Data Collection and IPR in Multilingual Parallel Corpora
Dutch Parallel Corpus

Orphée De Clercq[1,2] and Maribel Montero Perez[3]
LT3, Language and Translation Technology Team, University College Ghent[1]
Groot-Brittanniëlaan 45, 9000 Gent Belgium
orphee.declercq@hogent.be
Dept. of Applied Mathematics and Computer Science, Ghent University[2]
Krijgslaan 281 (S9), 9000 Gent Belgium
K.U. Leuven - Campus Kortrijk[3]
Etienne Sabbelaan 53, 8500 Kortrijk Belgium
maribel.monteroperez@kuleuven-kortrijk.be

Abstract

After three years of work the Dutch Parallel Corpus (DPC) project has reached an end. The finalized corpus is a ten-million-word high-quality sentence-aligned bidirectional parallel corpus of Dutch, English and French, with Dutch as central language. In this paper we present the corpus and try to formulate some basic data collection principles, based on the work that was carried out for the project. Building a corpus is a difficult and time-consuming task, especially when every text sample included has to be cleared from copyrights. The DPC is balanced according to five text types (literature, journalistic texts, instructive texts, administrative texts and texts treating external communication) and four translation directions (Dutch-English, English-Dutch, Dutch-French and French-Dutch). All the text material was cleared from copyrights. The data collection process necessitated the involvement of different text providers, which resulted in drawing up four different licence agreements. Problems such as an unknown source language, copyright issues and changes to the corpus design are discussed in close detail and illustrated with examples so as to be of help to future corpus compilers.

1. Introduction

The creation of a corpus consists of two crucial steps: besides the effort put in data processing, a considerable amount of time has to be devoted to acquiring text material and clearing copyrights. Although data collection is the crucial starting point in every project of corpus compiling, there is currently no universal approach of dealing with this task (Xiao, 2010:152).

This paper is an attempt 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 been published on 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 on very briefly.

When examining existing parallel corpora, we notice that some are freely available but lack text type balance such as Europarl (Koehn, 2005), and that others include several text types but are not freely accessible to the research community due to copyright restrictions, 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 to the best of our knowledge, instead the project websites are usually provided with a notice that anyone offended can file a request for removing specific documents from the corpora (Baroni et al, 2009).

The Dutch Parallel Corpus does exhibit 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 providers of high-quality text material, i.e. published and/or revised by professional translation services, which fits 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 attempt to formulate in this paper some general principles about data acquisition and permission clearance that might be re-used in other projects involving data acquisition.

The remainder of this paper 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 Nederlandse Taalunie (NTU: Dutch Language Union). The compilation of aligned parallel corpora was one of the programme's priorities, since high-quality parallel corpora with Dutch as the central language were scarce and, if existing, not accessible to the research community, due to copyright restrictions (Odijk et al., 2004). The Dutch Human

Language Technology Agency (HLT-agency) [1] is responsible for the distribution of the DPC.

The finalized Dutch Parallel Corpus is an annotated ten-million-words parallel corpus of Dutch, English and French. All the text material included in DPC has been standardized, sentence aligned, tokenized and annotated with linguistic information (lemmata and part-of-speech tags).

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.

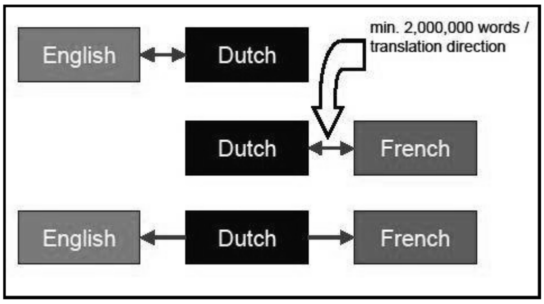


Figure 1: Translation directions

Secondly, the corpus offers a great variety of text material coming from different domains divided into five major text types.

[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

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 fine-tune his/her search. This approach led the DPC team to opt for a two-level typology as 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 (Macken et al.,2007 and Rura et al.,2008).

