Chapter 11
Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French

Hans Paulussen, Lieve Macken, Willy Vandeweghe, and Piet Desmet

11.1 Introduction

Parallel corpora are a valuable resource for researchers across a wide range of disciplines, i.e. machine translation, computer-assisted translation, terminology extraction, computer-assisted language learning, contrastive linguistics and translation studies. Since the development of a high-quality parallel corpus is a time-consuming and costly process, the DPC project aimed at the creation of a multifunctional resource that satisfies the needs of this diverse group of disciplines.

The resulting corpus—the Dutch Parallel Corpus (DPC)—is a ten-million-word, sentence-aligned, linguistically enriched parallel corpus for the language pairs Dutch-English and Dutch-French. As the DPC is bidirectional, the corpus can also be used as comparable corpus to study the differences between translated versus non-translated language. A small part of the corpus is trilingual. The DPC distinguishes itself from other parallel corpora by having a balanced composition (both in terms of text types and translation directions), by its availability to the wide research community thanks to its copyright clearance and by focusing on quality rather than quantity.

To guarantee the quality of the text samples, most of them were taken from published materials or from companies or institutions working with a professional translation division. Care was taken to differentiate kinds of data providers, among them providers from publishing houses, press, government, corporate enterprises, European institutions, etc. To guarantee the quality during data processing, 10 %

H. Paulussen (✉) · P. Desmet
ITEC-IBBT KULeuven Kulak, Kortrijk, Belgium
e-mail: hans.paulussen@kuleuven-kulak.be; piet.desmet@kuleuven-kulak.be

L. Macken · W. Vandeweghe
Language and Translation Technology Team (LT^3), University College Ghent and Ghent University, Gent, Belgium
e-mail: lieve.macken@hogent.be; willy.vandeweghe@hogent.be

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_11, © The Author(s) 2013

of the corpus has been manually verified at different levels, including sentence splitting, alignment and linguistic annotation. On the basis of these manually verified data, spot-checking and automatic control procedures were developed to verify the rest of the corpus. Each sample in the corpus has an accompanying metadata file. The metadata will enable the corpus users to select the texts that fulfil their specific requirements. The entire corpus is released as full texts in XML format and is also available via a web interface, which supports basic and complex search queries and presents the results as (parallel) concordances.

The remainder of this paper is structured as follows. Section 11.2 focuses on corpus design and data acquisition, while Sect. 11.3 elaborates on the different corpus processing stages. Section 11.4 contains the description of the two DPC exploitation formats along with the first exploitation results of the corpus in different research domains. Section 11.5 ends with some concluding remarks.

11.2 Corpus Design and Data Acquisition

The design principles of DPC were based on research into standards for other parallel corpus projects and a user requirements study. Three objectives were of paramount importance: balance (Sect. 11.2.1), quality of the text samples and IPR clearance (Sect. 11.2.2).

11.2.1 Balanced Corpus Design

The Dutch Parallel Corpus consists of two language pairs (Dutch-English and Dutch-French), has four translation directions (Dutch into English, English into Dutch, Dutch into French and French into Dutch) and five text types (administrative texts, instructive texts, literature, journalistic texts and texts for external communication). The DPC is balanced both in terms of text types and translation directions.

In order to enhance the navigability of the corpus, a subdivision was imposed on the five text types resulting in the creation of a finer tree-like structure within each type. This subdivision has no implications for the balancing of the corpus. The introduction of subtypes 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 select documents and search the corpus. A division could also be made between two main data sources: commercial publishers versus institutions and companies (cf. Table 11.1). For a detailed description of the DPC corpus design and text typology, we refer to [17, 24].

All information on translation direction and text types has been stored in the metadata files, complemented with other translation- and text-related information such as the intended audience, text provider, etc.

Table 11.1 DPC text types and subtypes according to data source

Source	Text type	Subtype
Institutions / Companies	Administrative texts	Legislation
		Proceedings of debates
		Minutes of meetings
		Yearly reports
		Official speeches
	External communication	(Self-)presentation
		Informative documents
		Promotion/advertising
		Press releases
		Scientific texts
	Instructive texts	Manuals
		Legal documents
		Procedure descriptions
Publishers	Journalistic texts	News reporting articles
		Comment articles
	Literature	Novels
		Essayistic texts
		(Auto)biographies
		Expository works

The Dutch Parallel Corpus consists of more than ten million words, distributed over five text types, containing 2,000,000 words each. Within each text type, each translation direction contains 500,000 words. In order to preserve a good balance, the material of each cell (i.e. the unique combination of text type and translation direction) originates from at least three different providers. The exact number of words in DPC can be found in Table 11.2.[1] When compiling DPC, we were forced to make two exceptions to the global design:

- Given the difficulty to find information on translation direction for instructive texts, the condition on translation direction was relaxed for this text type.
- For literary texts, it often proved difficult to obtain copyright clearance. For that reason, the literary texts are not strictly balanced according to translation direction, but are balanced according to language pair.

