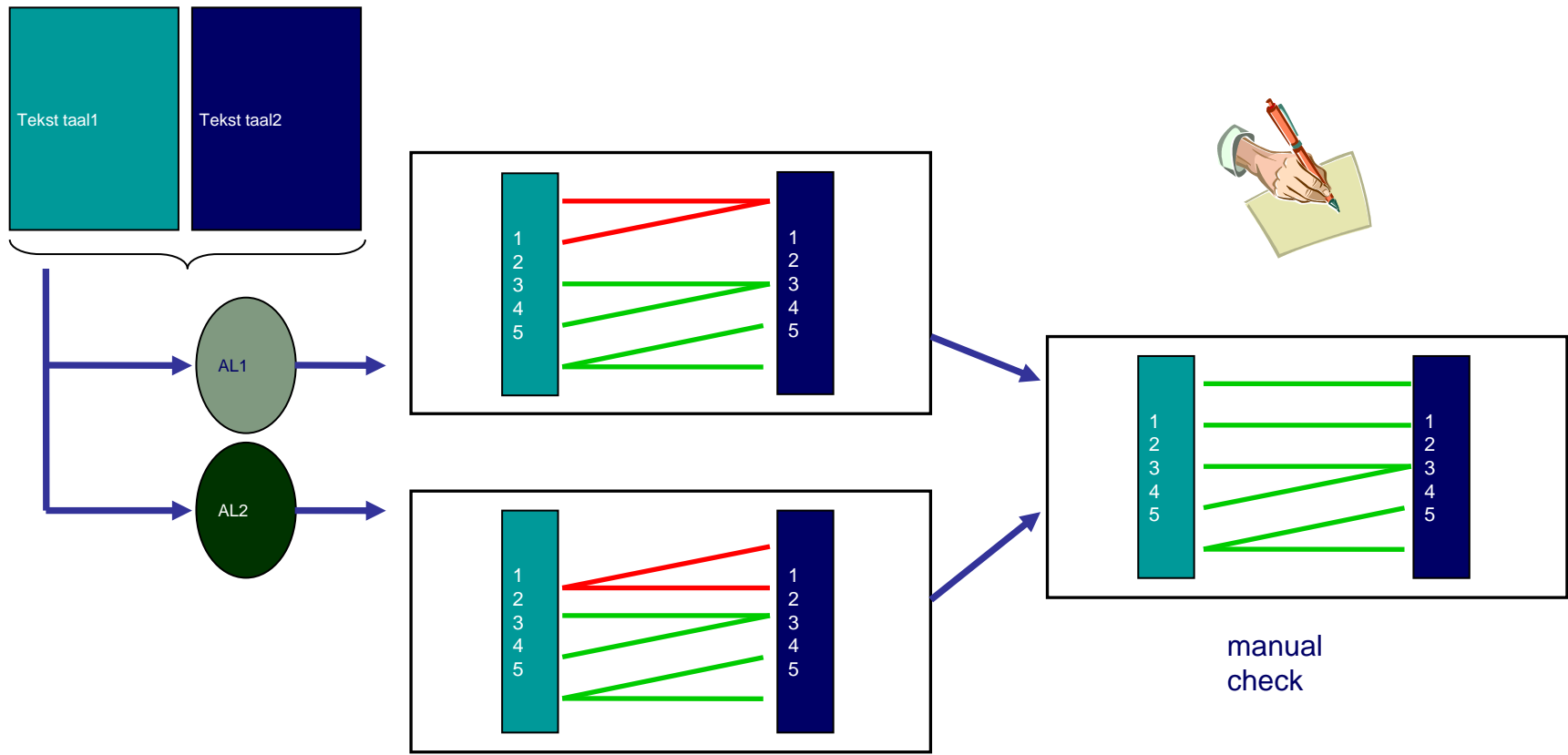
– **Smooth Injective Map Recognizer (SIMR)**

- Melamed 1997
- Based on word correspondences and cognates
- Used NI-Translex bilingual dictionaries

– Microsoft Bilingual Aligner (Moore 2002)

