Annotating the Dutch Parallel Corpus

Hans Paulussen*
ITEC, K.U.Leuven Campus Kortrijk

Lieve Macken†
LT^3, University College of Ghent and Ghent University

Abstract

The Dutch Parallel Corpus (DPC) is a translation corpus containing Dutch, English and French text samples aligned at sentence level. Next to sentence alignment, the corpus has also been grammatically annotated, thus improving exploitation for different domains, including natural language processing, translation research or CALL (computer-assisted language learning). In this paper, we describe the compilation of DPC and the alignment procedures used. This is followed by a description of the annotation task for the three languages, which required different tools and different tag sets. Finally the impact of different grammatical annotations on multilingual corpus exploitation is discussed.

1 Introduction

Over the last decade it has become clear that aligned parallel corpora are indispensable resources for a wide range of multilingual applications. These include different domains, such as machine translation (especially corpus-based MT such as statistical and example-based MT), computer-assisted translation tools, cross-lingual information extraction, multilingual terminology extraction, and computer-assisted language learning.

For some time, high-quality parallel corpora with Dutch as the central language did not exist or were not readily accessible for the research community. This was mainly due to copyright restrictions.

The output of the DPC project is a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French (Paulussen *et al.* [16], Macken *et al.* [10]). The corpus is a multilingual translation

*email: fistname.lastname@kuleuven-kortrijk.be
†email: fistname.lastname@hogent.be

corpus that not only is aligned at sentence level, but that has also been annotated grammatically and lemmatised for all three languages involved. The combination of aligned sentences and an enriched grammatical annotation layer makes DPC a very useful instrument for multilingual corpus exploitation.

This article starts with a general description of the Dutch Parallel Corpus, pointing out in which way DPC differs from similar parallel corpora. Then the alignment procedures and the annotation procedures used during the compilation of the corpus are described. This is followed by a discussion on the consequences and the impact of the use of different grammatical annotation sets with regard to parallel corpus exploitation.

2 A well-balanced parallel corpus

An important drawback of many parallel corpora is their lack of text balance. For example, the MLCC parallel corpus[1] only covers a selection of the Debates of the European Parliament. Similarly, the EU ACQUIS parallel corpus is solely devoted to European legal texts (Erjavec *et al.* [7]).

One of the main tasks of the DPC project consisted in compiling a well-balanced translation corpus. Therefore, special attention was paid to select a representative sample for each text type. The texts were selected from different domains in compliance with the requirements of the user group. The 10,000,000 word corpus covers translations of the following five text types: literature, journalistic texts, instructive texts, administrative texts and external communication. The texts were selected from different types of text providers including publishing houses, press, government, commercial companies and content brokers (Rura *et al.* [17]).

In order to guarantee the quality of the corpus, a number of validation stages were incorporated during the compilation process, and this for every step of the compilation task: corpus design, text sample selection, cleaning and structuring the data, alignment and annotation of the corpus. For the annotation step, special attention was paid to carefully comply with the annotation protocols proposed by the researchers of the D-Coi project, who compiled a 50-million-word pilot corpus of contemporary Dutch[2].

One of the main tasks of the DPC project consisted in solving all copyright issues (De Clercq and Montero Perez [5]). Thanks to monitoring this delicate task, the corpus is now freely available for the whole research community. The Dutch HLT-agency[3] is in charge of the distribution of the corpus.

[1] URL: `http://www.elda.org/catalogue/en/text/W0023.html`

[2] The D-Coi project also uses the annotation procedures developed for compiling CGN (Corpus Gesproken Nederlands), the Dutch Spoken Corpus.

[3] URL: `http://www.inl.nl/nl/tst-centrale`

3 Sentence alignment

In the DPC project, the alignment was carried out with three different aligners. The basic aligner was the vanilla aligner (Danielsson and Ridings [4]), which is an implementation of the sentence-length-based statistical approach of Gale and Church ([8]). The vanilla aligner has some practical limitations, since it expects texts to have the same number of paragraphs in source and target texts, so that some preprocessing was required.

The second aligner is Melamed's GMA tool (Geometric Mapping and Alignment) ([12]), an implementation of the *Smooth Injective Map Recognizer* (SIMR) algorithm, which is based on word correspondences and sentence length, and relies on finding cognates (tokens with the same meaning and similar spelling) in parallel texts to suggest word correspondences. Optionally translation lexicons can also be used. In the DPC project, we used the NL-Translex translation lexicons ([9]) as an additional source for establishing word correspondences.