2.2 Final Corpus

The corpus as it was eventually composed is presented in Table 2. The word counts are based on cleaned data (i.e. text of which all figures, tables and graphs are removed). As can be deducted from this table the finalized corpus contains five text types represented with around 2,000,000 words each: administrative texts, texts used for external communication, literature (both fiction and non-fiction), journalistic texts and instructive texts. Within each text type, each translation direction (Dutch-English, English-Dutch, Dutch-French and French-Dutch) is represented by 500,000 words. This brings the total number of words to 10,795,175 words, exceeding our prime objective of ten million words.

A closer look at Table 2 reveals that, although a balanced composition was achieved, there is one text type - literature - that is not completely balanced over the translation directions. Some translation directions are underrepresented - such as Dutch-English within the Instructive texts - because of source language problems. These are changes to the original design that can all be attributed 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

One could expect a data collection process to consist of four steps: (i) a researcher finds adequate text material to 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 there is considerably more to it, especially in the case of parallel corpus compilation, since more parties are involved (author, translator, publisher, foreign publisher). Negotiations on IPR matters may drag on for months and exceptionally even years.

3.1 Data Collection

Considering the necessity to allocate enough time to data collection (Schuurman et al., 2004) and in view of the fact that the Dutch Parallel Corpus had to be distributable for

[1] The Dutch-Flemish agency for management, maintenance and distribution of Dutch digital language resources. See http://www.tst.inl.nl

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
	X→D/E	21,679	20,118	0	41,797	8.36
	X→D/E/F	14,192	14,953	15,743	44,888	8.98
	Total	1,035,434	563,165	536,421	2,132,020	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
	X→D/F	59,791	0	73,758	133,549	27.71
	X→D/E	299,996	296,698	0	596,694	119.34
	X→D/E/F	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

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 obtaining high quality translations was important we first contacted translation divisions and professional translators. Another objective we formulated was to collect at least three different text providers per text type . Following the design we could make a division between two main data sources, *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 (fictional as well as non-fictional) literature.

The same division is relevant when describing the difficulties 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. A different approach was thus necessary to be able to persuade both parties to participate.

3.1.1. Institutions

In a bilingual country like Belgium, many official texts need to be available in both Dutch and French. As most multinationals also have a local branch in Flanders or the Netherlands, a lot of English text material gets translated into Dutch. Due to this high level of multilingualism, we were able to collect sufficient 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 grant permission for including instructive texts in the corpus, it sometimes proved hard or even impossible to find out in which language a particular 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. As can be seen in Table 2, for approximately one million words (is 10%) it was impossible to establish the source language (SRC = X).

3.1.2. Commercial publishers

Convincing publishers to participate in the project was the most difficult part of the entire acquisition process. The reasons for this can be grasped intuitively: for publishers, producing text is the main source of income, a core activity. Therefore, they are on their guard against illegitimate competition.

When negotiating with publishers there are at least four stakeholders: the author, the home publisher, the translator and the foreign publisher. This brings along a complex and lengthy negotiation process. In one case, negotiations lingered for more than a year and a half.

For journalistic texts, we were unable to meet the quantitative goals set out in the corpus design until the last month of the project.

Acquiring fictional texts is equally laborious (Geyken, 2007). Nearly every single publishing house in Belgium and the Netherlands was contacted, but the difficult part is getting the request by the corpus compilers to the desk of the deciders, which turned out to be virtually impossible without a helping hand from 'above': some high officials of the DLU[2] had to be called in for help before a breakthrough could be achieved.

The difficulties encountered during the acquisition of fictional literature, forced us to adapt the original design 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. As shown in Table 2, it was easier to convince local publishers than foreign ones. For the text type Literature the translation directions English-Dutch and French-Dutch are slightly underrepresented as a consequence.

3.2 Copyright Clearance

A clear definition of copyright in the context of 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."

This means that the data collection process can only be concluded when permission clearance is obtained for text X furnished by data provider Y.

