Meta
Journal des traducteurs
Translators' Journal

META

Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus

Lieve Macken, Orphée De Clercq and Hans Paulussen

Volume 56, Number 2, June 2011

Les corpus et la recherche en terminologie et en traductologie
Corpora and Research in Terminology and Translation Studies

URI: https://id.erudit.org/iderudit/1006182ar
DOI: https://doi.org/10.7202/1006182ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Macken, L., De Clercq, O. & Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta*, *56*(2), 374–390. https://doi.org/10.7202/1006182ar

Article abstract

This paper presents the Dutch Parallel Corpus, a high-quality parallel corpus for Dutch, French and English consisting of more than ten million words. The corpus contains five different text types and is balanced with respect to text type and translation direction. All texts included in the corpus have been cleared from copyright. We discuss the importance of parallel corpora in various research domains and contrast the Dutch Parallel Corpus with existing parallel corpora. The Dutch Parallel Corpus distinguishes itself from other parallel corpora by having a balanced composition and by its availability to the wide research community, thanks to its copyright clearance. All texts in the corpus are sentence-aligned and further enriched with basic linguistic annotations (lemmas and word class information). Approximately 25,000 words of the Dutch-English part have been manually aligned at the sub-sentential level. Rich metadata facilitates the navigability of the corpus and enables users to select the texts that satisfy their needs. The entire corpus is released as full texts in XML format and is also available via a web interface, which supports basic and complex search queries and presents the results as parallel concordances. The corpus will be distributed by the Flemish-Dutch Human Language Technology Agency (TST-Centrale).

Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus

LIEVE MACKEN
University College Ghent and Ghent University, Ghent, Belgium
lieve.macken@hogent.be

ORPHÉE DE CLERCQ
University College Ghent and Ghent University, Ghent, Belgium
orphee.declercq@hogent.be

HANS PAULUSSEN
University of Leuven, Kortrijk, Belgium
hans.paulussen@kuleuven-kortrijk.be

RÉSUMÉ

Le présent article décrit un corpus parallèle de grande qualité en néerlandais, en français et en anglais contenant 10 millions de mots (DPC, pour *Dutch Parallel Corpus*). Les différents types textuels, au nombre de cinq, sont équilibrés, ainsi que les différentes directions de traduction. Tous les problèmes relatifs aux droits d'auteurs ont été résolus. L'importance de la disponibilité des corpus parallèles dans plusieurs domaines de recherche est discutée et nous comparons le DPC avec d'autres corpus multilingues actuels. Le DPC se distingue par sa composition équilibrée et par le fait qu'il est offert à l'ensemble des chercheurs, car il est libre de droits. Les textes sont alignés au niveau de la phrase et enrichis avec des annotations linguistiques (lemme, étiquettes morphologiques). De plus, environ 25 000 mots (dans la partie néerlandais-anglais) ont fait l'objet d'un alignement manuel sous-phrastique. La richesse des métadonnées permet d'effectuer un certain nombre de sélections adaptées aux besoins de l'utilisateur. L'exploitation se fait de deux manières : d'une part, il est possible d'accéder à l'intégralité du corpus et de s'en servir en format XML. D'autre part, le corpus est consultable à travers une interface web qui autorise des requêtes simples ou complexes et présente les résultats sous forme de concordances parallèles. Le corpus sera distribué par l'Agence néerlandaise et flamande pour le traitement automatique des langues (TST-Centrale).

ABSTRACT

This paper presents the Dutch Parallel Corpus, a high-quality parallel corpus for Dutch, French and English consisting of more than ten million words. The corpus contains five different text types and is balanced with respect to text type and translation direction. All texts included in the corpus have been cleared from copyright. We discuss the importance of parallel corpora in various research domains and contrast the Dutch Parallel Corpus with existing parallel corpora. The Dutch Parallel Corpus distinguishes itself from other parallel corpora by having a balanced composition and by its availability to the wide research community, thanks to its copyright clearance. All texts in the corpus are sentence-aligned and further enriched with basic linguistic annotations (lemmas and word class information). Approximately 25,000 words of the Dutch-English part have been manually aligned at the sub-sentential level. Rich metadata facilitates the navigability of the corpus and enables users to select the texts that satisfy their needs. The entire corpus is released as full texts in XML format and is also available via a web interface, which supports basic and complex search queries and presents the results as parallel concordances. The corpus will be distributed by the Flemish-Dutch Human Language Technology Agency (TST-Centrale).

MOTS-CLÉS/KEYWORDS

corpus parallèle, traductologie fondée sur corpus, linguistique de corpus, affranchissement des droits d'auteurs, interface web

parallel corpus, corpus-based translation studies, corpus linguistics, copyright clearance, web interface

> Through corpora, we can observe patterns in language which we were unaware of before or only vaguely glimpsed.
>
> (Johansson 2007: 1)

1. Introduction

In the last decades corpus linguistics and corpus-based translation studies have developed considerably and many monolingual and multilingual corpora representing both well-studied and less-studied languages have emerged.[1] Although many conflicting opinions exist as to the usefulness of corpora in linguistic analysis, language teaching and language learning, corpora have been widely accepted as valuable linguistic resources (for a discussion, see McEnery, Xiao *et al.* 2006: 243-258).

There are already thousands of DIY[2] corpora that were principally created for specific research projects and consequently are not publicly available or do not satisfy a specific user's needs (Xiao 2010: 147). One might ask therefore why time, effort and money should be invested in corpus creation when the product might not be accessible to those interested?