The creation of a corpus that is balanced both in terms of text types and translation directions relies on a rigorous data collection process, basically consisting of two phases:

- Finding text providers who offer high-quality text material in accordance with the design prerequisites and convincing them to participate in the project.
- Clearing copyright issues for all the texts that are integrated in the corpus.

[1]The word counts are all based on clean text, meaning that all figures, tables and graphs were removed. "X" stands for unknown source language.

Table 11.2 DPC word counts per text type and translation direction

Text type	SRC ⇒ TGT	DU	EN	FR	Total
Administrative texts	EN ⇒ DU	255,155	246,137		501,292
	FR ⇒ DU	307,886		322,438	630,324
	DU ⇒ EN	249,410	257,087		506,497
	DU ⇒ FR	280,584		301,270	581,854
	Total	1,093,035	503,224	623,708	2,219,961
External communication	EN ⇒ DU	278,515	272,460		550,975
	FR ⇒ DU	233,277		250,604	483,881
	DU ⇒ EN	246,448	255,634		502,082
	DU ⇒ FR	241,323		270,074	511,397
	X ⇒ D/E	21,679	20,118		41,797
	X ⇒ D/E/F	14,192	14,953	15,743	44,888
	Total	1,035,434	563,165	536,421	2,135,020
Instructive texts	EN ⇒ DU	340,097	327,543		667,640
	FR ⇒ DU	40,487		42,017	82,504
	DU ⇒ EN	19,011	20,696		39,707
	DU ⇒ FR	110,278		115,034	225,312
	X ⇒ D/F	59,791		73,758	133,549
	X ⇒ D/E	299,996	296,698		596,694
	X ⇒ D/E/F	138,673	145,103	166,836	450,612
	Total	1,008,333	790,040	397,645	2,196,018
Journalistic texts	EN ⇒ DU	262,768	264,900		527,668
	FR ⇒ DU	240,785		265,530	506,315
	DU ⇒ EN	250,580	259,764		510,344
	DU ⇒ FR	314,989		340,319	655,308
	Total	1,069,122	524,664	605,849	2,199,635
Literature	EN ⇒ DU	148,488	143,185		291,673
	FR ⇒ DU	186,799		186,620	373,419
	DU ⇒ EN	346,802	361,140		707,942
	DU ⇒ FR	323,158		348,343	671,501
	Total	1,005,247	504,325	534,963	2,044,535
Grand total		5,211,171	2,885,418	2,698,586	10,795,175

11.2.2 Data Collection and IPR

An ideal data collection process consists of three or maybe four steps: a researcher finds adequate text material that should be included in the corpus, he/she contacts the legitimate author and asks his/her permission, the author agrees and both parties sign an agreement. As experienced during the whole project period, this process

is in reality far more complicated[2] and negotiations lasting 1–2 years were not exceptional.

As was briefly mentioned before, two main data sources can be distinguished on the basis of text provider type, namely commercial publishers versus institutions and companies. This main distinction can be considered as an anticipator on the difficulties encountered during data collection. When text production is a text provider's core business (e.g. newspaper concerns, publishing agencies, etc.), one can intuitively expect longer negotiation cycles.

Throughout the project period, clearing copyright issues proved a difficult and time-consuming task. For all IPR matters, the DPC team worked in close collaboration with the HLT agency that drew up the agreement templates.

Due to the heterogeneity of text providers (55 text providers donated texts to DPC) different types of IPR agreements were made: a standard IPR agreement, an IPR agreement for publishers, a short IPR agreement and an e-mail or letter with permission. Although specific changes often had to be made in the agreements, all texts included in the corpus were cleared from copyrights at the end of the project period. Using different agreements was a great help in managing negotiations with text providers and bringing them to a favourable conclusion. For a detailed description of data collection, IPR agreements, practical guidelines and advice, we refer to [7].

11.3 Corpus Processing

After collecting the different texts and normalizing the format, the actual processing of the corpus can start. The main task consisted in aligning the texts at sentence level (Sect. 11.3.1). The second task involved an extra layer of linguistic annotation: all words were lemmatized and grammatically tagged (Sect. 11.3.2).

The different processing stages were carried out automatically. For reasons of quality assurance, each processing stage was checked manually for 10 % of the corpus. For the other part, spot-checking and automatic control procedures were developed.

11.3.1 Alignment

The main purpose of aligning a parallel corpus is to facilitate bilingual searches. Whereas in monolingual corpora you look for a word or a series of words, in a parallel corpus you also want to retrieve the corresponding words in the other language. This kind of search is only possible if the corpus is structured in such a way that all corresponding items are aligned. During alignment a particular text

[2] In the case of a parallel corpus more parties are involved: author, translator, publisher, and foreign publisher.

chunk (e.g. a sentence) in one language is linked to the corresponding text chunk(s) in the other language. The following alignment links are used within the DPC: 1:1, 1:many, many:1, many:many, 0:1 and 1:0. Many-to-many alignments are used in the case of overlapping or crossing alignments. Zero alignments occur when no translation could be found for a sentence in either the source or the target language.

