Dutch Parallel Corpus:
MT corpus and translator's aid

Lieve Macken[1], Julia Trushkina[2], Lidia Rura[1]

Language and Translation Technology Team – Ghent, Belgium [1]
K.U. Leuven – Campus Kortrijk, Belgium [2]
Lieve.Macken@hogent.be, Yulia.Trushkina@kuleuven-kortrijk.be, Lidia.Rura@hogent.be

Abstract

This paper reports on the development of the Dutch Parallel Corpus: a high quality sentence-aligned parallel corpus of 10 million words for the language pairs Dutch-English and Dutch-French. The corpus is composed of different text types. All steps of processing the corpus including alignment and linguistic annotation undergo quality control on different levels. Four categories of potential users of the DPC can be distinguished: developers of HLT-applications, linguists conducting more fundamental research, human translators and language learners. This paper focuses on two types of intended users: MT developers and human translators. The paper describes different characteristics of the corpus relevant for such users, concentrating on corpus design, processing of the corpus data and the exploitation of the corpus.

Introduction

This paper highlights the development of the Dutch Parallel Corpus (DPC): a high quality sentence-aligned parallel corpus of 10 million words for the language pairs Dutch-English and Dutch-French.

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 a diverse group of potential users. The following characteristics of the corpus contribute to fulfilling this aim:

Firstly, the DPC is being aligned on a sentence and, partially, on a sub-sentence level to guarantee its usability in such research areas as Machine Translation, Computer-Assisted Language Learning, Computer-Assisted Translation and other multilingual applications. Secondly, the DPC will be enriched with linguistic annotations to broaden the scope of its application. Thirdly, the corpus has a balanced composition of different text types. And last but not least, all data processing steps undergo quality control on different levels, including automatic and semi-automatic verification, as well as consistent manual checks.

At the moment of the paper submission, the DPC project is going through its second stage, concentrating on data alignment. The next step will be linguistic annotation and the development of corpus exploitation tools.

The paper is structured as follows: in the initial part we elaborate on the multifunctionality of the DPC, while the main part of the paper concentrates on the description of the corpus design, processing of the corpus data and the exploitation of the corpus.

Multifunctional purpose

Aligned parallel corpora are an indispensable resource for a wide range of multilingual applications: machine translation, 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 (Olohan, 2004).

Generally speaking, four categories of users can be distinguished: developers of HLT-applications, linguists conducting more fundamental research, human translators and language learners. 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. In this paper we will focus on two types of intended users: MT developers and human translators.

Machine Translation

Aligned parallel corpora are used in MT as training and test material for corpus-based MT systems (SMT or EBMT). 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.

The DPC, which is currently being compiled, contains texts from a wide range of text types (fiction and non-fiction), and diverse domains.

Full text corpora as translator's aid

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. With a bilingual concordancing system, translators can query a large corpus of aligned translated material in order to identify the more appropriate target language equivalents and idiomatic expressions for a difficult source language passage. The sentences matching the search query are retrieved and displayed together with their aligned translation.

It is only recently that the potential of full text corpora as translation aid has been recognized. According to Bowker and Barlow (2004), bilingual concordancing systems in conjunction with aligned parallel corpora can be seen as complementary to translation memories. The decision on which tool to use depends on a number of factors, among others the nature of the job, the text type, the translator's working style and the translator's experience.

According to Simard and Macklovitch, TransSearch processes thousands of queries every day, submitted by professional translators. 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.

Corpus Design

The design principles of the DPC 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[1]. All members of the predefined user group, 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, of whom 17 are computational linguists.

The analysis confirmed a strong need for a parallel corpus with Dutch as a 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, kind and degree of annotations and required metadata. The details are presented below.

Language pairs and translation directions

As stated earlier, the DPC 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.

An important balancing criterion in a parallel corpus is the translation direction. The authors are not aware of any study investigating the impact of the translation direction on MT systems. However, translated texts tend to show certain idiosyncrasies. In translation studies, where among others the differences between translated and non-translated texts are studied as a means to study the translation process, these idiosynchrasies are also known as *translation universals*. Baker (1995) 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.

Therefore, the DPC will be balanced according to the translation direction. Information about the translation direction of the texts included in the corpus, and about how the texts were translated (human translation, computer-assisted translation or machine translation corrected by a human) is documented in the metadata.