This paper will attempt to counter this view by presenting a new resource that, thanks to its multifunctional design, aims to be of use in both the linguistically-oriented and more technological fields of corpus linguistics. The Dutch Parallel Corpus (DPC) is a multifunctional and bidirectional parallel corpus of Dutch, English and French with Dutch as a central language. It contains more than ten million words, is completely cleared from copyrights and all the text material is aligned at sentence level and annotated with linguistic information (lemmas and part-of-speech tags). Throughout the entire data collection process and data processing steps, four objectives were of paramount importance: a balanced corpus design, high quality, easy access and widespread availability.

Dutch, the language spoken in the Netherlands and Flanders, the northern part of Belgium, has long been under-represented in the rapidly evolving language industry. For this reason, the STEVIN programme,[3] a Flemish/Dutch human language technology research programme, was set up to strengthen the economic and cultural position of Dutch in the modern ICT-based society. One of its key objectives was building a parallel corpus with Dutch as a central language.

Building a parallel corpus fulfilling all the above-mentioned needs presented a challenging task: based on other corpus projects (Section 2) and a user requirements study, a balanced design was created and effectuated (Section 3); maintaining high quality was achieved by actively keeping track of as many metadata as possible; aligning everything at sentence level and a small part even sub-sententially (Section 4). Last but not least, a user-friendly web interface was built so as to assure the corpus' ease of use and wide availability (Section 5).

2. Parallel Corpora in Translation Studies

With the introduction of a corpus-based methodology in the field of translation studies (Baker 1993) and the increased availability of large monolingual and multilingual corpora, empirical studies have been conducted to examine the fundamental characteristics of translated text. The results of these studies has offered us insights into both the nature of translated language and the translation process. Translation universals, the ideology of translation and stylistic differences between translators, to name only a few, have been and will continue to be extensively discussed. Apart from translation studies, corpora have proven to play an important role in other linguistically-oriented research areas as well as in natural language processing research (NLP).

The role of parallel corpora is very diverse, ranging from more technological NLP applications to methodological approaches in linguistics and translation studies. Aligned parallel corpora play a fundamental role in developing corpus-based statistical MT (Koehn 2005) and example-based MT (Carl and Way 2003). Apart from machine translation, they are also a helpful resource for computer-assisted translation tools (Hutchins 2005) and computer-assisted language learning (Deville, Dumortier *et al.* 2004). Parallel corpora have proven especially useful when studying translated text (Halverson 1998) and when it comes to contrastive linguistics, they are often combined with comparable corpora to validate research hypotheses.[4] Bernardini (2010)[5] and McEnery and Xiao (2008) emphasize that monolingual comparable corpora are useful for highlighting overall features of translated texts but that parallel corpora are ideal for observing translation shifts. This implies that a new corpus structure should emerge: a corpus should not only contain originals in a particular language with their translations, but also a set of comparable texts in the source and target language.

According to Ebeling (1998: 604), creating a bidirectional parallel corpus is already one way of making such a corpus, because "the effect of translationese is averaged out to some extent." McEnery and Xiao (2008: 23), however, warn that to achieve this, the same sampling frame must be used for selecting source data in both languages, because "any mismatch of proportion, genre, or domain, for example, may invalidate the findings derived from such a corpus."

Given its bidirectional and balanced design, the Dutch Parallel Corpus can be perceived as a corpus that is both parallel and comparable. The central role of Dutch makes it a corpus that can be placed within the category of corpora that are being developed for less-widespread languages. A brief analysis of other parallel corpus projects is given below.

2.1. Parallel Corpus Projects

Regardless whether parallel corpora include a Dutch component, we can observe two major drawbacks in existing parallel corpora. First of all, numerous corpora lack text type balance, such as the Europarl corpus,[6] the Canadian Hansard corpus[7] and the European Corpus Initiative.[8] These all contain one or maximum two text types and the corpus builders often limit themselves to readily-available data. The texts in the Europarl corpus, for example, consist only of the proceedings of parliamentary

debates. Since all these debates have been translated to all languages of the member states and are stored electronically, it is rather easy to compile a corpus including these texts. A related problem is that texts originating from the institutions of the European Union can be problematic "since it can be difficult to assign the status of 'source texts' to one of the language versions, documents may be written in more than one language and, once translations exist, there is nothing to distinguish a source texts from the 'other language variations'" (Koskinen 2000: 55). The same problem arises with manuals drawn up by large multinationals written and translated into various languages.

On the other hand, there exist corpora with a balanced design but that are unavailable to the research community because of Intellectual Property Rights (IPR) issues, such as the English-Norwegian Corpus.[9] Although there is currently no universal approach dealing with copyrights, this does not mean that copyrights should not be cleared. Moreover, the general rule is: "Whenever in doubt, seek permission" (Xiao 2010: 153).

Another issue is that not all corpora are processed in the same way. Users of parallel corpora are interested in retrieving instances in the source language together with their translations (Olohan 2004: 25). In order to achieve this, a parallel corpus has to be aligned at sentence level. For most parallel corpora, sentence alignment was done automatically without any or with very little manual verification. The addition of linguistic information such as part-of-speech tags or lemmas is also important to ensure the multidisciplinary character of the corpus. Nevertheless, only few parallel corpora (for example OPUS[10]) are provided with these annotations.

Keeping these disadvantages of existing parallel corpora in mind, it was decided to create a corpus that would solve all these problems for Dutch and thus to create a balanced corpus that is completely cleared from copyrights and that is aligned at sentence level and enriched with linguistic annotation, whilst ensuring high quality.

3. Corpus Design, Copyright Clearance and Metadata

From other corpus projects and parallel corpora used in translation studies we learned that creating a balanced and representative corpus is of paramount importance. The process needed to achieve this, however, is subject to various pitfalls.

