# Data Collection and IPR in Multilingual Parallel Corpora

# Dutch Parallel Corpus

**Orphée De Clercq[1,2] and Maribel Montero Perez[3]**
[1]LT3, University College Ghent
Groot-Brittanniëlaan 45, Gent Belgium
orphee.declercq@hogent.be
[2] Dept. of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281, Gent Belgium
[3]K.U. Leuven - Campus Kortrijk
Etienne Sabbelaan 53, Kortrijk Belgium
MaribelMonteroPerez@kuleuven-kortrijk.be

**Abstract**

Building a corpus is sometimes a difficult and time-consuming. Especially when every text sample included in the corpus had to be cleared from copyrights. In this paper we describe the data collection process of the Dutch Parallel Corpus in close detail and try to formulate some basic principles that might be useful for other project that have corpus compilation as one of its tasks. The Dutch Parallel Corpus is a ten-million-word high-quality sentence aligned corpus for the language pairs English-Dutch-French. It contains five text types and all the text material was cleared from copyrights. Contacting different data providers resulted in drawing up four different licence agreements. Throughout the data collection process some problems were encountered and solutions to these problems were found. All this hard work resulted in the Dutch Parallel Corpus which might help other corpus builders. This is all formulated in the paper below where every step of the data collection process is desribed and illustrated with various examples for future work help.

## 1. Introduction

The creation of a corpus consists of two crucial steps: beside the effort put into data processing, a considerable amount of time is allocated to acquiring text material and clearing copyrights. Data collection can be seen as the crucial starting point in every project that has corpus compilation as one of its tasks and yet there is no universal approach of dealing with this sometimes difficult and time-consuming process.

In this paper we try to formulate some basic data collection principles, based on the experience gained during the creation of the Dutch Parallel Corpus (DPC). Many textbooks have already appeared dealing with corpus linguistics as a discipline. Kennedy (1998), Wynne (2005) and McEnery et al. (2006), to name only a few, all devote one or more chapters to corpus compilation and design principles. Issues of data collection and more specifically copyright clearance, however, are only touched very briefly.

When looking at other parallel corpus projects, existing parallel corpora are either freely available but lack text type balance such as Europarl (Koehn, 2005) or include several text types but are not accessible for the research community due to copyright restictions, e.g. the English-Norwegian corpus (Johansson, 1999/2002).

New corpus compiling methods such as the web as corpus initiative WacKy do not deal with copyrights at all but place information on their website so that anyone offended can request to remove specific documents from the corpora (Baroni et al, 2009).

The Dutch Parallel Corpus does have text type balance and is available for the entire research community. These two objectives were actually the prerequisites of the data collection process, which consisted of two crucial steps:

- Finding potential text providers of high-quality text material that fit in the corpus design and convincing them to participate in the project;
- Obtaining copyright clearance for all texts included in the corpus for both commercial and non-commercial purposes.

Since these two challenges can be transferred to any other corpus project, we try to formulate throughout this paper some general principles about data acquisition and permission clearance that might be re-used in other data acquisition tasks.

The remainder of this article is structured as follows: Section 2. presents the Dutch Parallel Corpus project and describes its balanced design. Section 3. gives an overview of the entire data collection process, copyright clearance and focuses on problems that arise during Intellectual Property Rights (IPR) negotiations. Section 4 concludes the paper.

## 2. Dutch Parallel Corpus

The DPC project was carried out within the framework of the STEVIN programme of the Dutch Language Union (NTU). Since high-quality parallel corpora with Dutch as the central language did not exist before or were not accessible for the research community, due to copyright restrictions, the compilation of aligned parallel corpora was one of its priorities (Odijk et al., 2004).

The Dutch Agency for Human Language Technologies[1] (TST-centrale) is responsible for the distribution of DPC. The finalized Dutch Parallel Corpus is a high-quality annotated parallel corpus of ten million words Dutch,

[1] http://www.tst.inl.nl

English and French. All the text material included in DPC has been standardized, sentence aligned, tokenized and annotated with linguistic information For more information we refer to (Paulussen & Macken, 2010).