- Three-step hybrid approach
 - 1) Sentence-length based alignment
 - 2) Train statistical word alignment model on sentences aligned with high probability in step 1
 - 3) Realign based on word alignments
- Generates only 1:1 alignments

Alignment merge



Performance Dutch-English

Data set: 16563 sentence pairs

Alignment characteristics:

0:1	253	1:1	15197
1:0	261	n:m	852

Results for different aligners:

Vanilla: correct 15269 wrong 917

GMA: correct 14957 wrong 1228

MS: correct 14294 wrong 669 missed 1823

Results for combined aligners:

All three aligners : correct: 13700 (82 %) wrong: 218 (1.3%)

Performance Dutch-French

Data set: 15667 sentence pairs

Alignment characteristics:

0:1	165	1:1	14088
1:0	158	n:m	1256

Results for different aligners:

Vanilla: correct 14552 wrong 952

GMA: correct 13709 wrong 1634

MS: correct 12266 wrong 1456 missed 2737

Results for combined aligners:

All three aligners : correct: 11901 (76%) wrong: 133 (<1%)

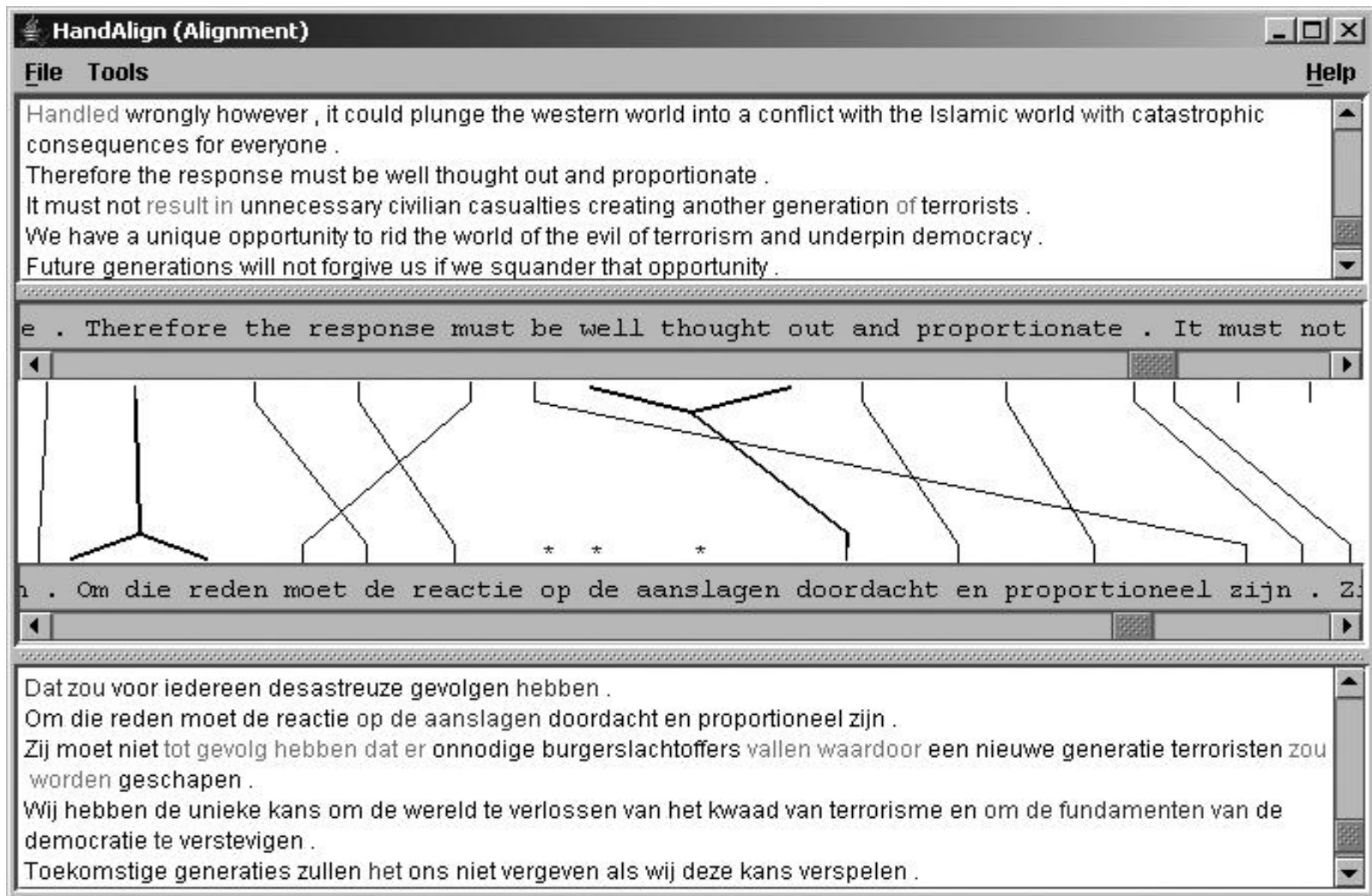
Spotchecking heuristics

- Paragraph alignment
 - Manually verified paragraph alignment if
 - Different number of paragraphs
 - Large discrepancies in length of corresponding paragraphs
- Sentence alignment
 - Manually verified sentence pairs that were not linked by all three alignment tools

Vizualisation tool Paraconc

ParaConc - [Alignment Dutch (Standard) - English (United Kingdom) (dpc-eup-000020-nl-al-three.txt - dpc-eup-000020-en-al-three.txt)]...	
File Alignment Search Frequency Window Info	
<p>XXXXXXXX Het alternatief van ECOFIN draait evenwel de klok terug. [eos] Het druist in tegen de budgettaire evidentie, tegen geheel de logica van het grondwettelijk verdrag en tegen onze jarenlange inspanningen om onze werkzaamheden via meer parlementair toezicht democratischer te maken. [eos]</p>	<p>XXXXXXXX The ECOFIN alternative is a step backwards; it flies in the face of the budgetary evidence, the whole logic of the Constitutional Treaty, and our long-term efforts to democratise our work through increased parliamentary scrutiny. [eos]</p>
<p>Ik verwerp niet zomaar enkele ideeën uit de recentste voorstellen, maar ik moet jullie wel waarschuwen voor de grote gehechtheid van het Parlement aan deze kwestie omdat begrotingsbevoegdheden - de controle over de geldbeugel - tot de kern van de parlementaire democratie behoren, niet alleen in Europa maar ook in al onze lidstaten. [eos]</p>	<p>I do not reject out-of-hand some of the ideas contained in the most recent proposals, but I have to alert you to the depth of feeling in Parliament on this issue, because budgetary powers, the control of the purse-strings, go to the heart of parliamentary democracy, not only in Europe but also in all our Member States. [eos]</p>
