CILL 32.1-4 (2006), 269-285

DUTCH PARALLEL CORPUS : A MULTIFUNCTIONAL AND MULTILINGUAL CORPUS

H. PAULUSSEN[*], L. MACKEN[+], J. TRUSHKINA[*], P. DESMET[*], W. VANDEWEGHE[+]
([*]) K.U.Leuven Campus Kortrijk,
([+]) LT[3], University College Ghent

0. INTRODUCTION

Nowadays, text corpora play an important role in language research and all fields involving language study, including theoretical and applied linguistics, language technology, translation studies and CALL (Computer Assisted Language Learning). Multilingual corpora, especially translated corpora, are not always readily available for Dutch. Much depends on the private initiative of individuals, and the data are often restrictedly available. The DPC-project (Dutch Parallel Corpus), which is carried out within the STEVIN program (Odijk *et al.* 2004), intends to fill the gap for this type of corpora for Dutch. This paper gives an overview of the DPC project. First, an overview and a discussion is given of the main parallel corpora containing Dutch. Then the DPC project is described, focusing on those aspects that make the DPC different from existing parallel corpora. Finally, the choice of an XML based format is explained.

1. DUTCH IN PARALLEL CORPORA

The aim of the DPC-project is to develop a high-quality state-of-the-art, multilingual corpus, with Dutch as central language. The DPC mainly differs from other existing parallel corpora in the following five aspects: quality control, level of annotation, balanced composition, availability and Dutch kernel. This section first describes the parallel corpora with a Dutch component and then discusses each of the five aspects separately.

1.1. State of the art

There are a number of available multilingual corpora that contain a Dutch component. However, many of the multilingual corpora are comparable corpora[1], or contain only few translated texts. MULTEXT[2] (Ide and Véronis 1994) and PAROLE (Kruyt 1998, de Does and van der Voort/van der Kleij 2002) are typical examples of projects that focus on harmonization of multilingual corpus standards, but they contain no translations for the Dutch text samples.

Table 1 gives an overview of the main presently available parallel corpora containing a Dutch component[3]: the Namur Corpus (Paulussen 1999), the European Corpus Initiative Multilingual Corpus I (ECI/MCI) corpus[4], the MLCC corpus[5], the Scania corpus (Tjong Kim Sang 1996), the Oslo Multilingual Corpus[6] (Johansson 2002a, Johansson 2002b), the Europarl corpus (Koehn 2005), and the OPUS[7] corpus (Tiedemann and Nygaard 2004). The corpora are sorted according to their creation period.

For each corpus, the number of Dutch words contained in the corpus is presented in the second column of the table. Except for the Europarl corpus and MLCC, the Dutch components of the parallel corpora contain less than 1,000,000 words. All the corpora listed have Dutch, French and English parallel samples, but the numbers in the table do not indicate which Dutch samples have been aligned with their English and/or French corresponding text samples.

The third column of the table provides details on domains of the corpora data. The Namur corpus contains both fiction and non-fiction (Unesco Courier and Debates of the European Parliament). Debates of the European Parliament make up two other corpora of the list: the MLCC corpus and the Europarl corpus. The ECI/MCI corpus represents a collection of EC Esprit program announcement

---

1 Comparable corpora contain texts in two or more languages on the same domain, but the texts are no translations; a parallel corpus contains translated texts.

2 MULTEXT contains a parallel component (MULTEXT JOC), but only for the following five languages: English, German, Italian, Spanish and French. Whenever Dutch is mentioned in the MULTEXT project, reference is made to the closely related MLCC project, which contains indeed a Dutch parallel component. Both MULTEXT and MLCC are part of MLAP, the European "Multilingual Action Plan" of the nineties.

3 There are a number of other projects on parallel corpora mentioning Dutch, but the information is unclear or ambiguous: e.g. PEDANT, ETAP (Borin 1999).

4 The ECI/MCI corpus contains 21,527,223 words of multilingual data, but only a small portion is parallel data (214,210 words). See http://www.elsnet.org/resources/eciCorpus.html

5 See http://www.elda.org/catalogue/en/text/W0023.html

6 See http://www.hf.uio.no/ilos/OMC/

7 OPUS contains also the Europarl corpus, which gives a total of 30,074,511 Dutch words in OPUS.

texts. The Scania corpus is compiled of Scania truck manuals, whereas the OPUS corpus consists of OS software manuals.