The third aligner is the Microsoft Bilingual Aligner developed by Moore ([13]). This aligner uses a three-step hybrid approach to sentence alignment. The aligner uses sentence length and lexical correspondences, both of which are derived automatically from the source and target texts. In a first step, an initial set of high accuracy alignments is established using a sentence-length-based approach. In the second step, this initial set of alignments serves as the basis for training a statistical word alignment model ([2]). Finally, the corpus is realigned, augmenting the initial set of alignments with sentences aligned on the basis of the word alignments. The aligner outputs only 1:1 links and disregards all other alignment types.

During the DPC project we have tested the three aligners, and came to the conclusion that by combining the output of different aligners the amount of manual work necessary to achieve near 100% accuracy can be reduced significantly (Trushkina *et al.* [18]).

4 Linguistic annotation

To improve exploitation facilities of a parallel corpus, an additional linguistic annotation layer was added to the sentence aligned DPC corpus, consisting in grammatical annotation (adding Part-of-Speech tags) and lemmatization[4].

Because of the available tools and the existing PoS tag sets, the job was carried out differently for each of the three languages. Although aiming at complying with annotation standards, such as specified by EAGLES ([6]), we also had to comply with *de facto* standards.

For English, grammatical annotation and lemmatization was performed by the combined memory-based PoS tagger/lemmatizer, which is part of the MBSP tools ([3]). The English memory-based tagger was trained on data from the Wall Street

[4]Note that the sentence alignment task and the annotation task were carried out concurrently. Only a the end of the project, the resulting files were fused into one output format.

Journal corpus in the Penn Treebank ([11]), and uses the Penn Treebank tag set, which contains 45 distinct tags.

For Dutch, the D-Coi tagger was used ([19]). This tagger, which uses the CGN PoS tag set ([20]), is an ensemble tagger that combines the output of different machine learning algorithms. The CGN PoS tag set ([20]) is characterized by a high level of granularity. Apart from the word class, the CGN tag set codes a wide range of morpho-syntactic features as attributes to the word class. In total, 316 distinct full tags are discerned.

```
<seg type="sent" n="seg.p1.s4">
  <seg type="original">
    De fles wordt in geopende toestand met een staalkabel
    in zee neergelaten.
  </seg>
  <s n="p1.s4">
    <w ana="LID(bep,stan,rest)" lemma="de">De</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="fles">fles</w>
    <w ana="WW(pv,tgw,met-t)" lemma="worden">wordt</w>
    <w ana="VZ(init)" lemma="in">in</w>
    <w ana="WW(vd,prenom,met-e)" lemma="openen">geopende</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="toestand">toestand</w>
    <w ana="VZ(init)" lemma="met">met</w>
    <w ana="LID(onbep,stan,agr)" lemma="een">een</w>
    <w ana="N(soort,ev,basis,zijd,stan)"
         lemma="staalkabel">staalkabel</w>
    <w ana="VZ(init)" lemma="in">in</w>
    <w ana="N(soort,ev,basis,zijd,stan)" lemma="zee">zee</w>
    <w ana="WW(vd,vrij,zonder)" lemma="neerlaten">neergelaten</w>
    <w ana="LET()" lemma=".">.</w>
  </s>
</seg>
```

Figure 1: DPC sample annotation Dutch: dpc-bmm-001099-nl

For French, we used a modified version of TreeTagger. This approach was motivated by the fact that the basic PoS tag set of TreeTagger is rather limited for French. Instead of using the basic small PoS tag set, covering only 33 labels[5], we opted for the richer GRACE tag set ([15]). A parameter file based on GRACE was provided by the LIMSI research team, who had created a corpus using the GRACE tag set ([1]). Although this parameter file contained the grammatical information, it did not contain any lemmatized data. In order to solve this problem we opted for a two step annotation cycle. In a first run, the basic parameter file was used and lemmata were added, together with tagging probabilities. The lemmata were then updated and detailed morphosyntactic information was added using a separate tool (FLEMM, [14]). In a second run, the LIMSI parameter file was used, based on the GRACE tag set, covering 312 distinctive tags. Finally, the output of both

[5]The French TreeTagger tagset can be found at URL:
http://www.ims.uni-stuttgart.de/šchmid/french-tagset.html

runs were compared and put together. Only the GRACE tags were retained. The original TreeTagger tags were only used for comparison. The comparison of both runs was also used for spot checking possible errors where the two outputs differ in one way or another.

At the final stage of the DPC project, the output of both main tasks (i.e. sentence alignment and grammatical annotation) were fused together into one XML file, as illustrated in Figure 1, which will be explained in the following section.

5 Discussion

In this section, we discuss the implications of the three PoS tagging formats when used in exploitation of parallel corpora. The tagging codes used are illustrated in example (1), where the token-PoS pairs are given of the Dutch sample shown in Figure 1, together with the token-PoS pairs for the corresponding French and English sentences. A more legible format of the three sample sentences is shown in example (2).

