Dutch Parallel Corpus: A Multilingual Annotated Corpus

Lieve Macken,[1] Julia Trushkina,[2] Hans Paulussen,[2]
Lidia Rura,[1] Piet Desmet[2] and Willy Vandeweghe[1]

1 Introduction

Aligned parallel corpora form an indispensable resource for a wide range of multilingual applications, including, among others, machine translation (MT), especially corpus-based MT like statistical MT (Koehn, 2005) and example-based MT (Carl and Way, 2003), computer-assisted translation tools (Hutchins, 2005), multilingual information extraction and computer-assisted language learning (Desmet and Paulussen, 2005).

Apart from the more technological applications, parallel corpora can be used to conduct more fundamental research in the fields of contrastive linguistics and translation studies (Baker, 1995; Laviosa, 2002; Olohan, 2004).

Since high-quality parallel corpora with Dutch as a central language do not exist or are not accessible for the research community due to copyright restrictions, the compilation of aligned parallel corpora with Dutch as a central language was one of the priorities of the STEVIN program (Odijk et al., 2004).

The Dutch Parallel Corpus (DPC) project aims at fulfilling this need. Within the DPC project, a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French is being compiled. The corpus will be enriched with linguistic annotations: part-of-speech and lemmatization information for the whole corpus and syntactic analysis for a subpart of the corpus.

As the corpus 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). A part of the corpus is trilingual and contains Dutch texts translated into both English and French.

To guarantee the quality of the corpus and its multifunctional availability for the wide research community, each step in compiling, structuring and annotating the corpus is being validated by a user group of specialists in linguistics and language technology. In order to make the corpus accessible for the whole research community, a copyright clearance for all samples included in the corpus is being obtained.

The DPC-project started in May 2006 and runs until March 2009.

The remainder of the paper is organized as follows: Section 2 deals with the specific needs of the different intended users. Section 3 describes in detail the corpus design. Section 4 focuses on the more technical issues of text normalization, alignment and linguistic annotation. Section 5 elaborates on quality control and Section 6 concludes the paper.

[1] LT[3], University College Ghent, Belgium
e-mail: firstname.lastname@hogent.be
[2] KU Leuven – Campus Kortrijk, Belgium
e-mail: firstname.lastname@kuleuven-kortrijk.be

2 Intended users

Aligned parallel corpora form an indispensable resource for a wide range of multilingual applications and can be used in different research fields. Generally speaking, four categories of users can be distinguished: linguists conducting more fundamental research in contrastive linguistics and translation studies, language teachers and language learners, human translators, and developers of HLT-applications.

Each of these four groups has its own requirements relating to corpus design, kind and degree of annotation and required metadata of a parallel corpus. Since the development of a parallel corpus is time-consuming and costly, the DPC project aims at the creation of a multifunctional resource to satisfy the needs of the diverse group of potential users.

2.1 Contrastive Linguistics and Translation Studies

Large aligned parallel corpora represent a valuable source of empirical material both for Contrastive Linguistics and Translation Studies. Contrastive Linguistics focuses on the translation product and tries to discern similarities and differences between the language systems by comparing two parallel texts. The generalized conclusions drawn from this comparative study can be used among other things for developing new, more effective methods for second-language teaching.

In Translation Studies, parallel texts are mostly used for the descriptive analysis of the translation process and its impact on the final linguistic product, for analyzing different translation strategies, the motivation behind choices in translation and the role of translation as a means of communication and mediation between cultures.

Interestingly, the DPC can also be used as a monolingual comparative corpus as it will contain original Dutch texts alongside with translated Dutch texts, which can be compared with each other. Translated texts tend to show certain idiosyncrasies, like deviations from the target language conventions under the source language influence, but also the effacing of creative elements from the original. These idiosyncrasies are also known as ‘translation universals’, referring to the features typical of translated texts and inherent in the translation process that are found in translations from and into different languages regardless of the text type. Baker (1996) mentions four features typical of a translated text: ‘simplification’ of the language or the message, ‘explicitation’, ‘normalization’, i.e. using only typical patterns of the target language and ‘levelling out’ variations in the source text by converging towards the middle.