Corpus name	**Size in words**	**Domains**	**Aligned**	**Markup**	**PoS tagged**
Namur	700,000	Fiction + Non Fiction (Unesco Courier + Debates of the European Parliament)	P	custom	-
ECI/MCI	25,000	EC Esprit program announcement text	-	TEI	-
MLCC	7,100,000	Debates of the European Parliament	-	TEI	-
Scania	216,424	Scania Truck manuals	S	TEI	-
OMC	170,000	Fiction	S	TEI	-
Europarl	29,188,340	Debates of the European Parliament	S	XML	-
OPUS	886,171	OS software manuals	S	XCES	Yes

Table 1: Main parallel corpora available with Dutch component

The fourth column of the table indicates whether the corpora are aligned and, if yes, on which level: "**P**" stands for paragraph alignment, "**S**" stands for sentence alignment, "**-**" stands for no alignment. The Namur corpus is aligned at paragraph level. The ECI/MCI and MLCC corpora are not aligned at all. The remaining corpora are aligned at sentence level.

The fifth column gives information on the markup of the corpora. The Namur corpus uses only a customized markup. The ECI/MCI and MLCC corpora are the first two corpora in which XML markup is used. More specifically, the TEI standard is used for those two corpora, whereas OPUS uses the XCES standard. Both XCES and TEI are XML protocols specifically written for corpus annotation. Note that the Europarl uses XML, without further specification of XCES or TEI[8].

The last column of the table shows that, apart from OPUS, none of the parallel corpora has any systematic encoding of PoS tags.

[8] Both XCES and TEI are described further under section 3 of this paper.

1.2. Quality control

The development of a high-quality state-of-the-art multilingual corpus of reasonable size is a challenge. The existing parallel corpora are either very large (hence lacking quality assurance) or smaller in size. The Europarl corpus, covering more than 29 million words for Dutch alone, is a typical example of a large-scale parallel corpus. This type of parallel corpora is certainly useful for statistical analysis, but the alignment quality can no longer be verified in detail, which can be a drawback for many other applications. Also in the context of machine translation (where statistical data are favoured), a more qualitative resource would be very welcome to improve the results of the statistical tools. CALL applications using parallel corpora as resource of authentic text will also benefit from a qualitative parallel corpus such as DPC[9].

In order to guarantee corpus quality, a considerable part of the DPC corpus is checked manually at different levels, including sentence splitting, linguistic annotation and alignment. A quality label is used to mark the level of verification. The introduction of a fine-tuned system of quality labels improves the selection of corpus samples considerably.

1.3. Level of annotation

Apart from sentence boundaries, all parallel corpora in Table 1 (except OPUS) lack any form of linguistic annotation. The DPC corpus is being sentence-aligned, PoS-tagged and lemmatized. The annotation and linguistic processing are produced by state-of-the-art tools. For Dutch, we adhere to the D-COI conventions as much as possible, strengthening the standards[10]. For English and French we adhere to internationally accepted standards, as defined by EAGLES and similar guidelines. Since Dutch is the central language, the annotation schemes of the other languages have to be compatible with the Dutch part.

1.4. Balanced composition

Another important drawback of the existing parallel corpora is their lack of text type balance. Most of the corpora shown in Table 1 cover a small set of domains or text types, mainly focusing on European Commission texts. For

---

9 An application illustrating the usefulness of parallel corpora in a CALL application is the NEDERLEX project, which resulted in a web reading tool for Dutch using a Dutch-French parallel corpus showing aligned paragraphs (Deville *et al.* 2004).

10 The D-Coi project is a preparatory project which aimed to produce a blueprint and the tools needed for the construction of a 500-million-word reference corpus of contemporary written Dutch. Cf. http://lands.let.ru.nl/projects/d-coi/

example, the MLCC parallel corpus only covers a selection of the Debates of the European Parliament. The parallel part of the MLCC corpus only contains texts from the Official Journal of the European Commission. Table 2[11], giving an overview of the subcorpora in the OPUS corpus (sorted by number of words in Dutch[12]), shows that OPUS only consists of open source software manuals and extracts from the European Parliament[13]. The EU ACQUIS parallel corpus, which has recently been compiled, is solely devoted to European legal texts (Erjavec *et al.* 2005).