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.

Text types

The DPC 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.

The data in the corpus originates from two main sources: commercial publishers and institutions (both profit and non-profit), and this division is 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.

[1] http://www.kuleuven-kortrijk.be/dpc

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 differences in language and translation style used in three different text types are illustrated below:

1. Instructive texts, translated extremely accurately, almost word-for-word are characterized by their dry style devoid of idiomatic expressions or figurative language. The sentence structure remains almost always intact:

En: *Another shortcoming of the 2-tier client-server model became apparent with scale is the amount of resources that are consumed by such applications. Deploying hundreds or thousands of fat clients, as 2-tier clients are often called, increased demands on processing power and capacity of each client workstation.*

Nl: *Een andere beperking van het tweelaagsmodel die bij de schaalvergroting naar voren kwam, is de hoeveelheid resources die wordt gebruikt door grootschalige toepassingen. Het gebruik van honderden of duizenden fat clients, zoals tweelaagsclients vaak worden genoemd, vergt steeds meer van de verwerkingscapaciteit van een clientwerkstation.*

2. The translation of literary texts (both fictional and factual) is marked by deviations from the original as far as sentence structure and lexical units are concerned, because the translator tries not only to convert words into another language, but also to convey the style, the flavour and the ideas of the author. Such texts often contain omissions as well as slight shifts in meaning.

En: *When the Belgian bishops, in 1966, as heads of the Catholic university of Leuven* ***to which the Belgian state pays 90 % of the amount it spends on each of the two State universities,*** *Leuven made an ex cathedra pronouncement about the constitution of the university which completely disregarded informed advice, indignation was so vigorous* ***that their advice was ignored****.*

Nl: *De Belgische bisschoppen, die aan het hoofd staan van de katholieke universiteit, kondigden in 1966 ex cathedra wijzigingen aan in het statuut van de Leuvense universiteit en legden elk weloverwogen advies naast zich neer. De verontwaardiging daarover was toen zo groot dat het* ***hun nadien veel moeite heeft gekost om hun verklaring te verantwoorden****.*

3. Administrative texts are typical examples of idiomatic translations that should convey the message, but sound as naturally as possible and therefore easy to understand. In the following example, the sentence structure is completely changed.

En: *We will send you a new statement whenever there is a change in the amount of your pension or benefit. This means that in any case, you will receive statements for the usual pension and benefit adjustments of January and July, as well as a statement for May, when your annual holiday allowance is paid.*

Nl: *Bij elke wijziging in uw pensioen of uitkering krijgt u een specificatie. Dus in ieder geval in januari en juli vanwege de nieuwe AOW- en Anw-bedragen. En in mei ontvangt u een specificatie vanwege de jaarlijkse uitbetaling van het vakantiegeld.*

A further subtypology is used in the metadata to characterize the source texts.

Besides language pair and translation direction, the DPC will also be balanced proportionally with respect to text type. However, it cannot be ignored that obtaining enough material for certain text types in some translation directions may prove extremely problematic, for instance, newspaper material is hardly ever translated from Dutch into English.

To guarantee the quality of the text samples, most of them come from published materials[2] 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.

Quality control

The development of a high-quality, state-of-the-art multilingual corpus of a reasonable size is a challenging task. The existing parallel corpora are either very large (hence sacrificing quality assurance) or smaller in size.

Three forms of quality control are envisaged for the DPC data:

1. Manual verification
2. Spot-check
3. Automatic control procedures

Manual verification of each processing step will be guaranteed for minimally 10% of the whole corpus. On the basis of an error analysis of the manually verified data, a spot-check module will be developed. Additionally, automatic control procedures are used, such as the automatic comparison of the output from different alignment programs.

A quality label is used to mark the level of verification. With the introduction of a fine-tuned system of quality labels the user can control the selection of corpus samples.

[2] High-quality websites (such as the Belgian portal site www.belgium.be) are also considered to represent published materials.

Copyright clearance

In order to make the corpus available for the whole research community, copyright clearance is being obtained for all samples included in the corpus. The license agreements needed to guarantee availability 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 Dutch Agency for Human Language Technologies (TST-centrale).