<p>Hierbij wil ik nog opmerken dat tijdens onze ontmoetingen vorige week met nationale parlementsleden van de Conventie een consensus bestond over het feit dat de rechten van het Parlement, met name wat de begrotingsprocedure betreft, niet in het gedrang mogen komen. [eos]</p>	<p>I note that when we met with national parliamentarians from the Convention last week, there was a consensus that Parliament's rights, particularly in the budgetary procedure, must not be jeopardised. [eos]</p>
<p>Mijn eerste taak bestaat er uiteraard in de positie van het Parlement te verdedigen: zijn politieke voorrechten, maar ook zijn efficiëntie. [eos]</p>	<p>My primary duty is clearly to defend Parliament's position, its political prerogatives, but also its efficiency. [eos]</p>
<p>Dit Parlement, met zijn zware werklast op wetgevend gebied en zijn toezichhoudende rol op de begroting en de Commissie, moet echter ook een bestuurbaar orgaan blijven. [eos]</p>	<p>This Parliament, with its heavy legislative workload and its role of scrutiny on the budget and over the Commission, must also be a manageable body. [eos]</p>
<p>Deze werkzaamheden kunnen enkel worden verricht door een parlement, niet door een volkerencongres. [eos]</p>	<p>This work can only be done by a Parliament, not a Congress of Peoples. [eos]</p>
<p>736 zetels - zowel in de voorstellen van de Conventie als in het Italiaanse compromis - benadert enorm dicht de grenzen van wat organisatorisch haalbaar is voor een werkbaar Parlement. [eos]</p>	<p>736 seats - in the Convention proposals and in the Italian compromise - is pretty well at the limits of what is organisationally operational for a working Parliament. [eos]</p>
<p>Jullie weten uiteraard dat er naar de zetelverdeling moet worden gekeken, met name om rekening te houden met de bekommernissen van de kleinste lidstaten, en misschien ook om de bekommernissen van anderen over hun vertegenwoordiging in andere instellingen te erkennen, maar dit mag niet leiden tot opoffering van het beginsel van degressieve proportionaliteit, en evenmin tot toevoeging van zetels aan een reeds bijzonder hoog totaal. [eos]</p>	<p>You may well perceive that there is a need to look at the seat distribution, particularly to take account of the concerns of the smallest Member States, and perhaps to recognise the concerns of others about their representation in other Institutions, but this must not involve sacrificing the principle of degressive proportionality, nor should it add seats to what is already a very high total. [eos]</p>
<p>XXXXXXXX Dit is niet louter een pleidooi van het Parlement. [eos] Wil het Europese project slagen, dan moeten de Europese instellingen efficiënt werken. [eos]</p>	<p>XXXXXXXX This is not simply pleading from the Parliament; for the European project to work, the European Institutions must be efficient. [eos]</p>
<p>De zetels in het Parlement mogen niet worden gebruikt als inzet in een goktent. [eos]</p>	<p>Seats in the Parliament should not be used as a playing chip in a gambling saloon. [eos]</p>