Corpus	**EN**	**FR**	**NL**
EuroParl	28,842,367	33,238,913	29,188,340
KDE	2,238,452	1,067,751	476,807
EUconst	164,697	177,162	167,945
PHP	522,603	382,407	146,540
KDEDoc	41,521	419,241	94,879
OpenOffice	478,654	496,780	0

Table 2: Number of words (EN, FR and NL) in the OPUS corpus

There is a great need for more diversity in the types of texts compiled. Paulussen (1999) has shown that some meanings of prepositions and particles are only found in specific types of text. This result was based on the Namur corpus, which covers both fiction and non-fiction. Macken (2007) examined the problem of translational correspondence in different text types (user manuals, press releases and proceedings of plenary debates) and showed that this correspondence is harder to pinpoint in text types adopting a more free translation style. The need for diversity is particularly important for applied linguistic studies, including the development of CALL applications. The DPC therefore contains texts from a wide range of text types (fiction and non-fiction), and diverse domains.

---

11 OpenOffice, KDE (K Desktop Environment: a graphical desktop environment for Unix workstations), KDEDoc and PHP refer to software manuals. EuroParl and EUconst refer to documents from the European Parliamant.

12 For the naming conventions of the language names in Table 2, we use the two letter codes defined by the ISO 639-2 standard which is generally applicable for internet applications. This explains why NL is the abbreviation for Dutch. See also: http://www.loc.gov/standards/iso639-2/

13 The European Parliament extracts are borrowed from the Europarl corpus.

1.5. Availability

The availability of corpora is often problematic. In some cases, the compilation of a corpus is only possible within the context of a PhD thesis (cf. the Namur corpus). In other cases, the corpus is only available within the private company that compiled the corpus. For example, the Scania corpus "*(...) is unlikely to ever become available, since the material is 'commercial in confidence'*."[14] In order to maximize research on parallel corpora, the DPC will be made available to the research community via the Agency for Human Language Technologies (the TST-centrale)[15].

1.6. Dutch kernel

A final drawback of the parallel corpora available is the minor position of Dutch. For example, the OMC contains almost 170,000 words of Dutch translations, but no Dutch source texts[16]. In the case of the software manuals (cf. OPUS), too, many of the Dutch texts are translations from English or other languages. Even if it is true that there is more translation from English into Dutch than the other way around, it is important for language study in general and translation studies in particular to have representative samples where Dutch is the source language. The DPC will consist of two bidirectional bilingual parts and one trilingual part (see Table 3).

EN <-	**NL** -> FR
EN <->	**NL**
	NL <-> FR

Table 3: DPC translation directions

2. DUTCH PARALLEL CORPUS

In comparison with the parallel corpora described in the previous section, the DPC project intends to compile a parallel corpus for Dutch that will offer added value not yet present or minimally present in the existing parallel corpora. Moreover, the approach followed will result in a qualitative corpus, which will also be very useful for corpus exploitation which is not limited to the automatic

---

14 http://spraakbanken.gu.se/pedant/parabank/parabank.html

15 The copyright issues are being solved in close collaboration with the TST-Centrale. See also section 2.1.3 IPR.

16 http://www.hf.uio.no/ilos/OMC/English/Subcorpora.html

processing of the data. The following subsections focus on corpus design and corpus data processing of DPC[17].

2.1. Corpus design

The design principles of the DPC were based on two sources: the information available about other parallel corpus projects, and the analysis of requirements stated by a predefined group of possible users who represent specialists in linguistics and language technology, which was carried out within the DPC project.

To identify the requirements of the user group with respect to corpus design, a questionnaire has been composed in close collaboration with language experts from a research partner group. The questionnaire analysis confirmed a strong need for a freely available parallel corpus with Dutch as a kernel language. The analysis has also shown that the quality of text materials as well as the quality of alignments and linguistic annotations are crucial for the users in corpus applications. The users opted for a high variety of text types and rich metadata and, in general, stated that inclusion of full texts is not a necessary condition for them as long as fragments of different text types are present.

Based on the user requirements analysis, motivated choices have been made regarding the balancing criteria, text typology, sampling criteria, and kind and degree of annotations and required metadata. An overview of the different criteria of the corpus design are presented below. Further details are presented in Macken *et al.* (2007).