### 2.1 Balanced Design

The corpus is balanced in two ways. It contains an equal amount of text material in all four translation directions, with a minimum of 2,000,000 words per translation direction. A small part of the corpus is trilingual. this is represented in Figure 1.

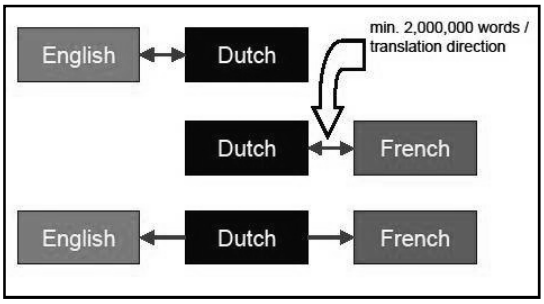



Figure 1: Translation directions

Secondly, the corpus offers a great variety of text material coming from different domains divided into five text types. The typology and structure of the initial design were based on the prototype approach by David Lee. In order to prevent having overbroad categories containing heterogeneous material Lee (2001) advocates using a prototype approach based on the basic-level category and thus creating a multi-level typology. This means introducing subcategories and adding this information to the metadata which allows the user to perform a finer-tuned search.

| [SUPERORDINATE] | [BASIC LEVEL] |
|---|---|
| **1. Literature** | 1.1 Novels |
| | 1.2 Essayistic texts |
| | 1.3 (Auto)biographies |
| | 1.4 Expository works |
| **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 debates |
| | 4.3 Minutes of meetings |
| | 4.4 Yearly reports |
| | 4.5 Official speeches |
| **5. External communication** | 5.1 (Self-)presentation |
| | 5.2 Informative documents |
| | 5.3 Promotion/advertising |
| | 5.4 Press releases |
| | 5.5 Scientific texts |

Table 1: DPC's two-level typology

To this purpose the DPC team has opted for a two-level typology which is presented in Table 1.
All this information is stored in the metadata, where it is further complemented with text and translation-related details such as the intended audience, type of text provider, translation direction and so forth. For a more detailed description of the corpus design, the project goals, applications and functionality we refer to (Rura et al., 2008).

### 2.2 Realized Corpus

The corpus as it was eventually realized in presented in Table 2 on the next page. As one can deduct from this table the finalized corpus contains five text types that each account for 2,000,000 words, namely administrative texts, texts used for external communication, literature (both fiction and non-fiction texts), journalistic texts and instructive texts. Within each text type each translation direction (Dutch-English, English-Dutch, Dutch-French and French-Dutch) contains 500,000 words. This brings the total number of words to ten million.

A closer look at the table reveals that although a balanced composition was achieved, there is one text type - literature - that is not completely balanced according to translation directions and some translation directions are underrepresented - Dutch-English within the Instructive texts - because of source language problems. These are changes to the original design that are all due to problems with data collection and copyright clearance. All of this is discussed in closer detail in the following sections.

## 3. Data Collection and IPR Clearance

An ideal data collection process consists of four steps: (i) a researcher finds adequate text material that should be included in the corpus, (ii) he/she contacts the legitimate author and asks his/her permission (iii), the author agrees and (iv) both parties sign an agreement.
In reality this process is far more complicated and time-consuming, especially in the case of parallel corpus compilation, since more parties are involved (author, translator, publisher, foreign publisher). Negotiations of one or even two years are not exceptional.

### 3.1 Data Collection

Considering the necessity to allocate enough time to data collection (Schuurman et al., 2004) and that the Dutch Parallel Corpus had to be distributable for both commercial and non-commercial purposes, data collection started in the first project term and continued throughout the whole project period.
The first step in the acquisition process consists in deciding where to find adequate text material and whom to contact. Since high quality was one of DPC's parameters we only contacted translation divisions and professional translator. Another objective was that for each text type a minimum of three different text provider had to be persuaded. Following the corpus design we could make a division between two main data sources:

| Text Type | SRC→TGT | DU | EN | FR | TOTAL | % |
|---|---|---|---|---|---|---|
| **Administrative Texts** | EN→DU | 255,155 | 246,137 | 0 | 501,292 | 100.26 |
| | FR→DU | 307,886 | 0 | 322,438 | 630,324 | 126.06 |
| | DU→EN | 249,410 | 257,087 | 0 | 506,497 | 101.30 |
| | DU→FR | 280,584 | 0 | 301,270 | 581,854 | 116.37 |
| | Total | 1,093,035 | 503,224 | 623,708 | 2,219,961 | 111.00 |
| **External Communication** | EN→DU | 278,515 | 272,460 | 0 | 550,975 | 110.19 |
| | FR→DU | 233,277 | 0 | 250,604 | 483,881 | 96.78 |
| | DU→EN | 246,448 | 255,634 | 0 | 502,082 | 100.42 |
| | DU→FR | 241,323 | 0 | 270,074 | 511,397 | 102.28 |
| | XDE- | 21,679 | 20,118 | 0 | 41,797 | 8.36 |
| | XDEF- | 14,192 | 14,953 | 15,743 | 44,888 | 8.98 |
| | Total | 1,035,434 | 563,165 | 536,421 | 5,132,020 | 106,75 |
| **Instructive Texts** | EN→DU | 340,097 | 327,543 | 0 | 667,640 | 133.53 |
| | FR→DU | 40,487 | 0 | 42,017 | 82,504 | 16.50 |
| | DU→EN | 19,011 | 20,696 | 0 | 39,707 | 7.94 |
| | DU→FR | 110,278 | 0 | 115,034 | 225,312 | 45.06 |
| | XD-F | 59,791 | 0 | 73,758 | 133,549 | 27.71 |
| | XDE- | 299,996 | 296,698 | 0 | 596,694 | 119.34 |
| | XDEF | 138,673 | 145,103 | 166,836 | 450,612 | 90.12 |
| | Total | 1,008,333 | 790,040 | 397,645 | 2,196,018 | 109.80 |
| **Journalistic Texts** | EN→DU | 262,768 | 264,900 | 0 | 527,668 | 105.53 |
| | FR→DU | 240,785 | 0 | 265,530 | 506,315 | 101.26 |
| | DU→EN | 250,580 | 259,764 | 0 | 510,344 | 102.07 |
| | DU→FR | 314,989 | 0 | 340,319 | 655,308 | 131.06 |
| | Total | 1,069,122 | 524,664 | 605,849 | 2,199,635 | 109.98 |
| **Literature** | EN→DU | 148,488 | 143,185 | 0 | 291,673 | 58.33 |
| | FR→DU | 186,799 | 0 | 186,620 | 373,419 | 74.68 |
| | DU→EN | 346,802 | 361,140 | 0 | 707,942 | 141.59 |
| | DU→FR | 323,158 | 0 | 348,343 | 671,501 | 134.30 |
| | Total | 1,005,247 | 504,325 | 534,963 | 2,044,535 | 102.23 |
| **Grand Total** | | **5,211,171** | **2,885,418** | **2,698,586** | **10,795,175** | **107.95** |

Table 2: Number of words included in DPC according to text type and translation direction

Institutions for finding the first three text types: administrative texts, texts treating external communication and instructive texts and commercial publishers for finding journalistic texts and literature (fiction and non-fiction).

The same division is relevant when describing the difficulties that were encountered during data collection. While institutions produce texts to inform and help their customers, commercial publishers publish text material as a core business. Institutions were thus more easily persuaded to hand over text material than commercial publishers, who are on the alert for undesired competition.

#### 3.1.1. Institutions

In a bilingual country like Belgium, every text has to be available in both Dutch and French. Many multinational companies also have a local branch in Flanders or the Netherlands so much English text material is translated into Dutch. Due to this high level of multilingualism we were able to find enough translated text material for the first three text types with Dutch both as a source and target language.

The instructive texts posed the first problem. Although it was rather easy to convince multinationals to include instructive texts in the corpus, it was difficult to find out in which language the text had originally been written. This led to a first adaptation of the original corpus design: it was decided to loosen the balance between the translation directions for this text type.