In general, there are two types of alignment algorithms: those based on sentence-length and those based on word correspondence. Very often a mixture of the two is used. The two types differ mainly in the method used: a statistical vs. a heuristic method [22]. The first type starts from the assumption that translated sentences and their original are similar in length. The correspondence between these sentences is either expressed in number of words (for example Brown et al. [2]) or in number of characters per sentence (for example Gale and Church [11]). On the basis of probability measures, the most likely alignment is then selected.

The second type of algorithms starts from the assumption that if sentences are translations of one another, the corresponding words must be translations as well. In this lexical approach the similarity of translated words is calculated on the basis of specific associative measures. To determine the degree of similarity between translated words, an external lexicon can be used, or a translation lexicon can be derived from the texts to be aligned [13]. In a more linguistic approach, one could look for morphologically related words or *cognates*, which can be very helpful for languages having similar word forms, as is the case for English and French [26].

Three different alignment tools were used to align all sentences of DPC, each of them having particular advantages and drawbacks.

The Vanilla Aligner developed by Danielsson and Ridings [6] is an implementation of the sentence-length-based algorithm of Gale and Church [11]. This tool aligns sentences within blocks of paragraphs, and therefore requires the same number of paragraphs for both languages, which can be a limitation, since the slightest shift in number of paragraphs blocks the whole alignment process. Therefore, in the DPC project, paragraph alignment has been carried out prior to sentence alignment by adopting a very pragmatic approach: only if the number of paragraphs or the size of the paragraphs differed, paragraph alignment was manually verified.

The Geometric Mapping and Alignment (GMA) developed by Melamed [19] uses a hybrid approach, based on word correspondences and sentence length. The system looks for cognates and can make use of external translation lexicons. The DPC project made use of the NL-Translex translation lexicons [12] as additional resources for recognizing word correspondences.

The Microsoft Bilingual Aligner developed by Moore [21] uses a three-step hybrid approach involving sentence and word alignment. In the first step, a sentence-length-based alignment is established. The output of the first step is then used as the basis for training a statistical word alignment model [3]. In the final step, the initial set of aligned sentences is realigned using the information from the word alignments established in the second step. The quality of the aligner is very good, but the aligner outputs only 1:1 alignments, thus disregarding all other alignment types.

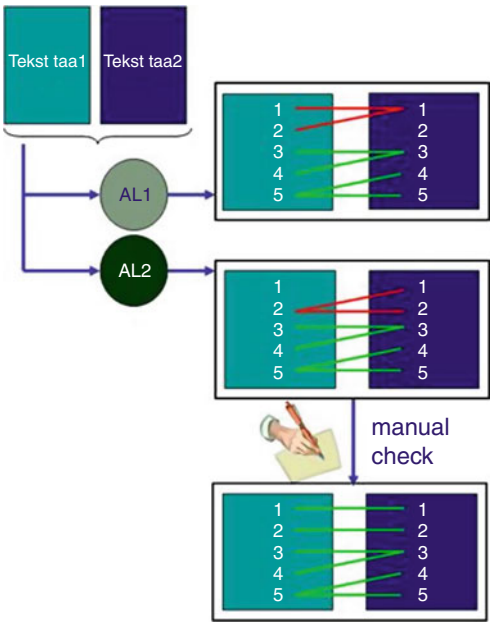


Fig. 11.1 Alignment spot check

Although each alignment tool has specific advantages and limitations, the combination of the three tools was a very helpful instrument in order to control the alignment quality of the DPC translations. Since the verification of a ten-million-word corpus is a time-consuming task, the manual verification could be limited to those cases where the three alignments diverged: when at least two aligners agreed, the alignment output could be considered of high quality. Thanks to this approach of alignment spot checks (cf. Fig. 11.1), only a small portion of the alignments was still to be checked by hand. More details on the performance of the different alignment tools used in the DPC project can be found in [15].

The entire corpus has been aligned at sentence level. The DPC also contains approximately 25,000 words of the Dutch-English part manually aligned at the sub-sentential level. These manually created reference alignments can be used to develop or test automatic word alignment systems. For more information on the sub-sentential alignments, we refer to [17].

11.3.2 Linguistic Annotation

Next to sentence alignment, the DPC data have been enriched with linguistic annotation, involving part-of-speech tagging and lemmatization to facilitate the linguistic exploration of any type of corpus. In the DPC project we have chosen to use annotation tools that are commonly available. In some cases, adaptation of the tools or pre-processing of the data was required.

Table 11.3 Performance of the PoS taggers and lemmatizers on a manually validated DPC sample

	Sample size (tokens)	Lemmata	PoS (full tag)	PoS (main category)
Dutch	211,000	96.5 %	94.8 %	97.4 %
English	300,000	98.1 %	96.2 %	N/A
French	330,000	98.1 %	94.6 %	97.4 %

For English, we opted for the combined memory-based PoS tagger/lemmatizer which is part of the MBSP tools set [5]. The English memory-based tagger was trained on data from the Wall Street Journal corpus in the Penn Treebank [18]. For Dutch, the D-Coi tagger was used [27], which is an ensemble tagger that combines the output of different machine learning algorithms. For French, we used an adapted version of TreeTagger [25].