2.1.1. Languages and translation directions

As stated earlier, the DPC contains the language pairs Dutch-English and Dutch-French and is bi-directional (Dutch as a source and a target language). A part of the corpus is trilingual, consisting of parallel texts in Dutch, English and French (see Table 3). A proportional distribution of text material between language pairs and translation directions is envisaged. For this purpose a target of minimally 2 million words per translation direction has been set.

[17] The DPC-project is carried out within the STEVIN program and runs from 2006 to 2009.

2.1.2. Text type and providers

The corpus is designed to represent as wide a range of translated Dutch texts as possible. In order to get a well-balanced corpus, texts are selected from different domains in compliance with the requirements of the user group.

The DPC corpus will have a balanced composition not only as far as translation directions are concerned but with respect to the text types as well. The data in the corpus originates from two main sources:

- *commercial publishers*, i.e. organisations whose income depends entirely on their publishing activities such as publishing houses and news agencies
- *institutions*, i.e. governmental en non-governmental organisations as well as private enterprises whose income does not directly come from the publishing business, who do not usually sell their texts as such but use them for other purposes, e.g. information, advertisement, instruction etc.

This division was used to separate the text material into two big groups according to the type of text provider.

Text type	**Text provider**
Fictional literature Non-fictional literature Journalistic texts	Commercial publishers
Instructive texts Administrative texts External communication	Institutions

Table 4: DPC text types

Each group has been subsequently divided into several text types but the criteria for this division are not of the same nature. Those coming from commercial publishers are established genres, i.e. groups of works characterized by a particular form, style, tone, content and purpose. The DPC includes the following genres: literature (both fiction and factual) and the journalistic genre. The institution texts were divided on the basis of their function and purpose: they instruct, document, inform and/or persuade. Table 4 summarizes the text types and providers of the DPC project[18].

[18] See also Macken *et al.* (2007)

2.1.3. IPR

In order to make the corpus accessible for the whole research community, copyright clearance is being obtained for all samples included in the corpus. The license agreements needed to guarantee accessibility and to protect the intellectual and economic property rights of the author and publishers of the texts are being developed in close collaboration with the Agency for Human Language Technologies (TST-centrale).

2.1.4. Metadata

The DPC metadata list consists of three groups: text-related data, translation-related data and annotation-related data.

The first group includes information on the text: language, author and/or translator, title, publishing information, intended outcome of the text (written to be read, or written to be spoken, or written reproduction of spoken language), on text type and topic, copyright information and statistical information (number of tokens, words, sentences and paragraphs).

The second group—translation-related data—indicates the translation direction (original, translated and intermediate texts) and points to other language versions of the same text. It also notes how the text was translated (human translation, translation by a human using translation memory or machine translation corrected by a human) and includes information on alignment tool and alignment quality.

The last group describes the additional annotation of the text. It provides details on tools used for tokenization, PoS tagging, lemmatization and syntactic annotation and the quality of the annotation steps.

2.2. Corpus data processing

The data received from providers come in different formats and need to be brought into conformity with the DPC standard. The unification procedure includes four steps. The following text normalization steps prepare data for further processing (linguistic annotation and alignment):

- conversion of texts to txt-format;
- assigning documents a unique standardized name and grouping documents if necessary;
- normalization of character encoding;
- cleaning the data:

    - content removal (tables of contents, tables, indexes, footnotes, headers and footers, images)
    - clarification of the structure if necessary (e.g. add tags for titles, epigraphs, chapters; group poem lines divided by vertical bars in one paragraph;
- sentence splitting;
- tokenization.

The texts are encoded in conformity with the TEI standards, adapted for aligned sentences. The texts will be stored in two ways: text files (for full text analysis and text interchange) and a relational database (for web queries). Characters are normalized to the Unicode standard UTF8. Only when certain tools require a different character set (e.g. ISO 8859-1) an intermediate character conversion is used temporarily.

2.2.1. Alignment

In sentence alignment, for each sentence of a source language text, an equivalent sentence or sentences of a target language text are found. The sentences linked by the alignment procedure represent translations of each other in different languages.
The following alignment links are legitimate in the DPC project:

- 1:1 (one sentence in a source language is aligned with one sentence in a target language);
- 1:many (one sentence in a source language is aligned with two or more sentences in a target language);
- many:1 (two or more sentence in a source language are aligned with one sentence in a target language);
- many:many (two or more sentence in a source language are aligned with two or many sentence in a target language);
- 0:1 (no alignment links for a sentence in a target language);
- 1:0 (no alignment links for a sentence in a source language).