Because of its interdisciplinary objectives the Dutch Parallel Corpus was designed to fit as many users as possible. To this purpose a user requirements study was carried out. A predefined user group – composed of academic and industrial specialists from different application and research domains – was asked to fill out a questionnaire. The main results of this questionnaire confirmed our anticipations:

- There was a strong need for a parallel corpus with Dutch as the central language;
- The quality of text material, alignments and linguistic annotations is of paramount importance;
- The variety of text types is more important than including full text;
- Rich metadata should be provided for every text included in the corpus.

Based on this user requirements study and after studying other parallel corpus projects, we were able to motivate our choices when it came to defining a balanced text typology, collecting copyright-cleared material and providing sufficient metadata so as to ensure the creation of a multifunctional corpus.

The finalized Dutch Parallel Corpus contains over ten million words and is balanced with respect to five text types (administrative texts, instructive texts, literature, journalistic texts and texts for external communication) and four translation directions (Dutch-English, English-Dutch and Dutch-French, French-Dutch). This implies that each text type should contain about two million words and that within each text type each translation direction should contain about 500,000 words.

The five text types were further subdivided in accordance with the prototype approach suggested by Lee (2001). This resulted in a two-level typology which is presented in Table 1. Having no implications for the balancing of the corpus, this subdivision is just a way of mapping the diversity of text material categorized under a particular text type, so as to allow the end user to correctly select documents. The text type information is also stored in the metadata.

Table 1

The two-level typology of the Dutch Parallel Corpus

Superordinate level	Basic level
1. Literature	1.1 Novels
	1.2 Essayistic texts
	1.3 (Auto)biographies
	1.4 Expository works of a general nature
2. Journalistic Texts	2.1 News reporting articles
	2.2 Comment articles
3. Instructive Texts	3.1 Manuals
	3.2 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 Official speeches
5. External Communication	5.1 (Self-)presentation of organizations
	5.2 Informative documents
	5.3 Promotion/advertising material
	5.4 Press releases
	5.5 Scientific texts

The finalized corpus contains approximately twelve million tokens and about ten million words (see Appendix for the details).

Another objective apart from this balanced design was to make sure that every text sample included in the corpus be cleared from copyrights. Together with experts from the Flemish-Dutch Human Language Technology Agency[11] (TST-Centrale, for Centrale voor Taal- en Spraaktechnologie), four different types of agreements were drawn up so as to assure that every sample in the corpus would be cleared from copyrights. For a detailed description of every step in the data collection process and the various problems that were encountered during copyright negotiations, we refer to De Clercq and Montero Perez (2010).

Rich metadata is an essential prerequisite to the optimal use of any corpus: "Metadata plays a key role in organizing the ways in which a language corpus can be meaningfully processed. It records the interpretive framework within which the components of a corpus were selected and are to be understood" (Burnard 2005: 46).

Each text in the DPC has an accompanying metadata file. During corpus creation, the metadata was of vital importance to build a balanced corpus that met the objectives that had been set in the corpus design phase. Similarly, during corpus exploitation the metadata will enable the corpus users to select the texts that fulfil their specific requirements.

The metadata stores different types of additional information. Firstly, the metadata contains information on the origins of the texts included in the corpus: publisher, translation direction, author or translator of the text, date of publication, and the like. Secondly, DPC project staff members added extra information to further characterize each text: text type and sub-type, domain and keywords, intended audience, and the like. Thirdly, the metadata indicates the type of IPR agreement that was concluded with the publishers as the IPR agreement determines the terms of use of the texts included in the corpus.

4. Alignment and Linguistic Annotation

It is generally accepted that parallel corpora become more valuable when the raw text material is enriched by different kinds of annotations. All texts of the DPC have been enriched with basic linguistic annotations (lemmas and word class information).

Additionally, the corresponding units in the source and target texts have been aligned. These correspondences can be established at the level of paragraphs, sentences, or words. While the quality of automatic sentence alignment programs is relatively high, considerably more manual effort is needed to establish high-quality sub-sentential translational correspondences. The entire Dutch Parallel Corpus has been aligned at sentence level. Furthermore, a small portion of the Dutch-English part has been aligned at sub-sentential level.

4.1. Sentence Alignment

Sentence alignment is the process of finding equivalent text chunks at the level of the sentence in parallel texts. The sentences linked by the alignment procedure represent translations of each other in different languages. An example is presented in Table 2. Most sentence alignments are one-to-one (1:1), one-to-many (1:m) or many-to-one (m:1) alignments. Null alignments are used to indicate deletions and additions; many-to-many alignments are used to model overlapping alignments. As crossing alignments cannot be handled by the automatic sentence alignment algorithms, they are grouped as many-to-many alignments.

TABLE 2

Sentence alignments extracted from the DPC

English text	Alignment links	Dutch text
The tiger's teeth	1:Ø	
	Ø:1	De werkplaats van de wereld
In 1980, three brothers in Qiaotou started up a business by picking buttons off the street.	1:2	In 1980 richtten drie broers in Qiaotou een bedrijfje op. Op die manier konden ze iets verdienen aan de knopen die ze op straat vonden.

Twenty-five years on, this remote town makes almost every zip and button we wear.	1:1	Vijfentwintig jaar later komt haast elke rits en knoop die aan onze kleren zit uit dat afgelegen stadje in China.
Just down the road, another has become a global centre for toothbrush-making, while a third is now the world capital of socks.	1:1	Een dorp in de buurt is uitgegroeid tot een mekka voor producenten van tandenborstels, en nog een ander stadje is de sokkenhoofdstad van de wereld.

Example of null (1:Ø; Ø:1), one-to-one (1:1) and one-to-many (1:m) sentence alignments (Example taken from English-Dutch subcorpus, text dpc-sta-002543)

Although a range of tools and algorithms are available for the task of sentence alignment, basically two different approaches to sentence alignment can be distinguished: the sentence-length-based approach and the word-correspondence-based approach.