Sub-sentential alignment



Small portion of the NI/En corpus: 25,000 words

- En-NI Instructive texts 7,536 words
- En-NI Journalistic texts 7,706 words
- NI-En Journalistic texts 10,480 words

- Preprocessing
- Metadata & matrix
- Processing principles
- Alignment
- **Linguistic annotation**
- Terminology

What?

- Sentence split into sequences of words
- Punctuation not belonging to the word form stripped off

Tools used:

- Nl & En: ILK Tokenizer (D-coi)
- Fr: adapted version of the TreeTagger preprocessing script

Errors corrected in ILK tokenizer

- abbreviations followed by digits:
 - approx . 6 g, Fig . 6
- unknown measures
 - 29.000m³ -> 29.000 m³
- telephone number split:
 - 059 - 23.43.01 -> 059-23.43.01

No spot-checking

- Obtained error rate reduction high enough to proceed automatically

What?

- Assigning part-of-speech code and base form to each token

Tools used:

- Memory-based PoS tagger/lemmatizer
- Treetagger

Tag set:

- Penn Treebank
- Coarse-grained tag set
- 45 distinct tags

Example

It it PRP

is be VBZ

tremendously tremendously RB

important important JJ

Performance tools (299,000 tokens):

- Precision MBSP PoS tagging: 96.2%
- Precision MBSP lemmatization: 98.1%
- Precision Treetagger PoS tagging: 95.2%
- Precision Treetagger lemmatization: 97.6%

Spotchecking: combined MBSP & Treetagger:

- Total number of identical PoS codes: 94.5%
- PoS Precision of identical PoS tags: 98.6%
- Total number of identical lemmata: 96.9%
- PoS Precision of identical lemmata: 99.4%

- Number of tokens to check (different PoS or different lemma): $\pm 10 \%$

Tools used:

- ILK Combitagger (D-Coi)

Tag set:

- CGN PoS tag set
- Fine-grained tag set
- 316 distinct tags

Example

Het LID(bep,stan,evon) het
kind N(soort,ev,basis,onz,stan) kind
krijgt WW(pv,tgw,met-t) krijgen
daardoor BW() daardoor
geheugenproblemen N(soort,mv,basis) ...