Corpus Processing

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

Text Normalization

The first part in compiling a corpus consists of cleaning and normalizing the text and standardizing character encoding. Splitting the text into sentences is also part of this preparatory step.

The incoming data are cleaned: tables of contents, figures, indexes, footnotes, headers and footers are removed.

For some text types, such as technical texts, especially manuals and patient information leaflets, only pdf-documents are available. These pdf-files often represent valuable material of excellent text and translation quality.

Pdf is widely used for final release versions and requires much less memory to store; the pre-release versions in processable formats are seldom kept after the document has been published. Besides, content providers seem to be more willing to allow their material be used if it can be downloaded from their sites rather than investing time and money in looking it up.

However, pdf-documents need to be converted before they can be normalized. Conversion results vary depending on the layout complexity, the document structure and the program used to convert the file into pdf. The team is therefore trying out different converters on a variety of documents in order to obtain texts with a minimum of noise.

Nevertheless, it does not seem to be possible to fully automate the process: manual verification and correction of e.g. false paragraph breaks are inevitable.

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 get an escaped coding format.

Alignment

For the alignment, (semi-)automatic procedures are being used. As the alignment is the main characteristic of the parallel corpus, the result of the sentence alignment process is verified in detail for a considerable part of the corpus. A small portion of the corpus will also be aligned at sub-sentential level.

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 through 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)

Below are some examples, illustrating possible alignment combinations, encountered in the corpus:

1:1

Nl: *De regent van Vlaanderen hielp hem daarbij door de twee meisjes aan hem uit te leveren, iets waar hij later spijt van zou hebben gehad.*

En: *He was aided and abetted by the regent of Flanders, who delivered the two little girls into his hands.*

1: many (2)

En: *Later the latter was said to have regretted this step, and to have begged the monks of the abbey where he lay dying to drag him through the streets by a rope tied round his neck to die like a dog, as he had lived like one.*

Nl: *In de abdij waar hij op sterven lag, zou hij de monniken hebben gevraagd om hem met een touw rond zijn nek door de stad te slepen. Hij had immers gehandeld als een hond en wou nu ook als een hond sterven.*

many (2): many (2)

En: *Two small daughters remained -- a golden opportunity for the wily old fox of France. It looked as if he were going to achieve his aim without violence.*

Nl: *Ze lieten twee dochtertjes achter. De sluwe oude vos van Frankrijk zag daarin een gouden kans om zonder geweld zijn doel te bereiken.*

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 1, a source language text and a target language text both consist of two sentences:

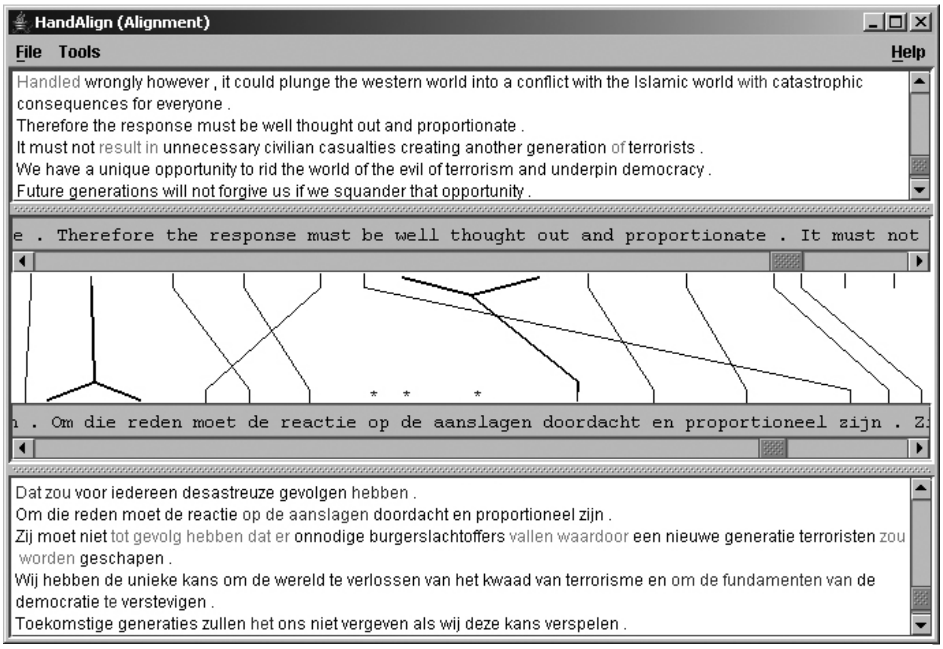
Source language text	*Target language text*
S_1: A, B, C	S'_1: A', B'
S_2: D, E	S'_2: C', D', E'

Table 1: 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.