Kennedy's (1996) statement that most copyright holders are willing to donate texts for research purposes, was not borne out by our own experience. The requirement that the corpus had to be available for both research and commercial purposes made most copyright holders reticent to donate text and certainly to sign an agreement in which this was stated explicitly.

Drawing up contracts on Intellectual Property Rights (IPR) was nevertheless necessary to avoid later discussions. Templates for these contracts were developed in close collaboration with the Dutch Human Language Technology Agency (HLT). Care was taken to both guarantee accessibility on the one hand, and protect the intellectual and economic property rights of the authors and publishers on the other hand.

Since creating a well-balanced corpus was one of the project ambitions, we contacted different groups of providers (cf. supra). This heterogeneity is inevitably reflected in the typology of licence agreements. Four standard licence agreements were drawn up: a standard agreement, one for publishers, a short version to speed up the negotiations and an e-mail or letter with permission.

- **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 arranges every possible dispute[3]. In order to reassure the text provider, it clearly stated that no competition is intended and that commercial use presupposes the text material to be unrecognizable as such. This implies that all text material can only be accessed via the corpus and that the text cannot be downloaded as such by the end user.

- **IPR for publishers**: this agreement is similar to the standard one for commercial use, but here the texts also have to be made partially recognizable for non-commercial purposes, which implies that also for research purposes the texts are only accessible by means of the corpus. Since most publishers feared undue competition, this feature was added to make it acceptable for them.

- **IPR short version**: while the negotiations proceeded we became aware that the standard agreement was a bit too long and that too much information on possible infringements was included, which alarmed some text providers. Therefore a short version of the standard IPR agreement was drawn up to simplify and accelerate negotiations by avoiding lengthy contract stipulations.

- **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 exceptional cases and when little text material was involved.

Aside from the text material subject to some form of agreement, quite some text material could be integrated without an IPR agreement at all because it belongs to the public domain. These texts can be published or copied, subject only to acknowledgement of the source.

In order to protect our text providers the corpus will be made available for end users by the Dutch HLT-agency

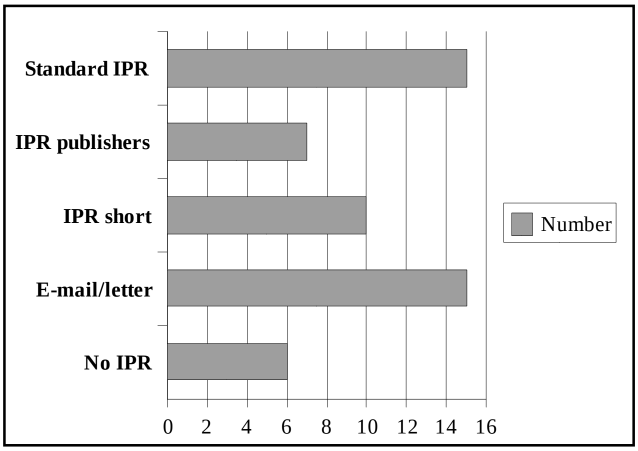
[2] Dutch Language Union/Nederlandse Taalunie was founded in 1980 to allow The Netherlands, Flanders and Suriname to cooperate on linguistic issues, language policy, language teaching and literature. The DLU can thus be perceived as an influential partner.

[3] http://taalunieversum.org/taal/technologie/stevin/documenten/model_C1.doc (example in Dutch).

after signing a user agreement.

After signing this agreement the corpus will be accessible either as a full text resource in XML-encoded format or by means of a password protected web interface that was specially developed for the Dutch Parallel Corpus by the Leuven Language Institute[4] of the K.U. Leuven.

Graph 1 presents the number of signed contracts for text material falling under one or other sort of agreement or under no agreement at all.



Graph 1: Quantitative proportions of agreement status

The short IPR agreement and the one for publishers were adopted in other corpus projects of the STEVIN programme, such as SoNaR. SoNaR aims at collecting a 500-million-word Dutch reference corpus, completely cleared from copyrights. For more information on this project we refer to (Oostdijk et al., 2008) and (Reynaert et al., 2010).