For the corpus to be suitable for the above-mentioned purposes, it has to contain high-quality representative data, be balanced by translation direction and contain information on the source/target language of the texts included in the corpus. This information will be available for the major part of the DPC and, where possible, be complemented by information on the translation mode (e.g. human translator, translation memory or machine translation).

2.2 CorpusCALL

In the context of language learning in general and CALL (Computer Assisted Language Learning) in particular, corpora have taken up an important position as resources for samples of authentic language usage. Initially, corpus samples were mainly used for the preparation of language exercises, but nowadays such samples are also used during the language learning activity itself (e.g. presenting the topic of a reading or writing exercise), or they illustrate typical language use in a feedback activity. Corpus samples are also the basic resources for language learning reference materials, including both learner dictionaries and learner grammars. The need for corpora in CALL applications has become so important, that a few years ago a new special interest group on corpusCALL has been created within EuroCALL (European Association for Computer Assisted Language Learning), the main European organisation involved in the promulgation of the development of ICT for language teaching.

Originally, the focus in CALL was on monolingual corpora, but there is nowadays also a growing need for parallel corpora. Compared to monolingual corpora, parallel corpora have the advantage of providing illustrative examples, which can only be retrieved from translated texts. When selecting a key word in the source language, one can retrieve a set of examples grouping synonymous words in the target language, thus showing a better illustration of the different usages of a word in the other language. An illustration of the use of parallel corpora in a didactic context is the NEDERLEX project (Deville, Dumortier and Paulussen, 2004), an electronic reading platform of Dutch texts for French speakers. The quality of this type of illustration material depends a lot on the size of the parallel corpora available, which explains why also in language learning and CALL larger parallel corpora are required.

Although quality of texts and translations is a requirement for of all of the above-mentioned applications, it is especially visible in the context of CALL applications. In some of the other applications, a parallel corpus is often considered the basic training material for improving other programs. The corpus as such is less visible to the end-user, or not visible at all. On the other hand, when using parallel corpora in CALL applications, the quality of the text, the translation and the alignment is directly accessible to the end-user. Moreover, the end-user is a language learner who is not a language specialist, and can only up to a certain extent evaluate the quality of the proposed samples. Therefore, quality of text samples and translations plays an important role in the compilation of a parallel corpus used for foreign language learning.

2.3 Full text corpora as translator's aid

Bilingual concordancing systems allow human translators to query a large corpus of aligned translated material. The sentences matching the search query are retrieved and displayed together with their aligned translation. Bilingual concordance searches can be seen as complementary to bilingual dictionaries, as they give valuable context information.

Concordance search capabilities are often integrated in Translation Memory systems, but they are also useful as stand-alone tools. The analysis of the TransSearch log files (Simard and Macklovitch, 2005) has shown that parallel corpora as such are a useful resource for professional translators to solve translation difficulties.

According to Simard and Macklovitch, TransSearch processes thousands of queries every day, submitted by professional English-French translators (the Canadian Government's Translation Bureau is one TransSearch users). Multitrans (Gervais, 2003) is another example of a translation support tool based on a repository of full text translations.

Full text parallel corpora are extremely useful for translators as they can retrieve translations of words in context. Human translators are very demanding users of a parallel corpus and expect high-quality translations and high-quality alignments. Information on translation direction does not seem that important for them.

2.4 HLT applications - Machine Translation

Aligned parallel corpora form an indispensable resource for the data-driven (mainly statistical) development of a wide range of multilingual applications, among others cross-lingual information extraction, multilingual terminology extraction, and machine translation.

Aligned parallel corpora are used in MT as training and test material for corpus-based MT systems (Statistical MT or Example-based MT). The most wide-spread parallel corpora used in MT cover a small set of domains or text types, and mostly contain texts of governments of multilingual countries, such as Canada (the Hansard Corpus English/French, consisting of the proceedings of the Canadian Parliament), or multinational institutions such as the United Nations (UN Parallel Text English/French/Spanish, containing archive documents of the Office of Conference Services in the period between 1988 and 1993) or the European institutions (Erjavec et al., 2005; Koehn, 2005).