Handled wrongly however , it could plunge the western world into a conflict with the Islamic world with catastrophic consequences for everyone .
Therefore the response must be well thought out and proportionate .
It must not result in unnecessary civilian casualties creating another generation of terrorists .
We have a unique opportunity to rid the world of the evil of terrorism and underpin democracy .
Future generations will not forgive us if we squander that opportunity .

e . Therefore the response must be well thought out and proportionate . It must not

n . Om die reden moet de reactie op de aanslagen doordacht en proportioneel zijn . Z

Dat zou voor iedereen desastreuze gevolgen hebben .
Om die reden moet de reactie op de aanslagen doordacht en proportioneel zijn .
Zij moet niet tot gevolg hebben dat er onnodige burgerslachtoffers vallen waardoor een nieuwe generatie terroristen zou worden geschapen .
Wij hebben de unieke kans om de wereld te verlossen van het kwaad van terrorisme en om de fundamenten van de democratie te verstevigen .
Toekomstige generaties zullen het ons niet vergeven als wij deze kans verspelen .

Figure 1: Sub-sentential alignment

In his search for the best method for sentence alignment tools, Rosen (2005) confirmed that quality of alignment depends to a large extent on properties of the input and alignment methods differ in their sensitivity to such properties. He concluded that word-correspondence methods are far better on noisy texts, where sentence-length methods give mixed results. In order to obtain the best possible alignments before manual verification, we opted to combine the results of different alignment tools.

The first alignment tool used in the DPC project is the Vanilla aligner (Danielsson and Ridings, 1997). The Vanilla aligner is an implementation of the Church and Gale (1993) algorithm, 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 with the ParaConc tool[3]. For short documents such as magazine articles, the whole document is assumed to be one paragraph.

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.

We are currently investigating whether adding a third alignment tool (Melamed, 1997) improves the precision of the sentence alignments.

Sub-sentential Alignment

A small portion of the corpus will be aligned at sub-sentential level. Reference corpora where sub-sentential translational correspondences are indicated manually are labour-intensive to create, and hence less widespread. Such manually created reference alignments – also called Gold Standards – have been used as an objective means for testing statistical word alignment systems (Melamed, 1998; Och and Ney, 2003).

[3] http://www.athel.com/para.html

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 will be used. In order to create a Gold Standard, an annotation style guide was created for the Dutch-English language pair. The annotation style guide was to a large extent based on the annotation guidelines of other word alignment projects (Melamed, 2001; Merkel, 1999; Véronis, 1998). An annotation style guide for the Dutch-French language pair will be developed.

To facilitate the annotation process, a graphical annotation tool, HandAlign (Daumé III and Marcu, 2005), will be used. The HandAlign annotation tool was originally developed for aligning articles and their summaries, but the tool offers enough flexibility to use it for other alignment purposes.

The annotator works in a graphical environment that consists of three panels (see Figure 1):

- The top text area contains the source text.
- The bottom text area contains the target text.
- The alignment area (in the middle) is where the source and target units can be selected and linked graphically.

As can be seen in Figure 1, the annotators can link different units (e.g. word, word groups, paraphrased sections, punctuation). The corresponding units are not necessarily contiguous.

As translations are characterized by both correspondences and changes, two different types of alignments were introduced: regular links were used to connect straightforward correspondences; fuzzy links for translation-specific shifts of various kinds (syntactic shifts, e.g. active-passive transformations, paraphrases, etc.). Null links were used for source text units that had not been translated or target text units that had been added. Null links are represented by means of an asterisk in the graphical annotation tool.

Linguistic annotation

It is generally accepted that corpora become more useful when the texts are enriched with additional linguistic annotations. The whole DPC will be tokenized, lemmatized and PoS-tagged. A small portion of the corpus will be further enriched with additional syntactic information (e.g. shallow parses).