#### 3.1.2. Commercial publishers

Convincing publishers to participate in the project was the most difficult part of the entire acquisition process. There are at least four interest groups: the author, the publisher, the translator and the foreign publisher. This involved a complex and lengthy negotiation process, in one case continuing for more than a year and a half. For journalistic texts, we achieved the quantitative goals set out in the corpus design only in the last month of the project.

Acquiring fictional texts is an equally hard task (Geyken, 2007). Nearly every single publishing house in Belgium and the Netherlands was contacted, but the problem is bringing the request of the corpus compilers on the desk of the deciders, which turned out to be virtually

impossible without a helping hand from 'above': some high officials of the DLU had to be called in for help before a breakthrough could be achieved.
The difficulties encountered in the acquisition of fictional literature, forced us to adapt the original design for a second time. Instead of having two literary text types (fiction and non-fiction), we had to bring together fictional and non-fictional literature in one group and partially loosen the balance between the translation directions.

### 3.2 Copyright Clearance

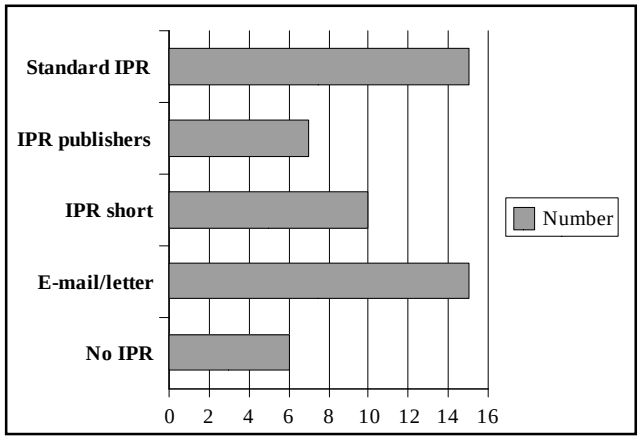
The acquisition process can only be concluded when permission clearance is obtained for text X furnished by data provider Y.
A clear definition of copyright in corpus-building can be found in Baker et al (2006): "The right to publish and sell literary, musical or artistic work. Corpus compilers need to observe copyright law by ensuring that they seek permission from the relevant copyright holders to include particular texts." Kennedy (1996) states that most copyright holders are willing to donate text for research purposes and in McEnery (2006) we read that there is no satisfactory solution to the issue of copyright in corpus-building.
Since we were dealing with copyright for building a corpus that had to be available for the whole research community but that also has commercial purposes, copyright clearance had to be obtained for all samples included in the corpus.
To this purpose we used Intellectual Property Rights (IPR) agreements to avoid later discussions. These license agreements need to guarantee accessibility and protect the intellectual and economic property rights of the authors and publishers. They were developed in close collaboration with the Agency for Human Language Technologies (HLT). Since DPC aimed at creating a well-balanced corpus, we contacted different groups of providers (cf. *supra*), which is inevitably reflected in the typology of licence agreements. Four standard licence agreements were drawn up. These agreements all have some features in common:

- Texts will not be altered, only metadata and language technology related information will be added (part-of-speech tags).
- Text material will be available both for commercial and non-commercial purposes.

We will now describe each agreement separately by pointing out the main differences.

1. **Standard IPR**: this is the standard version of the IPR agreement that was used for most text providers. It is about ten pages long and contains the nuts and bolts of permission seeking. As a way to reassure the text provider it was clearly stated that we intend no competition and that commercial use is only possible when the text material is not recognizable as such. This means that all the text material can only be accessed via the corpus and the text cannot be downloaded as such.
2. **IPR for publishers**: this is in sensu large the same agreement as the one for commercial use but here the texts also had to be made partially recognizable for non-commercial purposes. Since most publishers feared competition this feature was added to make it acceptable for them.
3. **IPR short version**: while the negotiations proceeded we became aware that the standard agreement was a bit to long and frightened some text providers Therefore this short version of the standard IPR agreement was drawn up to simplify and accelerate the negotiations.
4. **E-mail or letter with permission**: when a data provider wanted to participate in the project but was unable to sign an agreement stating this, it was decided that an e-mail or letter with permission could also be accepted. This was only possible in particular cases and when few text material was involved.