The English PoS tagging process differs a lot from both Dutch and French grammatical annotation, in the sense that for the former a limited set of only 45 distinct tags is used, whereas both Dutch and French require a more detailed set of tags, because of their morpho-syntactic structure. In the case of Dutch, the CGN PoS tag set [28] was used, which covers word categories and subcategories, coding a wide range of morpho-syntactic features, thus amounting to a set of 315 tags. For French, we used the GRACE tag set which consists of 312 distinctive tags [23].

The tagging process for French required some adaptation of the tools, because the language model lacked lemmatized data, so that we were obliged to run the tool twice: first using the original parameter file, providing lemmata but containing only a limited tag set, and then using the enriched parameter file (provided by LIMSI [1]), containing the GRACE tag set but lacking lemmatized forms. Although the tagging process implied different processing steps, the result was also the basis for the spot check task. Similar to the alignment procedure, the combination of two annotation runs gave the necessary information to automatically detect which tags had to be verified manually. For example, if both tagging runs resulted in the same PoS tag, no further manual check was required.

The performance of the part-of-speech taggers and lemmatizers is presented in Table 11.3. The automatically predicted part-of-speech tags and lemmata were manually verified on approximately 800,000 words selected from different text types. For Dutch and French, both the accuracy score on the full tags (containing all morpho-syntactic subtags) and the score on the main tags are given. The obtained scores give an indication of the overall tagging accuracy that can be expected in DPC.

11.4 Corpus Exploitation

The final task of the DPC project consisted in packaging the data in such a way that the corpus can easily be exploited. In order to meet the requirements of different types of users, it was decided to make the corpus available in two different

Table 11.4 DPC filename patterns

Filename pattern	Description
`dpc-xxx-000000-nl-tei.xml`	Monolingual Dutch file
`dpc-xxx-000000-yy-tei.xml`	Monolingual English or French file
`dpc-xxx-000000-nl-mtd.xml`	Dutch metadata file
`dpc-xxx-000000-yy-mtd.xml`	English/French metadata file
`dpc-xxx-000000-nl-yy-tei.xml`	Alignment index file

formats. First of all, the corpus is distributed as a set of structured XML data files, which can be queried by any researcher acquainted with basic text processing skills (Sect. 11.4.1). On the other hand, a special parallel web concordancer was developed, which can easily be consulted over the internet (Sect. 11.4.2). This section describes both application modes and gives an overview of the first exploitation results of DPC (Sect. 11.4.3).

11.4.1 XML Packaging

The data have been packaged in XML in accordance with the TEI P5 standard. The choice for XML was motivated by the fact that it is a transparent format which can easily be transferred to other types of formats depending on the tools available to the developer. The XML files are well-formed and validated. The former is a basic requirement for XML files, whereas the latter gives more control over the structure of the XML files. Each XML file complies with the specifications of a basic TEI P5 DTD[3] stipulating, for example, that each word should contain attributes for part-of-speech and lemma.

For each language pair five different files are involved (cf. Table 11.4). First of all we have a text file for each language (e.g. dpc-xxx-000000-nl-tei.xml and dpc-xxx-000000-en-tei.xml representing a Dutch source file and an English target file). These data files contain the annotated sentences, where each word is grammatically tagged and lemmatized. To each data file a metadata file is linked (e.g. dpc-xxx-000000-nl-mtd.xml is the metadata file for dpc-xxx-000000-nl-tei.xml). Finally, an index file is used which contains all aligned sentences for the selected language pairs: for example, the index file dpc-xxx-000000-nl-yy-tei.xml contains all indexes for dpc-xxx-000000-nl-tei.xml and dpc-xxx-000000-en-tei.xml. The link between the different files is illustrated in Fig. 11.2.

Thanks to the validated XML format, it is possible to exploit the data files in different ways. A nice example is the development of the DPC web concordancing program—the second application mode of DPC—which is explained in the following section.

[3] A DTD (Document Type Definition) could be interpreted as a kind of text markup grammar, defining which markup elements can be used in which order.