In the sentence-length-based approach, the alignment process is guided by the assumption that the lengths of corresponding sentences are highly correlated. In other words, short sentences tend to be translated by short sentences, and long sentences by long sentences or several short sentences. The sentence-length-based approach was introduced by Gale and Church (1991) and Brown, Lai *et al.* (1991). A probabilistic score is assigned to each proposed correspondence of sentences and the scores are used in a dynamic programming framework to find the maximum likelihood alignment of sentences. Structural information (headers, titles, paragraph information and the like) can be used to restrict the space of allowable alignments.

The word-correspondence-based approach is described in the seminal paper of Kay and Röscheisen (1993) and is based on the assumption that if sentences are translations of each other, the corresponding words must be translations as well. Their algorithm performs both sentence and word alignment and both processes reinforce each other. While Kay and Röscheisen used only text-internal information and derived the word correspondences from the texts to be aligned, an intuitive extension is the use of electronically available bilingual dictionaries (Melamed 1997). A second extension is the use of cognates (word tokens that are graphically identical or similar such as proper names, dates, certain symbols and the like) as corresponding words (Simard, Foster *et al.* 2000).

The performance of the individual alignment tools varies for different types of texts and language pairs and in order to guarantee high quality alignments, a manual verification step is needed. Macken (2010b) demonstrated that using the combined output of different alignment tools can drastically reduce this manual verification effort. In the Dutch Parallel Corpus project, we combined the output of three different alignment tools: the Vanilla Aligner (Danielsson and Ridings 1997), the Geometric Mapping and Alignment tool (Melamed 1997) and the Microsoft Bilingual Aligner (Moore 2002). Only the alignments that were not identified by at least two alignment tools were manually verified.

The percentages of the resulting sentence alignment types in the Dutch Parallel Corpus are presented in Table 3. In total, the DPC contains 293,163 sentence alignments, of which 87.4% are one-to-one alignments (1:1), almost 5% are null alignments (Ø:1 and 1:Ø) and 7.9% are one-to-many, many-to-one or many-to-many alignments (1:m. m:1, m:m).

TABLE 3

Rate of sentence alignment types in the DPC

Text Type	Sentence alignment (%)			
	Ø:1	**1:Ø**	**1:1**	**1:m, m:1, m:m**
Administrative texts	1.2	1.1	92.8	4.9
External Communication	5.4	5.6	81.0	8.0
Instructive texts	0.4	0.6	95.7	3.3
Journalistic texts	3.2	2.6	82.4	11.9
Literature	0.9	1.2	84.0	13.9
Total	**2.4**	**2.3**	**87.4**	**7.9**

If we take a closer look at the resulting sentence alignment types broken down per text type, the following trends can be observed:

- The administrative and instructive texts are translated rather literally, which is reflected by the high percentage of one-to-one alignments (92.8% and 95.7% respectively) and a very low percentage of null alignments (2.3% and 1.0% respectively);
- The texts dealing with external communication and – to a lesser extent – the journalistic texts exhibit a high percentage of null alignments (11% and 5.8%), which is an indication that the texts have been adapted during translation. Additions and deletions typically occur in the begin and end sections in the journalistic texts, but can occur at any place in the texts dealing with external communication;
- The literary and journalistic texts are characterized by a high percentage of one-to-many, many-to-one and many-to-many alignments (13.9% and 11.9% respectively), which means that quite a lot of sentences were split or merged during translation. This is tangible proof that the translators did not translate sentence by sentence, and might suggest that for some text types the sentence is not the key functional unit of translation (Zhu 1999) but that translators operate at the level of the paragraph. Two examples are presented in Table 4.

TABLE 4

Example of a many-to-many sentence alignment (English-Dutch) and a one-to-many sentence alignment (French-Dutch)

Source text	Alignment links	Target text
(1) "Magic moments have happened," said Ure. (2) "Justin [Hawkins] from the Darkness was standing there watching Dizzee Rascal do his rap bit in the middle and we were thinking, 'wow, that was fantastic.' (3) All the boundaries that musicians put up between them – I'm a rock star, you're a rap star – have disappeared."	3:3	(1) Hét magische moment, vond zowat iedereen, kwam van de jonge rapper Dizzee Rascal, die ter plekke een rap schreef en hem meteen insprak. (2) "Alle grenzen tussen genres vervaagden," mijmerde Justin Hawkins van The Darkness. (3) "Dat typische sektarisme – jij bent een rockster, ik ben een rapper – was helemaal verdwenen."

(1) Inscrit entre ville et littoral, bordé par le nouveau projet urbain Neptune, le LAAC (Lieu d'art et d'action contemporaine) a rouvert ses portes le 25 juin 2005, rénové par les architectes Grafteaux & Klein, qui ont su très subtilement transformer les défauts de cet édifice en qualité: éclairage, acoustique et mobilier.	1:3	(1) Het museum bevindt zich tussen de stad en de kust en ligt naast de Neptunussite, een grootschalig project voor ruimtelijke ordening. (2) Op 25 juni 2005 opende het LAAC (Lieu d'art et d'action contemporaine) opnieuw zijn deuren. (3) Het werd gerenoveerd door de architecten Grafteaux & Klein, die de oorspronkelijke tekortkomingen van het gebouw op een subtiele manier tot troeven wisten om te vormen (verlichting, akoestiek en meubilair).

4.2. Sub-sentential Alignment

Alignments below the level of the sentence are an even better means to reveal the differences in translation style in various text types. Therefore, in order to study the problem of translational correspondence below the level of the sentence, three sub-corpora were extracted from the English-Dutch part of the Dutch Parallel Corpus.

The texts were selected from three different text providers: (i) journalistic articles that were originally published in *The Independent* and translated into Dutch for *De Morgen*, a Flemish quality newspaper; (ii) newsletters from ING, a Dutch financial institution with diverse international activities, which brings financial news to private investors; and (iii) medical European Public Assessment Reports (EPARs) of one pharmaceutical company.