Beside these four agreements, some of the DPC text material could be integrated without an IPR agreement at all because the texts belong to the public domain. These texts can be published or copied, subject to acknowledgement of the source. Graph 1 presents the distribution of all agreements that were concluded during DPC.




Graph 1: Number of concluded agreements per IPR type

The short IPR agreement and the one for publishers are now also used in other corpus projects of the STEVIN programme, such as SoNaR[2] .

### 3.3 General Principles

There is no such thing as a universal recipe for the acquisition of text material and for obtaining permission clearance. Some guidelines already exist (Kilgariff, 2002), but we believe that, thanks to the experience gained within the DPC project, some more practical guidelines for the optimisation of data collection can be added.
Start data collection from day one, some negotiations

[2] SoNaR aims at developing a reference corpus for the Dutch language of 500 million words.

might take years;
Give sufficient information about the project itself and try to find examples that illustrate the need for data;
Stress the importance of available data for all kinds of purposes;
Use a different approach and IPR agreement for publishers and institutions ;
Be patient, repeating the same thing over and over again might be frustrating but is necessary;
Negotiate, the final product can also be useful for the institution or commercial provider;
Use influential partners (in our case: the Dutch Language Union) to negotiate on a high level.

## 4. Conclusion

Collecting text material and clearing copyrights is a difficult and time-consuming step in every corpus project. Thanks to the effort put in the collection of data and IPR for the DPC project, all copyright issues have been solved.
During the creation of DPC, valuable experience was gained for five different text types covering both the profit and non-profit sector. We hope that this experience and the principles described above might be useful for future corpus projects.

## 5. Acknowledgements

The DPC project was carried out within the STEVIN programme, which is funded by the Dutch and Flemish governments. Grant number STE-05-26. Its main collaborators were Piet Desmet (coordinator), Hans Paulussen, Maribel Montero Perez (K.U.Leuven Campus Kortrijk), Willy Vandeweghe (co-coordinator), Lieve Macken and Orphée De Clercq (School of Translation Studies, University College Ghent).

## 6. References

Baker, P., Hardie A. and McEnery T. (2006). A Glossary of Corpus Linguistics. Edinburgh: Edinburgh Univeristy Press, pp. 48

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3), pp. 209--226.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Felbaum (Ed.), *Collocations and Idioms: Linguistic, Lexicographic and Computational Aspects.* London, Continuum Press, pp .

Johansson, S., Ebeling, J., Oksefjell, S. (1999/2002). English-Norwegian Parallel Corpus: Manual. URL: http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html

Kennedy, G. (1998). An Introduction to Corpus Linguistics. London and New York: Longman, pp 76--78.

Kilgariff, A. (2002). Legal aspects of corpora compiling in *Corpora List Archive*. URL: http://helmer.hit.uib.no/corpora/2002-3/0253.html.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation in *Tenth Machine Translation Summit*, Phuket, Thailand, pp. 79--86.

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

McEnery, T., Xiao, R. and Yukio Tono (2006). Corpus-Based Language Studies: an advanced resource book. Oxon and New York: Routledge, pp. 77-79.

Odijk, J., Martens, J-P., van Eyde, F., Daelemans, W., Kenyon-Jackson, D., Vossen, P., van Hesse, A., Boves, L. and Beeken, J. (2004). Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie. STEVIN. Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands. The Hague: Nederlandse Taalunie.

Rura, L., W. Vandeweghe & M. Montero Perez (2008). Designing a parallel corpus as a multifunctional translator's aid in *Proceedings of XVIII FIT World Congress*, 4-7 August 2008, Shanghai.

Schuurman, I., W. Goedertier, H. Hoekstra, N. Oostdijk, R. Piepenbrock, M. Schouppe, (2004) Linguistic annotation