There is a need for more diversity in the types of texts compiled. Macken (2007) examined the problem of translational correspondence in different text types (user manuals, press releases and proceeding of plenary debates) in view of different heuristics used in existing sub-sentential alignment modules. She showed that for certain text types, it is sufficient to focus on contiguous translation units of maximally three words. However, the problem of translational correspondence was found to be more complex in text types where a freer or more target language-oriented translation style was adopted.

For statistical systems to be successful, they need a large amount of data. Given their size, most parallel corpora used for MT purposes are aligned automatically without manual verification. This does not pose a problem, as the alignment errors will be filtered out in the statistical process. Most parallel corpora used for the development of MT systems do not take into account the notion of translation direction, and often make use of indirect translations.

3 Corpus design

The design principles of the DPC corpus were based on two sources: on the one hand, the information available about other parallel corpus projects, and on the other hand the user requirements study, which was carried out within the DPC project.

To identify the requirements of the user group with respect to corpus design, a questionnaire was put online on the DPC-website. All members of the predefined user group - which is composed of academic and industrial specialists from different

application and research domains - were asked to fill in the form. In addition, other interested parties were invited to participate. In total 34 respondents completed the questionnaire.

The analysis confirmed a strong need for a parallel corpus with Dutch as central language. The analysis also showed that the quality of text materials as well as the quality of alignments and linguistic annotations are crucial for users of corpus applications. The users opted for a high variety of text types and rich metadata, and, in general, stated that inclusion of full texts is not a necessary condition for them as long as fragments of different text types are present.

Based on the user requirements analysis, motivated choices were made regarding the balancing criteria, text typology, sampling criteria, and kind and degree of annotations and required metadata. The details are presented below.

3.1 Language pairs and translation directions

The DPC corpus consists of two language pairs: Dutch-English and Dutch-French and is bi-directional (Dutch as a source and a target language). A part of the corpus will be trilingual and will contain Dutch texts translated into both English and French (see Table 1).

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

Table 1: DPC translation directions

As mentioned in section 2.1, especially for Translation Studies - where the translation process is being studied - translation direction is an important balancing criterion. The corpus will be balanced proportionally with respect to language pairs and translation directions. For this purpose the target figure of minimally 2 million words per translation direction has been set.

3.2 Text types and text providers

The corpus is designed to represent as wide a range of translated Dutch texts as possible. In order to get a well-balanced corpus, texts are selected from different domains. Thus, the DPC corpus will not only be balanced according to the language pair and translation direction but to the text type as well. However, it cannot be ignored that obtaining enough material for certain text types in some translation directions may prove extremely problematic. For example, newspaper material is hardly ever translated from Dutch into English.

The data in the corpus originates from two main sources: commercial publishers and institutions (both profit and non-profit). This division was used to separate the text material into two big groups according to the type of text provider. Each group has been subsequently divided into several text types but the criteria for this division are not of the same nature. Those coming from commercial publishers

are recognised genres: literature and journalistic texts. The institution texts were divided on the basis of their function and purpose: they instruct, document, inform and/or persuade.

In total, the corpus will contain the following six text types:

- Commercial publishers:
  - Fictional literature
  - Non-fictional literature
  - Journalistic texts
- Institutions:
  - Instructive texts
  - Administrative texts
  - External communication

The obtained six subtypes have been further subdivided into subtypes in order to create a finer tree-like structure within each type. However, this subdivision has no implication for the balancing of the corpus. It is merely a way of mapping the actual landscape within each text type and assigning accurate labels to the data in order to enable the user to correctly retrieve documents and navigate the corpus.

The subdivision is based on the prototype approach advocated by David Lee (Lee, 2001), and the subtypes are chosen from 'basic-level categories', a notion coming from cognitive linguistics and indicating cognitively salient and identifiable concepts, encountered in every-day language usage and easily distinguishable to the corpus compilers as to the corpus user (Ungerer and Schmidt, 1996). The result is a two-level typology, where the six main types represent superordinates each containing several basic-level categories.

For instance, non-fictional literature is an umbrella category uniting three basic-level categories: essays, (auto)biographies and expository works of a general nature; instructive texts consist of three basic-level categories as well, united as the word indicates by the instructive purpose of the document: manuals, legal documents (e.g. contracts, conditions, regulations etc) and procedure descriptions, i.e. documents dealing with all kinds of procedures.