In total, more than 25,000 words were manually aligned at the sub-sentential level. Table 5 summarizes the formal characteristics of the sub-corpora and presents the total number of words and the average sentence length of source (src) and target (tgt) sentences.

Table 5

Characteristics of below sentence aligned sub-corpora

Text Type	Total words	Sentence length (src)	Sentence length (tgt)
Journalistic articles	7,706	22.0	20.0
Newsletters	10,480	15.0	15.4
EPARs	7,536	17.2	17.7

Generally speaking, the minimal language units in the source text that correspond to an equivalent in the target text have been aligned. Three types of links were introduced: (i) regular links were used to connect straightforward correspondences; (ii) fuzzy links for translation-specific shifts of various kinds (divergent translations and paraphrases); and (iii) null links were used for source text units that had not been translated or target text units that had been added.

To make the manual annotations as useful as possible for different types of projects, a multi-level annotation scheme has been employed in the case of divergent translations: fuzzy links were used to connect divergent translations; regular links were used to connect corresponding words within the paraphrased sections. An example of an annotated sentence pair is presented in Figure 1.

FIGURE 1
Sub-sentential alignments

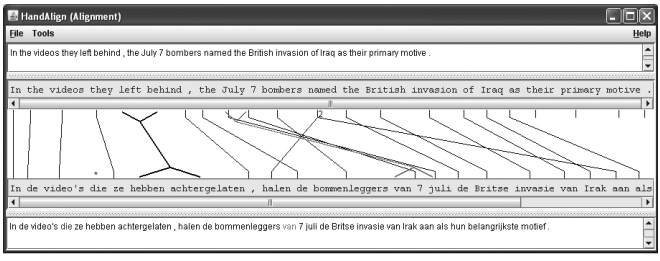

Table 6 gives an overview of the different types of links that were indicated in the three sub-corpora. As expected, a different degree of *freeness* can be observed, which is reflected in the percentage of fuzzy links and null links. A freer translation style is characterized by a high degree of fuzzy links and null links: the journalistic texts contain the highest number of fuzzy links (11.1%) and the highest number of null links (9.3%). Texts with a high degree of regular links follow more closely the forms of the source texts: the EPARs contain the lowest percentage of fuzzy (6.5%) and null links (3.9%). The newsletters are somewhere in between. More details regarding the annotation task can be found in Macken (2010a).

TABLE 6
Rates of links in below sentence aligned sub-corpora

Text types	Link rates (%)		
	regular	fuzzy	null
Journalistic articles	79.6	11.1	9.3
Newsletters	88.6	6.9	4.5
EPARs	89.6	6.5	3.9

4.3. Linguistic Annotation

The two most common forms of corpus annotation are part-of-speech tagging and lemmatization. Prior to adding these annotations, all texts have been automatically divided into sentences and tokenized. During tokenization, a sentence is split into sequences of words and all punctuation marks not belonging to the word form (punctuation marks that are not part of an abbreviation) are stripped off.

In the domain of natural language processing (NLP), part-of-speech tagging is a widely-researched and well-understood task (van Halteren, Zavrel *et al.* 2001), and nowadays programs that automatically assign part-of-speech tags are available for most languages. On the basis of a given word and its context, a part-of-speech tagger determines to which morpho-syntactic class (noun, verb, adjective) the word belongs. Reported accuracy scores of part-of-speech taggers typically fluctuate around 95%.

Lemmatization implies generation of a base form or lemma for each orthographic token. The typical base form for verbs is the infinitive, for nouns the singular. Lemmatization is often used to abstract over the word forms that appear in the corpus. It goes without saying that enriching a corpus with lemmas is more important

for highly inflected languages (for example German, the Romance and Slavic languages), than for weakly inflected ones (e.g., English).

Part-of-speech tagging and lemmatization are related tasks and are therefore often combined in one program. The lemmatizer makes use of the predicted part-of-speech tags to disambiguate ambiguous word forms, for example the Dutch word landen can be an infinitive (with base form landen) or the plural form of a noun (with base form land).

The entire Dutch Parallel Corpus has been tokenized, lemmatized and enriched with part-of-speech tags. Since these steps are language-dependent, different tools were used for each DPC language.

For Dutch, we made use of the combined D-Coi part-of-speech tagger/lemmatizer (van den Bosch, Schuurman *et al.* 2006), which uses the CGN part-of-speech tag set (Van Eynde, Zavrel *et al.* 2000). The CGN part-of-speech tag set is characterized by a high level of granularity. Apart from the word class (noun, adjective, verb), a wide range of morpho-syntactic features (singular, plural, case information, tense) are indicated as attributes to the word class. In total, 316 distinct full tags are discerned.

For English, part-of-speech tagging and lemmatization was performed by the combined memory-based part-of-speech tagger/lemmatizer, which is part of the MBSP tools (Daelemans and van den Bosch 2005). The English memory-based tagger was trained on data from the Wall Street Journal corpus in the Penn Treebank (Marcus, Santorini *et al.* 1993), and uses the Penn Treebank tag set, which contains only 45 distinct tags.

For French, we used Treetagger (Schmid 1994) with the LIMSI parameter file (Allauzen and Bonneau-Maynard 2008), which is based on the GRACE part-of-speech tag set (Paroubek 2000). The GRACE tag set is fine-grained and contains 312 morpho-syntactic tags.

In order to evaluate the performance of the annotation tools, we manually verified the accuracy of automatically predicted lemmas and part-of-speech tags on samples of the DPC containing texts of different text types. The results are presented in Table 7. Because the Dutch and French tag sets are so fine-grained two accuracy scores are presented for part-of-speech tagging: a first score is calculated on the fine-grained full tag (e.g. N(soort,ev,basis,zijd) which stands for a common singular basic male/female noun); a second score only takes into account the main category. (e.g. N) which stands for noun.