(1) a. De/LID(bep,stan,rest) fles/N(soort,ev,basis,zijd,stan)
wordt/WW(pv,tgw,met-t) in/VZ(init)
geopende/WW(vd,prenom,met-e)
toestand/N(soort,ev,basis,zijd,stan)
met/VZ(init) een/LID(onbep,stan,agr)
staalkabel/N(soort,ev,basis,zijd,stan) in/VZ(init)
zee/N(soort,ev,basis,zijd,stan)
neergelaten/WW(vd,vrij,zonder) ./LET()

b. La/Da-fs-d bouteille/Ncfs est/Vmip3s immergée/Afpfs
ouverte/Vmps-sf à/Sp ses/Ds3fps deux/Ak-fp extrémités/Ncfp
et/Cc fixé/Vmps-sm le/Da-ms-d long/Ncms d'/Sp
un/Da-ms-i câble/Ncms en/Sp acier/Ncms ./F

c. The/DT bottle/NN is/VBZ lowered/VBN into/IN the/DT
sea/NN on/IN a/DT steel/NN cable/NN ,/, open/JJ ./.

(2) a. De fles *wordt* in geopende toestand met een staalkabel in zee *neergelaten*.

b. La bouteille *est immergée* ouverte à ses deux extrémités et fixé le long d'un câble en acier.

c. The bottle *is lowered* into the sea on a steel cable, open.

First we give a brief explanation of the XML-format used for the sample sentences. DPC has been packed in TEI P5 format, in order to have a well-formed and validated format[6]. In the XML-formated samples, a sentence is represented twice.

[6]XML is only considered a wrapping format, in order to distribute the data easily; exploitation of the data can be carried out in whatever format the programmer prefers.

First, the original cleaned sentence is shown in a <seg> element of type original. Then the sentence is shown in an <s> element, including the tokenised format, whereby each word element (<w>) contains two attributes: the PoS tag (ana) and the lemmatized form (lemma). The original sentence can be helpful to reconstruct the original layout of a sentence, when token segmentation is ambiguous. It is also useful for skimming the XML-file quickly, since a flow of horizontally ordered words is easier to read than a list of words displayed in vertical format.

The sample sentences show some general differences, which can be interesting for translation studies. For example, the French version is more verbose than the Dutch and English sentences[7]. In Dutch, typically the auxiliary is placed at the beginning of the sentence whereas the main verb (*neergelaten*) is placed at the end. In French and English, the verb group — shown in italics — remains clustered at the beginning of the sentence. Another example is the word *staalkabel* which is translated as a noun group in English (*steel cable*) and a prepositional construction in French (*câble en acier*). Unfortunately, the underlying categorial labels cannot transparently be matched, which requires some processing. As shown in the three figures, each language uses a different set of PoS tags. In general, this is not such a big problem, but it can be annoying, if you really want to compare grammatical patterns.

When we take a closer look at the PoS tags in the three samples, one can immediately see that categorial and subcategorial information is intertwined in the English tags. In Dutch and French, on the other hand, a more systematic structure is used, which is related to the fact that the two latter languages are morphologically rich when compared to English, but probably also to the fact that English has been the first language for which a tagging scheme has been established.

The Dutch PoS tags have a clear pattern, whereby the tag always starts with the category label indicated in capital letters. The subcategorial features are summed up between brackets. For French, a similar approach is used: the first letter of each PoS tag indicates the grammatical category; all following letters refer to the subcategorial features. Because of the systematic structure of the Dutch and French PoS tags, it is not so difficult to select sentences in both languages based on a categorial filter.

A possible solution to handle the three different tag sets is shown in Table 1. The left column lists the generally accepted 10 basis categories, followed by punctuation labels and miscellaneous labels. The other columns contain the category labels for the three languages. In the case of English, we show the full label pattern, whereas for Dutch and French, only the categorial information is shown. The only exceptions for French are the preposition (Sp) — which is considered an adposition — and the numerals, which are considered a subcategorial feature.

The Dutch and French mapping are closest to the EAGLES proposals, whereas the English tags have a completely different system. The table may be a bit confusing, in the sense that only the English labels show a mixture of categorial and

[7] Note that in French, the bottle seems to be open at both sides: *ouverte à ses deux extrémités*

Basic categories	NL	EN	FR
Noun	N	NNS?, NNPS?	N
Verb	WW	VB[DGNPZ]? MD	V
Adjective	ADJ	JJ[RS]?	A
Adverb	BW	RB[RS]? WRB EX	R
Determiner	LID	DT, WDT	D
Numeral	TW	CD	[NAPD]k Ao
Pronoun	VNW	PRP$?, WP$?	P
Preposition	VZ	IN, TO	Sp
Conjunction	VG	CC, IN	C
Interjection	TSW	UH	I
Punctuation	LET		F
Miscellaneous	SPEC	SYM EX PDT, POS RP, FW LS	X ?

