

DPC: compiling a parallel corpus for CALL

Hans Paulussen, Maribel Montero Perez & Piet Desmet

In this talk, we present DPC, the Dutch Parallel Corpus, a 10-million-word translation corpus of Dutch, English and French texts. This talk will cover different aspects of the compilation of DPC and illustrates how a parallel corpus can be used for computer assisted language learning (CALL).

From the very beginning, text corpora and derived lexicons have been used as indispensable resources for research and development in language and speech technology. Over the last two decades, corpora have become the main resource for language research in general, and are now being used as authentic resources for data driven language learning. The introduction of multilingual corpora opens new perspectives for language learning applications in the field of corpusCALL. Parallel corpora offer reliable resource of information on word usage and translation.

DPC, which has been compiled as part of the STEVIN program, is an example of a parallel corpus especially compiled for corpusCALL. Whereas applications in NLP (natural language processing) are less demanding in quality of text selection and annotation, language teaching requires qualitative authentic corpus samples. This implies a well-balanced corpus design and the use of fine-grained metadata. The DPC-project compiled a high-quality state-of-the-art multi-lingual corpus, with Dutch as central language. The resulting product is a parallel corpus which offers added value not yet present or minimally present in existing parallel corpora. The approach followed resulted in a qualitative corpus, which is also very useful for corpus exploitation not limited to automatic data processing. The DPC corpus mainly differs from other existing parallel corpora in the following five aspects: balanced composition, level of annotation, quality control, availability and Dutch kernel.

(1) Balanced composition: Special attention was paid to the selection of text types: DPC is a balanced corpus, containing texts from a wide range of text types (fiction and non-fiction), and diverse domains. (2) Level of annotation: The corpus has been annotated at different levels: DPC is aligned at sentence level, and each word has been grammatically tagged and lemmatized, which facilitates linguistic searches. The annotation and linguistic processing is

Hans Paulussen
Itec KULeuven Campus Kortrijk
E-mail: hans.paulussen@kuleuven-kortrijk.be

produced by state-of-the-art tools. (3) Quality control: a considerable part of the DPC corpus has been checked manually at different levels and spot check procedures have been used. Additionally, automatic control procedures are performed, such as the automatic comparison of output from different alignment programs. (4) Availability: In order to make the corpus accessible for the whole research community, copyright clearance has been obtained for all samples included in the corpus. This was done in close collaboration with the Dutch HLT Agency (TST-centrale). (5) Dutch kernel: DPC is the first parallel corpus focusing on the translation of Dutch as kernel language.

DPC will soon be available in both XML-format (TEI P5) and via a web application. The first format is useful for NLP researchers acquainted with the technical aspects of file structure, where XML is considered a transparent transport format. The second format is useful for researchers who want to select directly authentic samples from the corpus without intervention of technical support. Thanks to an extensive metadata filter, the DPC user can easily restrict the sample selection to a predefined subcorpus selection. Such a filter can help language teachers to select appropriate samples for CALL exercises.

This presentation will show how DPC has been compiled, aligned and annotated, and gives an overview of the corpus exploration facility via web interface. The usefulness of parallel corpora for CALL will be illustrated.