Table 7

Lemmatization and part-of-speech tagging of corpora in Dutch, English and French: accuracy scores

		Accuracy scores		
	Sample size (n tokens)	**Accurate lemmas (%)**	**PoS full tag (%)**	**PoS main category (%)**
Dutch	211,000	96.5	94.8	97.4
English	300,000	98.1	96.2	N/A
French	330,000	98.1	94.6	97.4

5. Corpus Exploitation

In order to make the corpus as suitable as possible for further exploitation, special attention was devoted to the structuring of the data. A corpus is only useful if the data can be exploited in a transparent way; exploitation can be hampered considerably by a corpus without proper structuring, or by an ad hoc structuring of the data. Therefore, we adhered to internally accepted standards for corpus compilation.

All the data of the DPC have been structured in XML and are marked up according to the Text-Encoding Initiate P5 guidelines[12]: the source and the target texts are coded as monolingual XML files and contain the annotated sentences; the sentence alignments are stored in a separate index file, in which the indexes point to the sentences of the monolingual files. For programmers, it is relatively easy to convert the XML files to other formats, for example TMX,[13] a format that is used to exchange data in translation memory tools.

Additionally, a dedicated web interface has been developed to make the corpus easily accessible for users that are not familiar with programming or data processing techniques. The DPC web interface consists of a monolingual and a parallel concordancer, in which the user can search for words or word patterns. The results are presented in a split window, which displays sample sentences in both languages, similar to programs like ParaConc[14] or Multiconcord.[15] A first general difference between the DPC web interface and the above-mentioned parallel concordancers consists in the fact that the latter are stand-alone applications limited to one platform (usually Windows), whereas the DPC interface can be consulted via a browser from any place on the web. But more importantly, the DPC web interface offers additional selection criteria. First of all, the user can choose to either select the whole DPC corpus or a sub-corpus on the basis of the following metadata criteria: text type, domain, IPR agreement, and source language. By refining the selection criteria, the user can search for specific words within a predefined sub-corpus.

Once the corpus has been selected, the user can formulate search queries in the form of words or word patterns (either defined as word tokens or lemmas), part-of-speech tags or a combination of words and part-of-speech tags. All matches are retrieved and the result is displayed as a parallel KWIC (keywords in context) concordance. By pressing the button next to the sample, the user can get extra information: either the metadata record is shown or extra context is given, represented as some extra sentences before and after the sample sentence. The resulting output can be saved as an Excel sheet, in which the user can add extra annotations.

An example of a combined search is shown in the figure below, in which the user carried out a search in the Dutch-French language pair. The example concerns the use of past tenses in French and Dutch. We defined a bilingual search query to retrieve examples that contain the French verb avoir used as an auxiliary in the present tense followed by a past participle and whose translations contain a Dutch verb in the past tense.

Figure 2

KWIC concordance output of the web interface

A simple web interface allows for the DPC data to be searched for parallel sample sentences using different types of selections (words, word patterns, lemmas, and the like). This allows the user to quickly get some typical sample sentences. For more fine-tuned selections, the full corpus can be queried using programming tools adapted to search XML files.

6. Conclusion

The last two decades saw the rise of parallel corpora as part and parcel of various research domains such as contrastive linguistics, translation studies or machine translation. In all these research areas, considerable effort is spent on the creation of parallel corpora, which are either not publicly available or are limited in their scope and hence not useful for a wide range of research purposes. The two most severe limitations are lack of text type balance and lack of information on translation direction. When the subject under study is the translation product or the translation process, these constraints seriously hamper research in these fields.

As the creation of corpora is time-consuming and costly, the deliberate aim of the Dutch Parallel Corpus project was to create a multifunctional resource that would fill the needs of a diverse group of researchers. Copyright clearance has been obtained for all texts included in the DPC in order to guarantee the accessibility of the corpus. The result is a sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French of more than ten million words. To cover a wide range of phenomena that emerge from different writing and translation styles, the texts included in the corpus belong to five different text types. The corpus is balanced with respect to text type and translation direction. As the DPC is bidirectional (Dutch as source and target language), the corpus can also be used as a comparable corpus (to compare texts originally written in Dutch with translated Dutch texts).

At the moment of writing, the corpus has been delivered to the Dutch Human Language Technology Agency, which will be responsible for its distribution. The corpus will be distributed as full texts stored in XML format but can also be consulted via a dedicated web interface that supports basic and complex search queries and presents the results as parallel concordances.

ACKNOWLEDGMENTS

The DPC project has been carried out within the STEVIN programme, which is funded by the Dutch and the Flemish Governments. The DPC was created by a Flemish consortium (University of Leuven – Campus Kortrijk and the Faculty of Translation Studies of Ghent University College). The following researchers contributed to the DPC: Piet Desmet, Willy Vandeweghe, Hans Paulussen, Lieve Macken, Maribel Montero Perez, Orphée De Clercq, Lidia Rura, Julia Trushkina and Antoine Besnehard.

