

1- Multifunctional & multilingual

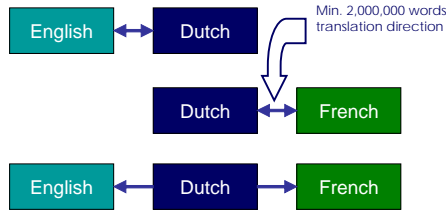
In the DPC project a 10-million-word, high-quality, sentence-aligned parallel corpus for Dutch-English and Dutch-French is being compiled.

The DPC project aims at the creation of a multifunctional resource to satisfy the needs of a diverse group of potential users:

- machine translation (MT)
- computer-assisted translation (CAT)
- computer-assisted language learning (CALL)
- research in contrastive linguistics and translation studies.

2- Corpus design

The DPC consists of two language pairs: Dutch-English and Dutch-French and is bi-directional. A part of the corpus will be trilingual and will contain Dutch texts translated into both English and French.



The corpus will offer a great variety of text types from different domains. To guarantee the quality of the text samples, most of them come from published materials or from companies or institutions working with a professional translation division. Besides language pair and translation direction, the DPC will also be balanced with respect to text type.

Commercial publishers

- Fictional literature, e.g. novels, short stories
- Non-fictional literature, e.g. essays
- Journalistic texts, e.g. news articles

Institutions

- Instructive texts, e.g. user manuals
- Administrative texts, e.g. meeting minutes
- External communication, e.g. promotion material, newsletters

3- Data acquisitions

DPC Matrix

The DPC matrix is a document that gives an instant up-to-date overview of the current state of data collection and processing. The matrix is a table comprising 4 columns (for each translation direction) and six rows (for every text type). Each cell should contain ideally 417.000 words. Collected data are put in respective cells, the amount of words per cell is stated in word count. Up till now, 2 cells of the matrix contain already 100%:
 -Administrative texts (Dutch → French)
 -External Communication (Dutch → French)

Providers – amount of data

Contacts with data providers started in Semester 1 and will continue throughout the whole project period. Up till now, numerous attempts have been made to contact data providers, most of these attempts were successful and led to collaboration agreements (Table 1).
 The DPC-corpus will contain 10,000,000 words by the end of the project. Up till now, the corpus consists of 4,000,000 words.

4- IPR Agreements

In order to make the corpus available for the whole research community, copyright clearance is being obtained for all samples included in the corpus.

Types of IPR-agreements:

- **IPR for Commercial use**
 → the standard DPC contract
 → French and English translations, validated by TST
- **IPR for publishers**
 → Based on standard IPR and details commercial and non-commercial use
 → French and English translations, validated by TST
- **Short version of the Standard IPR**
- **Letter or E-mail permission**
 → E-mail or letter with permission to use the data
 → Only for publicly accessible texts that are not a substantial part of the corpus

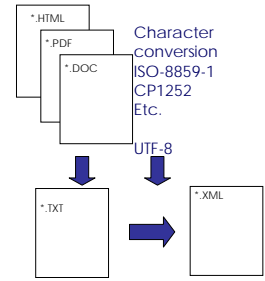
Up till now, 30 IPR agreements have been signed by data providers. The following table gives an overview of the data providers and the type of contract they signed:

Data provider	IPR agreement
Campuskrant FTPN Namur De Post Sociale verzekeringsbank Vlaamse Overheid BMM DNS FOD Justitie NMBS Group Quarterly Fortis IBM Gazette FOD Sociale zekerheid	<i>IPR for Commercial use</i>
Roularta Ons Erfdeel Transmed	<i>IPR for publishers</i>
Bosch Melexis	<i>Short version of IPR</i>
Electrolux Eli Lilly Ford Dodge Animal Health CisBio Orphan Europe Miscellaneous	<i>E-mail permission</i>
European Parliament Speeches Europarl Speeches Balkenende Speeches Kok Speeches from the throne Beatrix De Kamer De Senaat	<i>No contract Necessary (texts can be published, subject to acknowledgement of the source)</i>

Data Providers & IPR

5- Text Normalization

The text samples are collected from different sources in a wide range of text and character formats. A preliminary task consists in cleaning and normalizing the text and standardizing the character encoding. Tokenization and sentence splitting are also part of this preparatory work.



6- Alignment

The text samples are automatically aligned at sentence level. In order to improve alignment quality and to ease manual quality control, the alignment output of three aligners will be merged and the problematic alignments will be verified manually.

7- Annotation PoS & Lemma

The corpus will be PoS tagged and lemmatized. In order to keep annotation consistent with similar Dutch corpus projects, the PoS annotation scheme of D-COI will be used for Dutch.

- NL – DCOI
- EN – MBT
- FR – Treetagger/ Cordial

8- Manual verification

10 % of the corpus (1,000,000 words) will be checked manually. During the summer period of 2008, job students will manually check the cleaning, alignment and PoS tagging of the 10% corpus. The output of the manual verification will be used to validate the quality of the different tools, which will be optimized and adjusted where necessary.

9- Web Interface

A web Search Interface will be developed, enabling both simple and more elaborate queries, based on pattern matching of words and annotation labels. The query types include patterns like the following (SL = source language and TL = Target Language):

- Find word X in SL
- Find word X in SL that corresponds (or does not correspond) to word Y in TL
- Find word X having PoS = Q in SL
- Find word X having PoS = Q in SL corresponding to word Y having PoS = P in TL

PROJECT TEAM

- K.U. Leuven Campus Kortrijk
 Piet Desmet
 Hans Paulussen
 Julia Trushkina
 Maribel Montero Perez
- Hogeschool Gent
 Willy Vandeweghe
 Lieve Macken
 Lidia Rura

EXTERNAL VALIDATION

- Formal validation by CST (Copenhagen)
- Suitability test by Xplanation

DURATION: 2006-2009