As the whole corpus will be lemmatized, the human translator will be able to formulate his/her queries in a more intuitive way. The PoS-information and shallow syntactic information will be useful to study systematic structural changes that occur during translation.

We are currently investigating what state-of-the-art tools will be used for the annotation. To ensure compatibility with the Dutch monolingual corpus developed within the D-COI project (van den Bosch et al., 2006) and the DPC, the PoS tag set and combined tagger/lemmatizer of the D-

English word form	Penn Treebank PoS code	Dutch word form	D-COI PoS code
She	PRP	Ze	VNW(pers,pron,stan,red,3,ev,fem)
compared	VBD	vergeleek	WW(pv,verl,ev)
the	DT	het	(bep,stan,evon)
appearance	NN	uitzicht	N(soort,ev,basis,onz,stan)
of	IN	van	VZ(init)
Neferiti	NNP	Nefertiti’s	N(eigen,ev,basis,gen)
‘s	POS		
mummy	NN	mummie	N(soort,ev,basis,zijd,stan)
with	IN	met	VZ(init)
the	DT	de	LID(bep,stan,rest)
royal	JJ	koninklijke	ADJ(prenom,basis,met-e,stan)
Egyptian	JJ	Egyptische	ADJ(prenom,basis,met-e,stan)
fashion	NN	mode	N(soort,ev,basis,zijd,stan)
of	IN	uit	VZ(fin)
that	DT	die	VNW(aanw,det,stan,prenom,zonder,rest)
time	NN	tijd	N(soort,ev,basis,zijd,stan)
and	CC	en	VG(neven)
believes	VBZ	gelooft	WW(pv,tgw,met-t)
that	IN	dat	VG(onder)
the	DT	de	LID(bep,stan,rest)
mummy	NN	mummie	mummie N(soort,ev,basis,zijd,stan)
can	MD	kan	WW(pv,tgw,ev)
be	VB	worden	WW(inf,vrij,zonder)
identified	VBN	geïdentificeerd	WW(vd,vrij,zonder)
as	IN	als	VG(onder)
queen	NN	koningin	N(soort,ev,basis,zijd,stan)
Nefertiti	NNP	Nefertiti	N(eigen,ev,basis,onz,stan)
.	.	.	LET()

Tabel 2: English and Dutch PoS codes

COI project will be used. In the D-COI project, a 50-million-word pilot corpus of contemporary written Dutch was compiled.

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.

It is not possible, nor in our view desirable to use identical PoS tag sets across the different languages. Instead, rather than adapting the internationally accepted standards, it is better to define mapping tables so that the PoS codes of the different languages can be easily projected onto each other.
By way of illustration, the PoS codes for the following English and Dutch sentence are displayed in Table 2:

En: *She compared the appearance of Nefertiti's mummy with the royal Egyptian fashion of that time and believes that the mummy can be identified as queen Nefertiti.*
Nl: *Ze vergeleek het uitzicht van Nefertiti's mummie met de koninklijke Egyptische mode uit die tijd en gelooft dat de mummie kan worden geïdentificeerd als koningin Nefertiti.*

In the example, word-by-word correspondences can be indicated for all words, making it a perfect example to illustrate the differences in tokenization and PoS tagging. In the example a different treatment of the possessive marker ***'s*** in English and Dutch can be observed. According to the Penn Treebank conventions, the possessive marker '*s* is split off during tokenization, and a separate tag (*POS*) is assigned.
According to the D-COI conventions the possessive marker is not stripped off during tokenization, and the possessiveness of the proper noun *Nefertiti* is coded in the last attribute of *N(eigen,ev,basis,gen).* This last attribute contains case information (*gen* stands for "genitive", *stan* for "standard").

In general, the Dutch D-COI PoS tag set is more fine-grained. A mapping table will be developed to project the Penn Treebank codes to the D-COI codes and vice versa. For French a similar procedure will be followed.

Corpus exploitation

The DPC will be made available as a full text resource and through a web interface. The DPC web interface should be seen as a multilingual concordancer, with which the user can query the database at different levels. This interface will consist of a simple parallel KWIC concordance on the one hand, and a more advanced query tool that can handle more intricate linguistic patterns.