NOTES

1. For further information on available corpora, we refer to Xiao (2008), Ostler (2008) and the European Language Resources Association (ELRA): <http://www.elra.info/Language-Resources-LRs.html>, visited 3 May 2011.
2. DIY stands for *do-it-yourself*. Notion which appears in many textbooks treating corpus linguistics, usually when giving advice to corpus compilers for building their own corpus. See for example McEnery, Xiao *et al.* (2006: 71-76).
3. More information can be found at <http://taalunieversum.org/taal/technologie/stevin/english/>, visited 3 May 2011.
4. In order to prevent further confusion, a brief terminological note is in order. In accordance with Baker (1995: 230) and McEnery and Wilson (1996: 57), we define a parallel corpus as a collection of texts with their translations, in contrast to comparable corpora that can be seen as monolingual subcorpora using the same sampling frame (McEnery and Xiao 2008).
5. Bernardini, Silvia (2010): Parallel corpora and the search for translation norms/universals. Plenary talk given at the symposium MATS 2010 *Methodological Advances in Corpus-Based Translation Studies* Ghent, 8-9 January 2010.
6. <http://www.statmt.org/europarl/>, visited 3 May 2011.
7. <http://www.isi.edu/natural-language/download/hansard/>, visited 3 May 2011.
8. <http://www.elsnet.org/eci.html>, visited 3 May 2011.
9. <http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/>, visited 3 May 2011.
10. <http://urd.let.rug.nl/tiedeman/OPUS/>, visited 3 May 2011.
11. TST-Centrale: "The Dutch HLT Agency is the Dutch-Flemish agency for management, maintenance and distribution of Dutch digital language resources. Most resources are government-funded. The HLT Agency makes them available for education, research and development." See <http://www.inl.nl/en/tst-centrale>, visited 3 May 2011.
12. <http://www.tei-c.org/Guidelines/P5/>, visited 3 May 2011.
13. Translation Memory eXchange: <http://www.lisa.org/tmx/tmx.htm>, visited 3 May 2011.
14. <http://artsweb.bham.ac.uk/pKing/multiconc/l_text.htm>, visited 3 May 2011.
15. <http://www.athel.com/para.html>, visited 3 May 2011.

REFERENCES

Allauzen, Alexandre and Bonneau-Maynard, Hélène (2008): Training and evaluation of POS taggers on the French MULTITAG corpus. *In*: Nicolas Morales, Javier Tejedor, Javier Garrido, *et al.*, eds. *Proceedings of the Sixth Language Resources and Evaluation (LREC'08)*. (Language Resource and Evaluation Conference, Marrakech, 28-30 May 2008). European Language Resources Association (ELRA). Visited 3 May 2011, <http://www.lrec-conf.org/proceedings/lrec2008/>.

Baker, Mona (1993): Corpus linguistics and translation studies: implications and applications. *In*: Mona Baker, Gill Francis and Elena Tognini-Bonelli, eds. *Text and Technology: in honour of John Sinclair.* Amsterdam: Benjamins, 233-252.

Baker, Mona (1995): Corpora in translation studies: an overview and some suggestions for future research. *Target*. 7(2):223-243.

Brown, Peter F., Lai, Jennifer C. and Mercer, Robert L. (1991): Aligning sentences in parallel corpora. *In*: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. (Annual Meeting of the Association for Computational Linguistics, Berkeley, 18-21 June 1991), 169-176.

Burnard, Lou (2005): Metadata for corpus work. *In*: Martin Wynne, ed. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 30-46.

Carl, Michael and Way, Andy (2003): *Recent Advances in Example-Based Machine Translation*. Dordrecht: Kluwer Academic Publishers.

Daelemans, Walter and van den Bosch, Antal (2005): *Memory-based Language Processing*. Cambridge: Cambridge University Press.

Danielsson, Pernilla and Ridings, Daniel (1997): Practical presentation of a "vanilla" aligner. *In*: *Proceedings of the TELRI Workshop on Alignment and Exploitation of Texts*. (Workshop on Alignment and Exploitation of Texts, Ljubljana, 1-2 February 1997).

De Clercq, Orphée, Montero Perez, Maribel (2010): Data Collection and IPR in Multilingual Parallel Corpora. *In*: *Proceedings of the Seventh Language Resources and Evaluation (LREC'10)*. (Language Resource and Evaluation Conference, Valletta, 19-21 May 2010).

Deville, Guy, Dumortier, Laurence and Paulussen, Hans (2004): Génération de corpus multilingues dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère. *In*: Gérald Purnelle, Cédrick Fairon and Anne Dister, eds. *Le poids des mots: Actes des 7e journées internationales d'analyse statistique des données textuelles (JADT'04)*. (Journées internationales d'analyse statistique des données textuelles, Louvain-la-Neuve, 10-12 March 2004). Louvain-la-Neuve: Presses Universitaires de Louvain, 304-312.

Ebeling, Jarle (1998): Contrastive Linguistics, Translation and Parallel Corpora. *Meta*. 43(4):602-615.

Gale, William A. and Church, Kenneth W. (1991): A program for aligning sentences in bilingual corpora. *In*: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. (Annual Meeting of the Association for Computational Linguistics, Berkeley, 18-21 June 1991). 177-184.

Halverson, Sandra (1998): Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study. *Meta*. 43(4):494-514.

Hutchins, John (2005): Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*. 17(1-2):5-38.

Johansson, Stig (2007): *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam: Benjamins.

Kay, Martin and Röscheisen, Martin (1993): Text-Translation Alignment. *Computational Linguistics*. 19(1):121-142.

Koehn, Philipp (2005): Europarl: a parallel corpus for statistical machine translation. *In*: *Conference Proceedings: the tenth Machine Translation Summit*. (MT Summit X, Phuket, 13-15 September 2005). 79-86.

Koskinen, Kaisa (2000): Institutional Illusions. Translating in the EU Commission. *The Translator*. 6(1):49-65.

Lee, David Y.W. (2001): Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning and Technology*. 5(3):37-72.

Macken, Lieve (2010a): An annotation scheme and Gold Standard for Dutch-English word alignment. *In*: *Proceedings of the Seventh Language Resources and Evaluation (LREC'10)*. (Language Resource and Evaluation Conference, Valletta, 19-21 May 2010).