Zero alignments and many-to-many alignments are accepted in exceptional cases: Zero alignments are created when no translation can be found for a sentence of either the source or the target language, i.e. when a corresponding part of text is missing in the other language.

Many-to-many alignments are legitimate in two cases: overlapping alignments and crossing alignments. Overlapping alignments are cases of asymmetric sentence splitting in the two languages. For example, in Table 5, a source language text and a target language text both consist of two sentences: S_1, S_2 and S'_1, S'_2, respectively.

Source language text	Target language text
S_1: A, B, C;	S'_1: A', B'
S_2: D, E	S'_2: C', D', E'

Table 5: Overlapping alignments

Both sentence pairs in the two languages contain five elements A-E and A'-E' such that A' is a translation of A, B' is a translation of B, etc. S_1 and S'_1 cannot be aligned with each other, since translation of element *C* is absent from S'_1. Similarly, S_2 and S'_2 cannot be aligned with each other, since translation of element C' is absent from S_2. Therefore, a multiple alignment 2:2 has to be created (S_1, S_2 vs. S'_1, S'_2).

In the DPC project, we restrict ourselves to *non-crossing* alignments. Thus, if there is an alignment of text chunk N of a source language text and text chunk V of a target language text, then no alignment links can be made between chunk M of a source language text and chunk W of a target language text, such that M precedes N and W follows V. Crossing alignments are not allowed.

If cases of cross-translations occur in a text, multiple alignments (many-to-many) are introduced for the analysis: thus, a pair of sentence *m* and *n* will be aligned with a pair of sentences *v* and *w* in the example above.

Sentence alignment is preceded by text normalization and paragraph alignment. A small portion of the corpus will be aligned at sub-sentential level. The intended usage of the sub-sentential links will determine the granularity or level of the linking process, e.g. word-by-word linking to create a lexicon, or linking larger segments (e.g. constituents) for a more structural analysis of the texts. Motivated choices will be made based on the user requirements analysis.

2.2.2. Linguistic annotation

The whole corpus will be lemmatized and enriched with PoS tags. A small portion of the corpus will be enriched with syntactic annotations. To ensure compatibility between the Dutch monolingual corpus being developed in the D-COI project (van den Bosch, Schuurman and Vandeghinste 2006) and the DPC, the PoS tag set and tagger/lemmatizer of the D-COI team will be used. To increase the quality of the linguistic annotations, part of the processing will be manually verified. The manually validated texts will be added to the training corpus, and the tools will be regularly retrained to improve accuracy. The manual verification steps will be performed by students. A small portion of the corpus will be further enriched with shallow parses.

2.2.3. Quality control

Three forms of quality control are envisaged for the DPC data. The first one, traditional manual checking, guarantees high quality of resulting annotations. It is performed by qualified linguists with native and near-native language proficiency. Since manual checking of a 10-milion-word corpus is impossible, a spot checking method is used. Additionally, automatic control procedures are performed, such as the automatic comparison of output from different alignment programs.

3. XML AS BASIS FOR CORPUS EXPLOITATION

Part of the improvement of corpus compilation and exploitation is related to text and character standardisation. Also in the case of DPC, a standardised format based on XML will be used. After cleaning, annotating and aligning the text files, they will be stored in an XML wrapper, thus facilitating the further exchange and annotation of data.

Although closely related to HTML (the markup language for web pages), XML differs in a number of aspects, which makes it a more versatile markup language[19]. First of all, it is an *extensible* markup language, so that extra tags can be created when need be. HTML, on the other hand, is a closed set of markup labels, which are mainly restricted to layout information on the internet. Secondly, XML has a stricter syntax, which avoids possible confusion of related start and end tags, which reduces processing overload for analysing the consistency of the data.