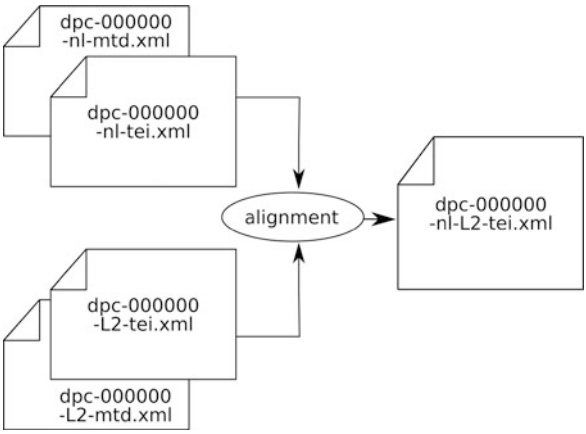


Fig. 11.2 DPC sentence-aligned files format

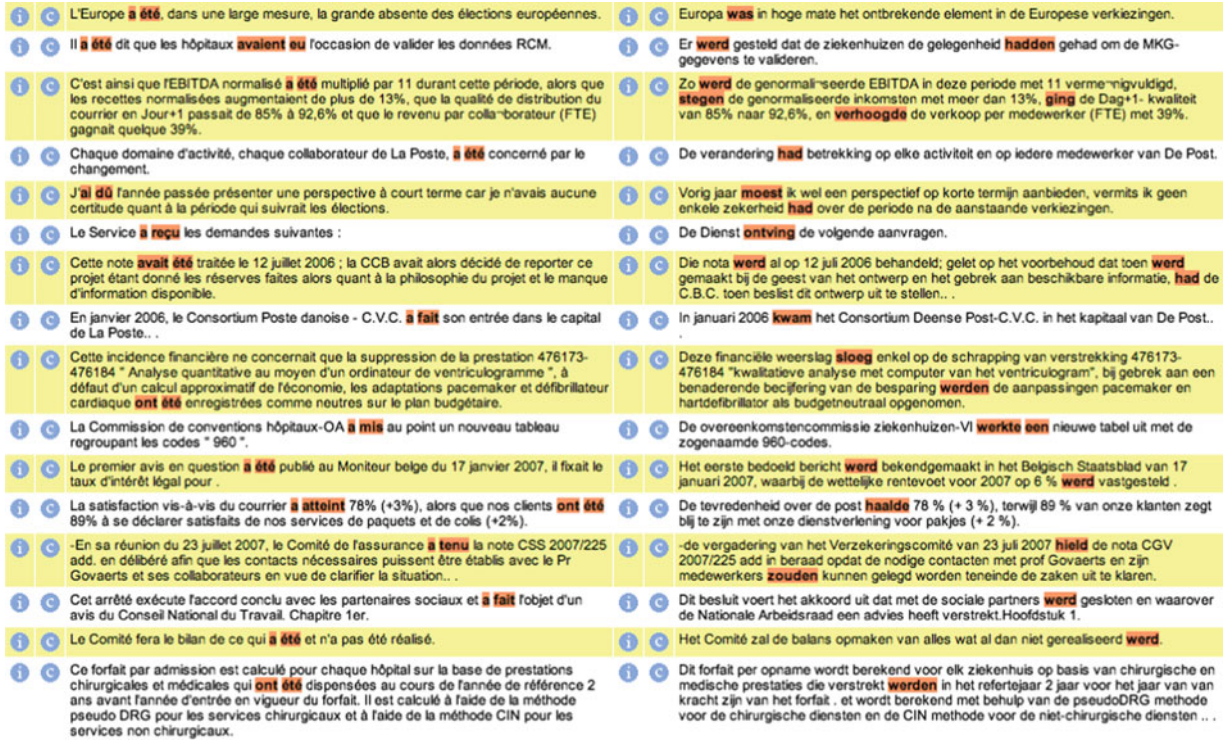
11.4.2 Parallel Web Concordance

A concordance program is a basic tool for searching a corpus for samples of particular words or patterns. Typically, the word or pattern looked for is presented in a context window, showing a certain number of context words left and right of the keyword. Therefore, such a concordancer is often called a KWIC-concordancer, referring to *keyword in context*. A parallel concordancer is a program written for displaying aligned data from a translation corpus. Since concordancers of this type are not as readily available as is the case with ordinary concordanders, and since they require a specific format, it was decided to develop a parallel concordancer especially for DPC.[4]

Parallel concordancers allows one to select words or patterns in one language and retrieve sample sentences from the selected language together with the corresponding aligned sentences in the other language. A better way consists in selecting words or patterns in the two languages. The DPC parallel concordancer is especially developed to make such an enriched bilingual search.[5] In Fig. 11.3 you can see the first output page of a combined query, which looks for French-Dutch text samples of the French *passé composé* matching the Dutch *verleden tijd* (simple past). The output is inevitably obscured by some noise—mainly due to complex sentence structure—, but the result is rich enough to allow researchers to further analyze the output, without having to call in the help of programmers. There is an exporting module to Excel, so that researchers can annotate the results in a more commonly used working format.

[4]The original web interface was developed by Geert Peeters and Serge Verlinde (ILT KU Leuven).