Besides the subtypes, the metadata will also contain extra information concerning the intended audience (broad external audience, limited internal audience, specialists), the type of the text provider (profit vs. non-profit) and domains with keywords indicating the field of human activities or the branch of science, to which a particular document belongs. This information, just as the subtypes, is only meant as an additional navigation tool, to enable the user to select documents of a certain type and has no bearing on the balancing of the corpus.

To guarantee the quality of the text samples, most of them come from published materials or from companies or institutions working with a professional translation division. The texts are selected from different types of data providers. These include providers from publishing houses, press, government, commercial companies and content brokers.

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

Collecting data thus presents two major challenges: persuading the text provider to participate in the project and obtaining permission clearances. The type of the text provider determines the line of negotiations. Permission clearances represent a rather delicate issue for commercial publishers who want guarantees that the corpus will not endanger their market position. Institutions on the other hand are often reluctant to commit themselves to the project because they are not prepared to spend time and money on looking up the data or retrieving information on for instance the source/target language. They are more inclined to collaborate if they are asked for the permission to use data already available on the web.

3.3 Metadata

It is not sufficient to compile a corpus. One should also be able to retrieve relevant information from the corpus. A basic means of exploring a parallel corpus consists in using a KWIC-concordance based on the selection of key words in the source language, and selecting the parallel equivalent text chunks in the target language. However, this type of corpus exploitation is rather crude and insufficient for detailed analysis. In order to improve retrieval criteria, and thus to fine-tune the selection of corpus samples, we will also annotate the corpus with additional metadata at different levels. The DPC metadata list consists of three parts: text-related data, translation-related data and annotation-related data.

The first part includes information on the text: language, author and/or translator, title, publishing information, intended outcome of the text. The text is also characterized according to its type and domain, as well as according to a type of institution, which produced the text (profit vs. non-profit) and according to an intended audience (internal communication, external communication for specialists, or external communication for general public). A list of relevant keywords is provided for the text, as well as information on copyright and on basic statistics (number of tokens, words, sentences and paragraphs).

The second part - translation-related data - indicates the translation direction, and links original and translated texts. It also notes how the text was translated (human translation, translation by a human using translation memory or machine translation corrected by a human) and includes information on alignment tool and alignment quality.

The last part describes the additional annotation of the text. It provides details on tools used for tokenization, PoS tagging, lemmatization and syntactic annotation and the quality of the above annotations. Since we will be using different types of annotation tools, these metadata can help to get a better idea of the quality of the tools used. The metadata on annotation quality is based on the quality control described in section 5 of this article.

The use of the three sets of metadata will improve text retrieval considerably. The metadata suits different type of users, who can select - according to their needs - a more fine-tuned sample set based on the combination of metadata tags. The exploitation of the DPC corpus will be twofold, and in both cases the metadata tags can be used for text retrieval. First of all, the corpus will be made available through a web interface. This interface will consists of a simple parallel KWIC concordance on the one hand and a more advanced query tool that can handle more intricate linguistic patterns. Secondly, the corpus will be made available (through the TST centrale) as

XML-files for researchers who have experience with text data manipulation. In both cases, the metadata can be used as extra filter.

4 Corpus data processing

The data received from providers come in different formats and need to be brought into conformity with the DPC standard. Section 4.1 describes text normalization steps that prepare the incoming texts for further processing: alignment (Section 4.2 and 4.3) and linguistic annotation (Section 4.4).

4.1 Text normalization

Text normalization steps include:

- conversion of texts to txt-format;
- assigning documents a unique standardized name and grouping documents if necessary;
- normalization of character encoding;
- cleaning the data:
- content removal (e.g. tables of contents, tables, indexes, footnotes, headers and footers, images)
- sentence splitting;
- tokenization.

The texts are encoded in conformity with the TEI standards, adapted for aligned sentences. Characters are normalized to the Unicode standard UTF8. Only when certain tools require a different character set (e.g. ISO 8859-1), an intermediate character conversion is used temporarily. Characters not available in the intermediate character set will get an escaped coding format.