Examples of possible queries are described by Simard and Macklovitch (2005) and include the following types: single words, continuous groups of words, and discontinuous groups of words. The queries can be language-independent, as well as language-specific and bi- or multilingual. All queries may be enriched with information on linguistic annotation, such as lemmas, parts of speech or syntactic functions of words. A web interface with such a query function is a helpful tool for human translators.

The second form in which the corpus will be made available, i.e. a full text resource, is needed for inductive language learning tools, e.g. Machine Translation. For each text pair two monolingual XML-files and one alignment file will be released.

Conclusion

Most corpus projects aim at collecting huge quantities of data, especially when quantitative and statistical results are involved. However, it is impossible to cover language as a whole: a corpus is always a snapshot of language usage. In this project we follow a complementary approach, focusing on quality rather than quantity.

Although the limited size of the corpus (10 million words) could be seen as a drawback, its certain strong points may prove invaluable for many researchers. First, the quality of the corpus texts is controlled on all levels, including corpus normalization, alignment and annotation. Second, restricting the corpus size to ten million words allows for a balanced composition of the corpus and for text type diversity. At the same time, the quality of the texts included in the corpus has been verified manually. And finally, the corpus minimizes a number of indirect translations and provides information on translation direction for most texts.

These characteristics make the corpus useful both for machine translation developers, e.g. for testing the accuracy of an MT system on different text types, and for human translators, providing a necessary base for translation aid.

Acknowledgements

The DPC project is carried out within the STEVIN program, which is funded by the Dutch and Flemish Governments. In more personalised terms, DPC is also Piet Desmet, Hans Paulussen, and Antoine Besnehard (KU Leuven – Campus Kortrijk) and Willy Vandeweghe (School of Translation Studies, Hogeschool Ghent).

References

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. Target, 7(2), 223-243.

Bowker, L., & Barlow, M. (2004). Bilingual concordancers and translation memories: A comparative evaluation. In Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training (Coling 2004) (pp. 52-61). Geneva, Switzerland.

Carl, M., & Way, A. (2003). Recent Advances in Example-Based Machine Translation (Vol. 21). Dordrecht: Kluwer Academic Publishers.

Danielsson, P., & Ridings, D. (1997). Practical presentation of a "vanilla" aligner. In Proceedings of the

TELRI Workshop on Alignment and Exploitation of Texts. Ljubljana.

Daumé III, H., & Marcu, D. (2005). Induction of word and phrase alignments for automatic document summarization. Computational Linguistics, 31(4), 505-530.

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

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger , R. (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 (L&T'05) (pp. 32-36). Poznan, Poland.

Gale, W.A., & Church, K.W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1), 75-102.

Gervais, D. (2003). Multitrans ^{TM} 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), 5-38.

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

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 (pp. 97-105). Leuven, Belgium.

Melamed, D.I. (1997). A Portable Algorithm for Mapping Bitext Correspondence. In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL) (pp. 305-312). Madrid, Spain.

Melamed, D.I. (1998). Empirical methods for MT lexicon development. In Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA '98). Langhorne PA, USA.

Melamed, D.I. (2001). Annotation style guide for the Blinker Project. In D. I. Melamed (Ed.), *Empirical methods for exploiting parallel texts* (pp. 169-182). Cambridge, Massachusetts: MIT Press.

Merkel, M. (1999). Annotation Style Guide for the PLUG Link Annotator. Retrieved 08-06-2005, from http://www.ida.liu.se/~magme/publications/pluglinkannot.pdf

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 (pp. 135-244). Tiburon, California.

Och, F.J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), 19-51.

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

Rosen, A. (2005). In search of the best method for sentence alignment in parallel texts. In Proceedings of the Third International Seminar on Computer Treatment of Slavic and East European Languages (SLOVKO 2005). Bratislava, Slovakia.

Simard, M., & Macklovitch, E. (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.

van den Bosch, A., Schuurman, I., & Vandeghinste, V. (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.

Véronis, J. (1998). Arcade. Tagging guidelines for word alignment. Version 1.0. Retrieved 15-11-2005, from http://www.up.univ-mrs.fr/veronis/arcade/arcade1/2nd/word/guide/index.html