Table 1: Mapping PoS in Dutch, English and French

subcategorial information. Some striking examples for English are the following:

All verb tags, except modal verbs (`MD`) start with `VB`. `WH`-words have a special status, since they are listed as adverb (`WRB`), determiner (`WDT`) and pronoun (`WP$?`). Strictly speaking, these cases are only formal variations of the same category, which only lead to cumbersome selections. Suppose, for example, that you want to check whether a selected sentence starts with a determiner in the three languages, you'll have to specify four different labels (`LID`, `DT` or `WDT` and `D`), which is a bit awkward.

There are two cases which are problematic. The `IN` label is ambiguously considered a *preposition or subordinating conjunction*. The second case is `TO` which is normally used to indicate *to* when linked to an infinitive verb, but which can also be used as an ordinary preposition. Moreover, `TO` as an infinitive indicator has no equivalent in French or Dutch. Depending on the kind of queries you are analyzing, further analysis of the selected material may be necessary.

6 Conclusion

This article presented DPC, a new parallel corpus for Dutch, English and French. DPC is a sentence aligned corpus with an additional linguistic annotation layer, encoding grammatical tags and lemmata. The extra layer makes the corpus more

suitable for fine-grained grammatical selections, at least for monolingual selections. In the case of multilingual selection, things can be rather complicated: the PoS tags differ for the three languages, even when in most cases, these tags are *de facto* standards for monolingual research. Also, the grammatical annotation schemes differ for the three languages.

A partial mapping of the PoS tags is possible, but this is mainly limited to the category level: combination of category and subcategory is not always possible. The labeling system behind the PoS tags for Dutch and French are better structured than the English PoS tags, which is related to the fact that English was the first language to be tagged, but also because English is morphologically poor, in comparison to Dutch or French.

Selection of text samples based on PoS tags are best carried out at category level. Multilingual selections may require mapping of PoS tags. Some tags are ambiguous and my need further analysis.

References

[1] Allauzen, A. and Bonneau-Maynard, H. (2008). Training and Evaluation of POS Taggers on the French MULTITAG Corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, pp. 28-30.

[2] Brown, P. F., Della Pietra, V. J., Della Pietra, S. A. and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics* 19(2), 263–311.

[3] Daelemans, W. and van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

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

[5] De Clercq, O. and Montero Perez, M. (2010). Data Collection and IPR in Multilingual Parallel Corpora: Dutch Parallel Corpus. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010)*, Valletta, Malta.

[6] [EAGLES] Expert Advisory Group on Language Engineering Standards (1996). *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-MAC/R. Version of March, 1996. (URL: `http://www.ilc.cnr.it/EAGLES/home.html`)

[7] Erjavec, T., Ignat, C., Pouliquen, B. and 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, Poznan, Poland.

[8] Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. In *Computational Linguistics* 19(1), 75-102.

[9] Goetschalckx, J., Cucchiarini, C. and Van Hoorde, J. (2001). *Machine Translation for Dutch: the NL-Translex Project*. Brussels/Den Haag, January 2001; 16pp.

[10] Macken, L., Declercq, O., and Paulussen, H. (forthcoming). Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. In *META*, 56(1).

[11] Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*. 19(2): 313-330.

[12] 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)*, Madrid, Spain, pp. 305-312.

[13] Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the fifth Conference of the Association for Machine Translation in the Americas (AMTA), Machine Translation: from research to real users*, Tiburon, California, pp. 135-244.

[14] Namer, F. (2000). FLEMM : Un analyseur flexionnel du français à base de règles. In *TAL, Traitement automatique des langues*. 41(2): 523-547.

[15] Paroubek, P. (2000). Language resources as by-product of evaluation: the multitag example. In *Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, pp. 151-154.

[16] Paulussen, H., Macken, L., Trushkina, J., Desmet, P. and Vandeweghe, W. (2006). Dutch Parallel Corpus: a multifunctional and multilingual corpus. In *Cahiers de l'Institut de Linguistique de Louvain, CILL*, Louvain-La-Neuve, 32.1-4 (2006), 269-285.

[17] Rura, L., Vandeweghe, W. and Montero Perez, M. (2008). Designing a parallel corpus as a multifunctional translator's aid. In *Proceedings of XVIII FIT World Congress*, August 2008, Shangai, pp. 4-7.

[18] Trushkina, J., Macken, L. and Paulussen, H. (2008). Sentence Alignment in DPC: Maximizing Precision, Minimizing Human Effort. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco.

[19] van den Bosch, A., Schuurman, I. and Vandeghinste, V. (2006). Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, Genoa, Italy, 2006.

[20] Van Eynde, F., Zavrel, J. and Daelemans, W. (2000) Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, pp. 1427-1434.