4.2 Sentence alignment

In sentence alignment, each sentence of the source language text is connected with the equivalent sentence or sentences of the target language text. The sentences linked by the alignment procedure represent translations of each other in the different languages.

The following alignment links are legitimate in the DPC project:

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

- *1 : 0* (no alignment links for a sentence in a source language)

Zero alignments are created when no translation can be found for a sentence of either the source or the target language, i.e. when the corresponding part of the text is missing in the other language.

Many-to-many alignments are legitimate in two cases: overlapping alignments and crossing alignments.

Overlapping alignments are cases of asymmetric sentence splitting in the two languages. For example, in
Table **2**, a source language text and a target language text both consist of two sentences:

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

Table 2: overlapping alignments

Both sentence pairs in the two languages contain 5 elements A-E and A′-E′ such that A′ is a translation of A, B′ is a translation of B, etc. S_1 and S'_1 cannot be aligned with each other, since translation of element C is absent from S'_1. Similarly, S_2 and S'_2 cannot be aligned with each other, since translation of element C′ is absent from S_1. Therefore, a multiple alignment 2:2 has to be created (S_1, S_2 vs. S'_1, S'_2).

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

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

Sentence alignment is preceded by text normalization (see Section 4.1 above) and paragraph alignment. Paragraph alignment is crucial for the normal functioning of one of the aligners used in the DPC project: the Vanilla aligner (Danielsson and Ridings, 1997). The Vanilla aligner is an implementation of the Gale and Church algorithm (1993), and aligns sentences based on sentence length. The Vanilla aligner requires prior alignment of paragraphs to reduce the search space. Paragraph alignment is performed by the linguists in a manual mode with ParaConc (Barlow, 2002). For short documents such as magazine articles, the whole document is assumed to be one paragraph.

In order to obtain the best possible alignments before manual verification, we opted to combine the results of different alignment tools. The second aligner used in the DPC project is the Microsoft Bilingual sentence aligner (Moore, 2002), which uses word correspondences - generated by a word translation model (IBM Translation Model 1) - to improve the initial alignment based on sentence length. The Microsoft Bilingual sentence aligner creates 1:1 links only.

Sentence alignment procedure involves the following steps:

- reformatting data in the format required by the aligners;

- automatic alignment:
    - with Vanilla aligner;
    - Microsoft Bilingual sentence aligner;
- combination of the output of two aligners;
- manual inspection and correction of the aligners' output;
- encoding of manual corrected alignments in the DPC format.

4.3 Sub-sentential alignment

A small portion of the corpus will be aligned at sub-sentential level. As the intended usage of the sub-sentential links will determine the granularity or level of the linking process, e.g. word-by-word linking to create a lexicon, or linking larger segments (e.g. constituents) for a more structural analysis of the texts, a multi-level annotation scheme as described in Macken (2007) will be used.

4.4 Linguistic annotation

The whole corpus will be lemmatized and enriched with PoS tags. A small portion of the corpus will be further enriched with additional syntactic information (e.g. shallow parses).

To ensure compatibility with the Dutch monolingual corpus developed in the D-COI project (van den Bosch, Schuurman and Vandeghinste, 2006) and the DPC, the PoS tag set and tagger/lemmatizer of the D-COI project will be used.

For English and French candidate tools and PoS tag sets are being evaluated. As the project aims at tagging standards that are compatible for the different languages, the lemmatizers, PoS tag sets and taggers will be selected based on several criteria: compatibility with the D-COI conventions, availability, license terms, and performance. IPR issues on training sets, tag sets and taggers will also be taken into account in the selection procedure.

To increase the quality of the linguistic annotations, part of the processing will be manually verified. If the accuracy of some tools does not meet the standards for some text types, the manually validated texts will be added to the training corpus, and the tools will be regularly retrained to improve accuracy. The manual verification steps will be performed by student-assistants.

5 Quality control

In order to guarantee corpus quality, a quality control system for each step in compiling, annotating and aligning the corpus will be developed. Three forms of quality control are envisaged for the DPC data:

- Manual verification: Traditional manual verification guarantees high quality data. It is performed by qualified linguists with native and near-native language proficiency. Manual validation of each processing step will be guaranteed for minimally 10% of the whole corpus.
- Spot checking: On the basis of an error analysis of the manually verified data, a spot-checking module will be developed.