Performance tools (79,229 tokens):

- Precision Combitagger PoS tagging: 95.3%
 - Only main category: 97.7%
- Precision Combitagger lemmatization: 96.9%
- Spot-check based on probabilities

Tools used:

- Treetagger + FLEMM

Tag set:

- GRACE tag set
- Fine-grained tag set
- 312 distinct tags

- Two cycle workflow:
 - Basic TreeTagger tagset
 - Lemmata
 - Probabilities
 - LIMSI tagset (based on GRACE)
- Spot check based on annotated version on output of both tagging cycles

cette	Dd-fs--	Déterminant démonstratif, féminin, singulier	ce
idée	Ncfs	Nom commun, féminin, singulier	idée
est	Vmip3s-	Verbe principal, indicatif, présent, 3, singulier	être
au	Sp+Da-ms-d	Préposition général	au
centre	Ncms	Nom commun, masculin, singulier	centre
d'	Sp	Préposition général	de
un	Da-ms-i	Déterminant article, masculin, singulier, indéfini	un
vaste	Afpms	Adjectif qualificatif, positif, masculin, singulier	vaste
débat	Ncms	Nom commun, masculin, singulier	débat

	A	B	C	D	E	F	G	H	I	
1	perc	percCode	catCode	lemmaCode	token	tag GRACE	correctie	vertaling	tag TT	lemma
2					<sent id="dpc-med-000674-fr-sen.txt.p.1.s.1.">					
3	1				"	F		Ponctuation	PUN(cit)	"
4	1				La	Da-fs-d		Déterminant article, féminin, singulier, défini	DET(ART):Da3fs---	le
5	1				reconnaissance	Ncfs		Nom commun, féminin, singulier	NOM:Nc-s--	reconnaiss
6	1				visuelle	Afpfs		Adjectif qualificatif, positif, féminin, singulier	ADJ:A--fs--	visuel
7	1				des	Sp+Da-mp-d		Préposition général	PRP(det):Sp+Da--p--d	du
8	1				visages	Ncmp		Nom commun, masculin, pluriel	NOM:Nc-p--	visage
9	1				"	F		Ponctuation	PUN(cit)	"
10					</sent>					
11					<sent id="dpc-med-000674-fr-sen.txt.p.2.s.1.">					
12	1				"	F		Ponctuation	PUN(cit)	"
13	1				Prof	Ncms		Nom commun, masculin, singulier	NOM:Nc-s--	prof
14	1				Raymond	Npms		Nom propre, masculin, singulier	NAM:Np----	Raymond
15	0.58	P01	CN01		Bruyer	Npms		Nom propre, masculin, singulier	ABR	bruyer
16	1				-	F		Ponctuation	PUN	-
17	0.68	P02	CN01		Université	Ncfs		Nom commun, féminin, singulier	ABR	université
18	1				catholique	Afpfs		Adjectif qualificatif, positif, féminin, singulier	ADJ:A---s--	catholique

code	description
CXX	The GRACE and TT tags belong to different categories (irrespective of percentage limit)
CN01	Noun check: The GRACE tag is a noun, the TT tags is not a noun.
CN02	Both GRACE and TT tags are noun, but percentage is less than default percentage limit (0.6)

- Preprocessing
- Metadata & matrix
- Processing principles
- Alignment
- Linguistic annotation
- **Terminology**

External validation

- Suitability test for terminology extraction
- In cooperation with Xplanation
- Created Gold Standard
 - Medical domain
 - Trilingual Dutch/English/French
 - » 615 aligned sentences (\pm 31,000 words)
 - Financial domain
 - Bilingual Dutch/English
 - » 571 aligned sentences (\pm 17,000 words)
 - Bilingual Dutch/French
 - » 355 aligned sentences (\pm 15,000 words)

Approaches

- Ask domain experts to manually indicate all terms
- Problems
 - Lack of consensus, low agreement scores
 - No domain experts available
- Use existing term banks and dictionaries as reference
 - IATE, Euramis, specialized dictionaries

Resulting Gold Standard

– Medical Domain

- NI/En/Fr
 - 450 terms

– Financial Domain

- NI-En
 - 400 terms
- NI-Fr
 - 350 terms

- Full text resource
 - annotated files
 - XML: D-COI
 - sentence aligned files
 - XML: TEI P5
- Web search interface
 - Parallel KWIC concordance
 - Monolingual & bilingual
 - Queries
 - Simple & extended