[5]A demo version of the parallel concordancer is available at the HLT agency via the following link: http://dpc.inl.nl/indexd.php



L'Europe **a été**, dans une large mesure, la grande absente des élections européennes.	Europa **was** in hoge mate het ontbrekende element in de Europese verkiezingen.
Il **a été** dit que les hôpitaux **avaient eu** l'occasion de valider les données RCM.	Er **werd** gesteld dat de ziekenhuizen de gelegenheid **hadden** gehad om de MKG-gegevens te valideren.
C'est ainsi que l'EBITDA normalisé **a été** multiplié par 11 durant cette période, alors que les recettes normalisées augmentaient de plus de 13%, que la qualité de distribution du courrier en Jour+1 passait de 85% à 92,6% et que le revenu par collaborateur (FTE) gagnait quelque 39%.	Zo **werd** de genormaliseerde EBITDA in deze periode met 11 vermenigvuldigd, **stegen** de genormaliseerde inkomsten met meer dan 13%, **ging** de Dag+1- kwaliteit van 85% naar 92,6%, en **verhoogde** de verkoop per medewerker (FTE) met 39%.
Chaque domaine d'activité, chaque collaborateur de La Poste, **a été** concerné par le changement.	De verandering **had** betrekking op elke activiteit en op iedere medewerker van De Post.
J'**ai dû** l'année passée présenter une perspective à court terme car je n'avais aucune certitude quant à la période qui suivrait les élections.	Vorig jaar **moest** ik wel een perspectief op korte termijn aanbieden, vermits ik geen enkele zekerheid **had** over de periode na de aanstaande verkiezingen.
Le Service **a reçu** les demandes suivantes :	De Dienst **ontving** de volgende aanvragen.
Cette note **avait été** traitée le 12 juillet 2006 ; la CCB avait alors décidé de reporter ce projet étant donné les réserves faites alors quant à la philosophie du projet et le manque d'information disponible.	Die nota **werd** al op 12 juli 2006 behandeld; gelet op het voorbehoud dat toen **werd** gemaakt bij de geest van het ontwerp en het gebrek aan beschikbare informatie, **had** de C.B.C. toen beslist dit ontwerp uit te stellen.. .
En janvier 2006, le Consortium Poste danoise - C.V.C. **a fait** son entrée dans le capital de La Poste.. .	In januari 2006 **kwam** het Consortium Deense Post-C.V.C. in het kapitaal van De Post.. .
Cette incidence financière ne concernait que la suppression de la prestation 476173-476184 " Analyse quantitative au moyen d'un ordinateur de ventriculogramme ", à défaut d'un calcul approximatif de l'économie, les adaptations pacemaker et défibrillateur cardiaque **ont été** enregistrées comme neutres sur le plan budgétaire.	Deze financiële weerslag **sloeg** enkel op de schrapping van verstrekking 476173-476184 "kwalitatieve analyse met computer van het ventriculogram", bij gebrek aan een benaderende becijfering van de besparing **werden** de aanpassingen pacemaker en hartdefibrillator als budgetneutraal opgenomen.
La Commission de conventions hôpitaux-OA **a mis** au point un nouveau tableau regroupant les codes " 960 ".	De overeenkomstencommissie ziekenhuizen-VI **werkte een** nieuwe tabel uit met de zogenaamde 960-codes.
Le premier avis en question **a été** publié au Moniteur belge du 17 janvier 2007, il fixait le taux d'intérêt légal pour .	Het eerste bedoeld bericht **werd** bekendgemaakt in het Belgisch Staatsblad van 17 januari 2007, waarbij de wettelijke rentevoet voor 2007 op 6 % **werd** vastgesteld .
La satisfaction vis-à-vis du courrier **a atteint** 78% (+3%), alors que nos clients **ont été** 89% à se déclarer satisfaits de nos services de paquets et de colis (+2%).	De tevredenheid over de post **haalde** 78 % (+ 3 %), terwijl 89 % van onze klanten zegt blij te zijn met onze dienstverlening voor pakjes (+ 2 %).
-En sa réunion du 23 juillet 2007, le Comité de l'assurance **a tenu** la note CSS 2007/225 add. en délibéré afin que les contacts nécessaires puissent être établis avec le Pr Govaerts et ses collaborateurs en vue de clarifier la situation... .	-de vergadering van het Verzekeringscomité van 23 juli 2007 **hield** de nota CGV 2007/225 add in beraad opdat de nodige contacten met prof Govaerts en zijn medewerkers **zouden** kunnen gelegd worden teneinde de zaken uit te klaren.
Cet arrêté exécute l'accord conclu avec les partenaires sociaux et **a fait** l'objet d'un avis du Conseil National du Travail. Chapitre 1er.	Dit besluit voert het akkoord uit dat met de sociale partners **werd** gesloten en waarover de Nationale Arbeidsraad een advies heeft verstrekt.Hoofdstuk 1.
Le Comité fera le bilan de ce qui **a été** et n'a pas été réalisé.	Het Comité zal de balans opmaken van alles wat al dan niet gerealiseerd **werd**.
Ce forfait par admission est calculé pour chaque hôpital sur la base de prestations chirurgicales et médicales qui **ont été** dispensées au cours de l'année de référence 2 ans avant l'année d'entrée en vigueur du forfait. Il est calculé à l'aide de la méthode pseudo DRG pour les services chirurgicaux et à l'aide de la méthode CIN pour les services non chirurgicaux.	Dit forfait per opname wordt berekend voor elk ziekenhuis op basis van chirurgische en medische prestaties die verstrekt **werden** in het refertejaar 2 jaar voor het jaar van van kracht zijn van het forfait . et wordt berekend met behulp van de pseudoDRG methode voor de chirurgische diensten en de CIN methode voor de niet-chirurgische diensten .. .