- Automatic control procedures: Additionally, automatic control procedures are used, such as the automatic comparison of the output from different alignment programs.

A quality label will be used to mark the level of verification. The introduction of a fine-tuned system of quality labels will enable the user to select samples based on quality criteria.

6 Conclusion

The objectives of the Dutch Parallel Corpus project have been described in this paper. The DPC mainly differs from other existing parallel corpora in the following aspects:

1. Quality control: in order to guarantee corpus quality, a considerable part of the DPC corpus is being checked manually at different levels, including sentence splitting, alignment and linguistic annotation. A quality label is used to mark the level of verification.
2. Level of annotation: the DPC corpus is aligned, tagged on part-of-speech level and lemmatized. The annotation and linguistic processing will be produced by state-of-the-art tools.
3. Balanced composition: the DPC will contain texts from a wide range of text types (fiction and non-fiction), and diverse domains. It contains two bidirectional bilingual parts and one trilingual part.
4. Availability: in order to maximize research on parallel corpora, the DPC will be made available to the research community via the Dutch Agency for Human Language Technologies (the TST-centrale).

Acknowledgement

The DPC project is carried out within the STEVIN program, which is funded by the Dutch and Flemish governments.

References

Baker, M. (1995) 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research'. *Target*, 7(2), pp. 223–43.

Baker, M. (1996) Corpus-based translation studies: The challenges that lie ahead, in H. Somers (ed.), Terminology, LSP and Translation, pp. 175–86. Amsterdam, Philadelphia: Benjamins.

Barlow, M. (2002) ParaConc: Concordance software for multilingual parallel corpora, in Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research., Las Palmas, Spain, pp. 20–24.

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

Danielsson, P. and D. Ridings (1997) Practical presentation of a "vanilla" aligner, in Proceedings of the TELRI Workshop on Alignment and Exploitation of Texts, Ljubljana.

Desmet, P. and H. Paulussen (2005) CorpusCALL: opportunities and challenges, in Proceedings of the CALICO congress, Michigan State University, USA.

Deville, G., L. Dumortier and H. Paulussen (2004) Génération de corpus multilingues dans la mise en oeuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère, in G. Purnelle, C. Fairon and A. Dister (eds.), Le poids des mots, Actes des 7es journées internationales d'analyse statistique des données textuelles, pp. 304–312. Louvain-la-Neuve.

Erjavec, T., C. Ignat, B. Pouliquen and R. Steinberger (2005) Massive multilingual corpus compilation; Acquis Communautaire and totale, in Proceedings of the 2nd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LandT'05), Poznan, Poland, pp. 32–36.

Gale, W. A. and K. W. Church (1993) 'A program for aligning sentences in bilingual corpora'. *Computational Linguistics,* 19(1), pp. 75–102.

Gervais, D. (2003) Multitrans ™ System Presentation. Translation Support and Language Management Solutions, in Proceedings of the MT Summit IX, New Orleans, USA.

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

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

Laviosa, S. (2002) Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam/New York: Rodopi.

Lee, D. 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), pp. 37–72.

Macken, L. (2007) Analysis of translational correspondence in view of sub-sentential alignment, in Proceedings of the METIS-II Workshop on New Approaches to Machine Translation, Leuven, Belgium, pp. 97–105.

Moore, R. C. (2002) Fast and accurate sentence alignment of bilingual corpora, in Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, pp. 135–244.

Odijk, J., 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.

Olohan, M. (2004) Introducing Corpora in Translation Studies. London/New York: Routledge.

Simard, M. and E. Macklovitch (2005) Studying the human translation process through the TransSearch log-files, in Proceedings of the AAAI Symposium on Knowledge Collection from volunteer contributors, Stanford, California, USA.

Ungerer, F. and H.-J. Schmidt (1996) An Introduction to Cognitive Linguistics. London and New York: Longman.

van den Bosch, A., I. Schuurman and V. Vandeghinste (2006) Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genua, Italy.