An illustration of the stricter XML requirements is the rule that says that tags (or elements) must be nested without overlap. In the following example, HTML will accept both case A and B, whereas XML will only consider case B as a *well-formed* construction:

A. **<bold><italic>some text</bold></italic>**
B. **<bold><italic>some text</italic></bold>**

In fact, the previous rule is based on the more general rule which stipulates that every element pair has to be nested. But the very first rule indicates that there is only one *root* element which contains all other elements. On the basis

[19] Both XML and HTML use related start tags and end tags, complying with the following basic format: (i) start and end tag use the same name (ii) both tags are placed between angular brackets, and (iii) the end tag is introduced by a slash: e.g. <tag> .. </tag>

of this simple set of rules, an XML document can be represented as a tree, and easily parsed.

XML validation is first of all based on the well-formedness of the document, but a second level of validation takes the syntax of the document into account. This type of validation is based on a kind of document grammar, called DTD (Document Type Definition), which defines the order and the number of elements used. If an XML document complies not only with the rules of well-formedness, but also with the rules of the related DTD, then the XML document is called a *valid* XML document.

Figure 1 shows a very simplified DTD for the structure of a book. This DTD grammar could be rewritten as follows: a book consists of a title, followed by one or more chapters; a chapter consists of a header, followed by one or more paragraphs. The rest of the DTD explains that all the elements consist of character data[20].

In principle, anybody can build his proper XML document format, consisting of the elements/tags you need, together with a customized DTD. However, a DTD can become rather complex. Therefore, it is better to start from existing standardisation formats which have been especially developed for your purpose, and which you can modify where necessary. On the basis of the general rules of the XML document structure, a number of standards have been developed for structuring documents concerning a particular domain: e.g. MathML (Mathematical Markup Language), CML (Chemical Markup Language), SMIL (Synchronized Multimedia Intergration Language). In the case of text standardisation, two formats have gained general acceptance as XML standard: TEI and CES[21]. Both standards are guidelines which define a grammar for describing how texts are constructed and propose names for their components.

---

[20] PCDATA refers to the fact that the characters have been *parsed* (PCDATA = parsed character data), meaning that the characters comply with the character encoding for this document defined. Note also that the plus sign indicates "one or more" elements, whereas the comma indicates the sequential order of the elements (e.g. first comes a <title> element, then one or more <chapter> elements; the other way round is not allowed.)

[21] Although TEI and CES are now often related to XML, the first implementation of both standards are based on SGML. In fact, the XML version of CES is called XCES (referring to *extensible* CES).

```
<!DOCTYPE book [
<!ELEMENT book          (title, chapter+)>
<!ELEMENT chapter       (heading, paragraph+)>
<!ELEMENT title         (#PCDATA)>
<!ELEMENT heading       (#PCDATA)>
<!ELEMENT paragraph     (#PCDATA)>
]>
```

Figure 1: simplified DTD sample for a book

The TEI[22] (Text Encoding Initiative) format was originally used to encode any type of text, which explains its rather extended format. TEI has become the de facto standard for scholarly work with digital text. CES[23] (Corpus Encoding Standard), on the other hand, was mainly focused on natural language processing applications, which explains why the initial element sets and DTD were smaller than those described along the TEI format. In this way, TEI format was mainly used for literary projects, and CES for NLP projects. This distinction is too extreme and no longer valid, since more and more corpus compilation projects are nowadays being compiled and structured in TEI format. Also in the case of DPC, the final format of the aligned corpus will be in TEI.

The use of XML has been an important improvement for the exchange of textual data over different platforms. However, it still remains mainly a transport format. Some types of exploitation, still require conversion to a binary format and construction of index tables, in order to speed up the consultation of the data in a more efficient way.

4. CONCLUSION

The Dutch Parallel Corpus[24] project has been described in this paper. The DPC mainly differs from other existing parallel corpora in the following aspects:

1. *Quality control*: in order to guarantee corpus quality, a considerable part of the DPC corpus is being checked manually at different levels, including sentence splitting, linguistic annotation and alignment. A quality label is used to mark the level of verification.
2. *Level of annotation*: the DPC corpus is aligned, tagged on part of speech level and lemmatized. The annotation and linguistic processing will be

22 http://www.tei-c.org/
23 http://www.cs.vassar.edu/XCES/
24 http:www.kuleuven-kortrijk.be/dpc

produced by state-of-the-art tools. For Dutch, we will adhere to the D-COI conventions as much as possible, strengthening the standards.