Fig. 11.3 Parallel concordancer output sample

Although it is possible to develop a full featured query interface, which allows for exploitation using regular patterns,[6] we have decided to restrict the interface to a small set of query patterns, transparent enough for non-experts to be able to find their way in exploring the parallel corpus without much hassle. Further exploitation is possible, if you analyse the XML source files, using XSLT or similar tools.

The DPC concordancer differs from similar parallel concordancers, in the sense that DPC has been provided with a extra annotation layers (PoS tags and lemmatization, and metadata), which allow for better selections, not possible in ParaConc or Multiconcord.[7] In the DPC concordancer, you can build subcorpora, based on metadata of text types and language filters. In the case of ParaConc, you cannot filter on extra annotation layers.

ParaConc and Multiconcord are platform specific. The first is available for Windows and Macintosh, the other only for Windows. The DPC concordancer is available over the internet and therefore not specifically linked to one platform. The DPC concordancer is freely available, but unlike the two others, adding new texts is not directly available.

11.4.3 First Exploitation Results of DPC

As mentioned in the introduction, it was the explicit aim of the DPC project to create a parallel corpus that satisfies the needs of a diverse user group. Since its (pre-)release DPC has been used in different research domains[8]:

- In the CAT domain, DPC has been used to select benchmarking data to evaluate different translation memory systems [14] and to extract language-pair specific translation patterns that are used in a chunk-based alignment system for English-Dutch [16].
- In the domain of CALL, DPC has been introduced as a valuable resource for language teaching. The corpus is being used as a sample repository for content developers preparing exercises for French and Dutch language learners [9]. Within CorpusCALL, parallel corpora like DPC are used as resources for data-driven language learning [20]. Parallel corpora are also useful instruments for rethinking the pedagogical grammaticography in function of frequency research. On the basis of such analysis one can find out, for example, how to teach the *subjonctif* for learners of French [29].

[6]Extended regular patterns are used in CQP (Corpus Query Processor) developed by IMS, and originally developed for CWB (Corpus Work Bench) (cf. also [4].)

[7]See http://www.athel.com/paraweb.pdf and http://artsweb.bham.ac.uk/pking/multiconc/lingua.htm for ParaConc and Multiconcord respectively.

[8]This is a non-exhaustive list as the authors are only aware of research making use of DPC conducted at their own institutions.

- In the framework of the DPC-project, a Gold Standard for terminology extraction was created. All terms (single- and multiword terms) were manually indicated in a set of texts belonging to two different domains (medical and financial). This Gold Standard has been used by Xplanation,[9] who as an industrial partner of the DPC project was partly responsible for the external validation of DPC.[10] It is also used in the TExSIS project[11] as benchmarking data to evaluate bilingual terminology extraction programmes.
- In the field of Translation Studies, DPC has been used as comparable corpus to study register variation in translated and non-translated Belgian Dutch [8] and [10]. More particulary, it was investigated to what extent the conservatism and normalization hypothesis holds in different registers of translated texts, compared to non-translated texts.
- Contrastive linguistics is another field where DPC has been used as a resource of authentic text samples. Vanderbauwhede [30, 31] studied the use of the demonstrative determiner in French and Dutch on the basis of corpus material from learner corpora and parallel corpora, including DPC. Although Dutch and French use the article and the demonstrative determiner in a quite similar way, parallel corpus evidence shows some subtle differences between both languages.

Furthermore, DPC is being used in a number of courses in CALL, translation studies and language technology. A substantial part of DPC has also been used for further syntactic annotation in the Lassy project (cf. Chap. 9, p. 147).

11.5 Conclusion

The DPC project resulted in a high-quality, well-balanced parallel corpus for Dutch, English and French.[12] Its results are available via the HLT Agency.[13] As part of the STEVIN objectives to produce qualitative resources for Dutch natural language processing, DPC is a parallel corpus that meets the requirements of the STEVIN programme.[14] The DPC corpus differs mainly from other parallel corpora in the following ways: (i) special attention has been paid to corpus design, which resulted in a well-balanced corpus, (ii) the corpus is sentence-aligned and linguistically annotated (PoS tagging and lemmatization), (iii) the different processing steps have been controlled in a systematic way and (iv) the corpus is available to the wide research community thanks to its copyright clearance.

[9] http://www.xplanation.com

[10] The Center for Sprogteknologi (CST) carried out a formal validation of DPC.

[11] http://lt3.hogent.be/en/projects/texsis/

[12] For further information, see the DPC project website: http://www.kuleuven-kortrijk.be/DPC

[13] http://www.tst-centrale.org

[14] A summary of the STEVIN requirements is given in the introduction of this book (cf. p. 1).

DPC is first of all used as a resource of translated texts for different types of applications, but also monolingual studies of Dutch, French and English can benefit greatly from it. The quality of the corpus—in content and structure—and the two application modes provided (XML and web interface) help to explain why the first exploitation results of DPC are promising.