Macken, Lieve (2010b): *Sub-sentential alignment of translational correspondences*. Doctoral dissertation, unpublished. Antwerp: University of Antwerp.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993): Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*. 19(2):313-330.

McEnery, Tony and Wilson, Andrew (1996): *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony and Xiao, Richard (2008): Parallel and Comparable Corpora: What is Happening? *In*: Gunilla Anderman and Margaret Rogers, eds. *Incorporating Corpora: The Linguist and the Translator.* Frankfurt: Multilingual Matters, 18-31.

McEnery, Tony, Xiao, Richard, Tono, Yukio (2006): *Corpus-Based Language Studies: An advanced resource book*. London: Routledge Taylor and Francis Group.

Melamed, Dan I. (1997): A Portable Algorithm for Mapping Bitext Correspondence. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. (Annual Meeting of the Association for Computational Linguistics, Madrid, 7-12 July 1997). California: Morgan Kaufmann Publishers, 305-312.

Moore, Robert C. (2002): Fast and accurate sentence alignment of bilingual corpora. *In*: Stephen Richardson, ed. *Machine Translation: from research to real users.* (Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, Tiburon, 8-12 October 2002). Berlin: Springer, 135-244.

Olohan, Maeve (2004): *Introducing Corpora in Translation Studies*. New York: Routledge.

Ostler, Nicholas (2008): Corpora of less studied languages. *In*: Anke Lüdeling and Merja Kytö, eds. *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: Mouton de Gruyter, 457-484.

Paroubek Patrick (2000): Language resources as by-product of evaluation: the Multitag example. *In*: *Proceedings of the Second Language Resources and Evaluation (LREC'00).* (Language Resources and Evaluation Conference, Athens, 30-02 May/June 2000). European Language Resources Association (ELRA), 151-154.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. *In*: Daniel B. Jones and Harold Somers, eds. *New Methods in Language Processing (Studies in Computational Linguistics).* (International Conference on New Methods in Language Processing, Manchester, 14-16 September 1994) London: Routledge, 154-163.

Simard, Michel, Foster, George, Hannan, Marie-Louise, *et al.* (2000): Bilingual text alignment: where do we draw the line? *In*: Simon Botley, Anthony McEnery and Andrew Wilson, eds. *Multilingual corpora in teaching and research*. Amsterdam: Rodopi, 38-64.

van Halteren, Hans, Zavrel Jakub and Daelemans, Walter (2001): Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics.* 27(2):199-229.

van den Bosch, Antal, Schuurman Ineke and Vandeghinste, Vincent (2006): Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development. *In*: *Proceedings of the Fifth Language Resources and Evaluation (LREC'06).* (Language Resources and Evaluation Conference, Genua, 22-28 May 2006). European Language Resources Association (ELRA).

Van Eynde, Frank, Zavrel, Jakub and Daelemans, Walter (2000): Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. *In*: *Proceedings of the Second Language Resources and Evaluation (LREC'00).* (Language Resources and Evaluation Conference, Athens, 30-02 May/June 2000). European Language Resources Association (ELRA), 1427-1433.

Xiao, Richard (2008): Well-known and influential corpora. *In*: Anke Lüdeling and Merja Kytö, eds. *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: Mouton de Gruyter, 383-457.

Xiao, Richard (2010): Corpus Creation. *In*: Nitin Indurkhya and Fred Damerau, eds. *Handbook of Natural Language Processing (2nd Revised edition).* Connecticut: Taylor & Francis, 147-165.

Zhu, Chunshen (1999): Ut once More: The Sentence as the Key Functional Unit of Translation. *Meta*. 44(3):429-447.

APPENDIX

Corpus size (number of tokens)

Text Type	Translation direction SRC→TGT	DU (n tokens)	EN (n tokens)	FR (n tokens)	TOTAL (n tokens)
Administrative Texts	EN→DU	255,155	246,137	0	501,292
	FR→DU	307,886	0	322,438	630,324
	DU→EN	249,410	257,087	0	506,497
	DU→FR	280,584	0	301,270	581,854
	Total	**1,093,035**	**503,224**	**623,708**	**2,219,961**
External Communication	EN→DU	278,515	272,460	0	550,975
	FR→DU	233,277	0	250,604	483,881
	DU→EN	246,448	255,634	0	502,082
	DU→FR	241,323	0	270,074	511,397
	X→D/E	21,679	20,118	0	41,797
	X→D/E/F	14,192	14,953	15,743	44,888
	Total	**1,035,434**	**563,165**	**536,421**	**2,132,020**
Instructive Texts	EN→DU	340,097	327,543	0	667,640
	FR→DU	40,487	0	42,017	82,504
	DU→EN	19,011	20,696	0	39,707
	DU→FR	110,278	0	115,034	225,312
	X→D/F	59,791	0	73,758	133,549
	X→D/E	299,996	296,698	0	596,694
	X→D/E/F	138,673	145,103	166,836	450,612
	Total	**1,008,333**	**790,040**	**397,645**	**2,196,018**
Journalistic Texts	EN→DU	262,768	264,900	0	527,668
	FR→DU	240,785	0	265,530	506,315
	DU→EN	250,580	259,764	0	510,344
	DU→FR	314,989	0	340,319	655,308
	Total	**1,069,122**	**524,664**	**605,849**	**2,199,635**
Literature	EN→DU	148,488	143,185	0	291,673
	FR→DU	186,799	0	186,620	373,419
	DU→EN	346,802	361,140	0	707,942
	DU→FR	323,158	0	348,343	671,501
	Total	1,005,247	504,325	534,963	2,044,535
Grand Total		**5,211,171**	**2,885,418**	**2,698,586**	**10,795,175**

The total number of tokens is presented per text type and translation direction. The word counts are all based on clean text, i.e without figures, tables and graphs. X stands for unknown source language.