3. *Balanced composition*: the DPC contains texts from a wide range of text types (fiction and non-fiction), and diverse domains.
4. *Availability*: in order to maximize research on parallel corpora, the DPC will be made available to the research community via the Agency for Human Language Technologies (the TST-centrale).
5. *Dutch kernel*: the pivotal language of the DPC corpus is Dutch: the corpus contains representative samples where Dutch is the source language. In general, DPC consist of two bidirectional bilingual parts and one trilingual part.

H. PAULUSSEN[*], L. MACKEN[+], J. TRUSHKINA[*],
P. DESMET[*], W. VANDEWEGHE[+]

(*) K.U.Leuven Campus Kortrijk
Subfaculteit Letteren
E. Sabbelaan 53
8500 Belgium
firstname.lastname@kuleuven-kortrijk.be

([+]) LT[3]
University College Ghent
Groot-Brittanniëlaan 45
9000 Gent
Belgium
firstname.lastname@hogent.be

5. REFERENCES

BORIN, L. (1999). The ETAP project - a presentation and status report. *Technical report*, Dept. of Linguistics, Uppsala University, 1999. ETAP research report etap-rr-01.

DOES, J. de & J. VAN DER VOORT VAN DER KLEIJ (2002), "Tagging the Dutch PAROLE Corpus", in: M. Theune et al. (eds.), *Computational Linguistics in the Netherlands 2001; Selected Papers from the Twelfth CLIN Meeting*. Rodopi, Amsterdam - New York, p. 62-76.

DEVILLE, G., DUMORTIER, L. & H. PAULUSSEN (2004), "Génération de corpus multilingues dans la mise en oeuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère", in Gérald P., Fairon, F. & A. Dister (eds.), *Le poids des mots, Actes des 7es journées internationales d'analyse statistique des données textuelles, JADT 2004*, Louvain-la-Neuve, March 2004, 304-312.

ERJAVEC, T., IGNAT, C., POULIQUEN B., & R. STEINBERGER (2005), "Massive multilingual corpus compilation; Acquis Communautaire and totale". In *Proceedings of the 2nd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland.

IDE, N. & J. VÉRONIS (1994), "MULTEXT: Multilingual Text Tools and Corpora". In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan, 588-92.

JOHANSSON, S. (2002a), Oslo multilingual corpus. URL: http://www.hf.uio.no/ilos/OMC/

JOHANSSON, S. (2002b), "Towards a multilingual corpus for contrastive analysis and translation studies". In *Parallel Corpora, Parallel Worlds*. Amsterdam: Rodopi.

KRUYT, J.G. (1998), "Elektronische woordenboeken en tekstcorpora voor Europese taaltechnologie". In *Jaarboek Lexicografie 1997-1998*. Trefwoord 12.

KOEHN, F. (2005), "Europarl: a parallel corpus for statistical machine translation". In *Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand.

MACKEN, L. (2007), "Analysis of translational correspondence in view of sub-sentential alignment". In *Proceedings of the METIS-II Workshop on New Approaches to Machine Translation*, Leuven, Belgium.

MACKEN, L., TRUSHKINA, J., PAULUSSEN, H., RURA, L., DESMET, P. & W. VANDEWEGHE (2007), "Dutch Parallel Corpus: a multilingual annotated corpus". In *Proceedings of The fourth Corpus Linguistics conference*, University of Birmingham.

ODIJK, J., MARTENS, J.P., VAN EYNDE, F., DAELEMANS, W., KENYON-JACKSON, D., VOSSEN, P., VAN HESSE, A., BOVES, L., & J. BEEKEN (2004), *Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie. STEVIN. Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands*. The Hague: Nederlandse Taalunie.

PAULUSSEN, H. (1999), *A corpus-based contrastive analysis of English "on/up", Dutch "op" and French "sur" within a cognitive framework*. Unpublished PhD, University of Gent.

TIEDEMANN, J. & L. NYGAARD (2004), "The OPUS corpus - parallel and free". In *Proceedings of the Fourth International Conference on Language resources and evaluation (LREC'04)*, Lisbon, Portugal.

TJONG KIM SANG, E. (1996). *Aligning the Scania Corpus*. Internal report, Department of Linguistics, Uppsala University.

VAN DEN BOSCH, A., SCHUURMAN, I., & V. VANDEGHINSTE (2006), "Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development". In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genua, Italy.