References

1. Allauzen, A., Bonneau-Maynard, H.: Training and evaluation of POS taggers on the French MULTITAG corpus. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC-08), Marrakech, pp. 28–30 (2008)
2. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, pp. 169–176 (1991)
3. Brown, P.F., Della Pietra, V.J., Della Pietra, S.A., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
4. Christ, O.: A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX, Conference on Computational Lexicography and Text Research, Budapest, pp. 23–32 (1994)
5. Daelemans, W., van den Bosch, A.: Memory-Based Language Processing. Cambridge University Press, Cambridge (2005)
6. Danielsson, P., Ridings, D.: Practical presentation of a vanilla aligner. In: Reyle, U., Rohrer, C. (eds.) The TELRI Workshop on Alignment and Exploitation of Texts, Ljubljana (1997)
7. De Clercq, O., Montero Perez, M.: Data collection and IPR in multilingual parallel corpora: Dutch parallel corpus. In: Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010), Valletta, pp. 3383–3388 (2010)
8. Delaere, I., De Sutter, G., Plevoets, K.: Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. Target, **24**(2), (2012)
9. Desmet, P., Eggermont, C.: FRANEL: un environnement électronique d'apprentissage du français qui intègre des matériaux audio-visuels et qui est à la portée de tous. Cahiers F: revue de didactique français langue étrangère / Cahiers F: didactisch tijdschrift Frans vreemde taal pp. 39–54 (2006)
10. De Sutter, G., Delaere, I., Plevoets, K.: Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling. In: Oakes, M., Ji, M. (eds.) Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide To Descriptive Translation Research, pp. 325–345. John Benjamins, Amsterdam (2012)
11. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Comput. Linguist. **19**(1), 75–102 (1993)
12. Goetschalckx, J., Cucchiarini, C., Van Hoorde, J.: Machine translation for Dutch: the NL-Translex project, Brussels/Den Haag, 16pp (2001)
13. Kay, M., Röscheisen, M.: Text-translation alignment. Comput. Linguist. **19**(1), 121–142 (1993)
14. Macken, L.: In search of the recurrent units of translation. In: Daelemans, W., Hoste, V. (eds.) Evaluation of Translation Technology. LANS 8/2009, pp. 195–212. Academic and Scientific Publishers, Brussels (2009)

15. Macken, L.: Sub-sentential alignment of translational correspondences. Ph.D. thesis, University of Antwerp (2010)
16. Macken, L., Daelemans, W.: A Chunk-Driven Bootstrapping Approach to Extracting Translation Patterns. In: Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (Iasi, Romania). Lecture Notes in Computer Science, vol. 6009, pp. 394–405. Springer, Berlin/ Heidelberg (2010)
17. Macken, L., De Clerq, O., Paulussen, H.: Dutch parallel corpus: a balanced copyright-cleared parallel corpus. Meta **56**(2), 374–390 (2011)
18. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Comput. Linguist. **19**(2), 313–330 (1993)
19. Melamed, D.I.: A portable algorithm for mapping bitext correspondence. In: Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL), Madrid, pp. 305–312 (1997)
20. Montero Perez, M., Paulussen, H., Macken, L., Desmet, P.: From input to output: the potential of parallel corpora for CALL. LRE (Submitted)
21. Moore, R.: Fast and Accurate Sentence Alignment of Bilingual Corpora. Machine Translation: From Research to Real Users. 2499, 135–144 (2002)
22. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**, 19–51 (2003)
23. Paroubek, P.: Language resources as by-product of evaluation: the multitag example. In: Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, pp. 151–154 (2000)
24. Rura, L., Vandeweghe, W., Montero Perez, M.: Designing a parallel corpus as a multifunctional translator's aid. In: Proceedings of XVIII FIT World Congress, Shangai, pp. 4–7 (2008)
25. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester (1994)
26. Simard, M., Foster, G., Hannan, M.L., Macklovitch, E., Plamondon, P.: Bilingual text alignment: where do we draw the line? In: Botley, S., McEnery, A., Wilson, A. (eds.) Multilingual Corpora in Teaching and Research, pp. 38–64. Rodopi, Amsterdam (2000)
27. van den Bosch, A., Schuurman, I., Vandeghinste, V.: 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 (2006)
28. Van Eynde, F., Zavrel, J., Daelemans, W.: 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, pp. 1427–1434 (2000)
29. Van Keirsbilck, P., Lauwers, P., Desmet, P.: Le subjonctif tel qu'il s'enseigne en Flandre et en France: bilan et perspectives. Travaux de didactique du FLE. 64, 131–145 (2010)
30. Vanderbauwhede, G.: Le déterminant démonstratif en français et en néerlandais à travers les corpus: théorie, description, acquisition. Ph.D. thesis, K.U. Leuven (2011)
31. Vanderbauwhede, G.: Les emplois référentiels du SN démonstratif en français et en néerlandais: pas du pareil au même. J. Fr. Lang. Stud. **22**(2), 273–294 (2012) doi:10.1017/S0959269511000020