3.3 General Principles

There is no such thing as a universal recipe for the acquisition of text material and for obtaining permission clearance. As stated above, many textbooks exist that deal with corpus compilation and design, but when it comes to data collection and copyright clearance much is left to the inspiration of the individual corpus builders.

We did find some guidelines (Kilgariff, 2002), but these only confirm that copyright law is still in its infancy when it comes to corpus building and that there is very little which is obviously legal or illegal. A widely accepted word of advice is that, whenever in doubt, seek permission (Xiao, 2010:153).

We believe that, thanks to the experience gained during the DPC project, some more practical guidelines for the optimisation of data collection and some useful advice can be formulated, this is done in Table 3.

This is not an exhaustive list and during corpus compilation many different situations arise, but we hope that this paper will help colleagues with some issues and difficulties they experience when building corpora and clearing text samples from copyrights.

Practical Guidelines
Start data collection from day one, some negotiations might take years; Give sufficient information about the project 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 on IPR agreement for publishers and for institutions; Contact multinationals, they often have a local branch and thus translate text into foreign languages; If possible, try to win high-level influential partners to facilitate negotiations. Be sure there is a budget for cases in which licences have to be paid for.
Advice
Be patient, repeating the same thing over and over again might be frustrating but is necessary; Use the argument that the final product may also be useful for the text provider; Anticipate possible questions ranging from What is a corpus? to How much money is my diary worth?; Remain polite, you are asking a favour for which (in most cases) no compensation is given.

Table 3: Practical guidelines and advice for data collection

4. Conclusion

Collecting text material and clearing copyrights is a difficult and time-consuming step in every corpus project for which there are no universal or clear-cut rules. Thanks to the experience that was gained during the DPC project, we were able to solve most copyright problems we were confronted with, and managed to collect a ten-million word parallel corpus Dutch-English-French that is available for research and commercial purposes.

Although changes and concessions had to be made in the design because of data collection problems, in the end we did manage to create a balanced corpus. Successful negotiation of copyright issues depends largely on using different agreements for different text providers.

It is our hope that the experience and the principles described above may 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, under 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

P. Baker, A. Hardie, and T. McEnery. 2006. *A Glossary of*

[4] http://ilt.kuleuven.be/english/

Corpus Linguistics. Edinburgh University Press, Edinburgh.

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

A. Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Felbaum, editor, *Idioms and Collocations: Corpus-based Linguistic and Lexicographic* Studies, pages 23-40. Continuum International Publishing Group, London.

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

G. Kennedy. 1998. *An Introduction to Corpus Linguistics*. Longman, London, New York.

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

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, Phuket, Thailand, pages 79-86.

D.Y.W. Lee. 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): 37-72.

L. Macken, J. Truskina, H. Paulussen, L. Rura, P. Desmet, and W. Vandeweghe. 2007. Dutch Parallel Corpus: A multilingual annotated corpus. In *On-line Proceedings of Corpus Linguistics*, July 2007, Birmingham, United Kingdom.

T. McEnery, R. Xiao, and Y. Tono. 2006. *Corpus-Based Language Studies: an advanced resource book*. Routledge, Oxon, New York.

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

N. Oostdijk, M. Reynaert, P. Monachesi, G Van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste. 2008. From DCoi to SoNaR: a reference corpus for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, May, Marrakech, Morocco. European Language Resources Association http://www.lrefconf.org/proceedings/lrec2008.

M. Reynaert, N. Oostdijk, O. De Clercq, H. Van den Heuvel, and F. de Jong. 2010. Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC-2010),* Valletta, Malta.

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

I. Schuurman, W. Goedertier, H. Hoekstra, N. Oostdijk, R. Piepenbrock, and M. Schouppe. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again… In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, Volume I pages 57-60.

R. Xiao. 2010. Corpus Creation. In N. Indurkhya, F. Damerau, editors, *Handbook of Natural Language Processing (2n Revised edition),* pages 147-165. Taylor & Francis, Connecticut, Cambridge.