DESIGNING A PARALLEL CORPUS AS A MULTIFUNCTIONAL TRANSLATOR'S AID

Lidia Rura
Ghent University College, Belgium
lidia.rura@hogent.be

Willy Vandeweghe
Ghent University College, Belgium
willy.vandeweghe@hogent.be

Maribel Montero Perez
K.U. Leuven, Belgium
maribel.monteroperez@kuleuven-kortrijk.be

Abstract: 2006 saw the start of a project for compiling a multifunctional parallel corpus with Dutch as a pivotal language: the Dutch Parallel Corpus (DPC). Among other things, parallel corpora can be a useful tool in translation business. They can help to improve idiomatic language usage, provide translation suggestions or serve for filling up translation memory with high-quality data. The advantages of parallel corpora over multilingual comparable corpora or dictionaries/ glossaries is search speed and reliability. They contain a great amount of aligned data, examples from which can be viewed in the surrounding context. Besides, parallel corpora offer their user the benefit of metadata with additional information allowing for a finer-tuned search of the corpus. The corpus design and text typology are crucial for the usability of the corpus. Insights from cognitive linguistics on basic-level categories have proven to be useful for elaborating such a design and typology assuring (i) text type diversity containing translation samples from different areas of expertise; (ii) high translation quality providing reliable translation solutions; (iii) a well-structured taxonomy for prompt data retrieval.

Key Words: parallel corpora; translation process; corpus design; text typology; metadata;

1. A PARALLEL CORPUS AS A USEFUL TRANSLATION TOOL

1.1 Background

1.1.1 Translation and parallel corpora

The last decennia have seen a growing interest for text corpora including parallel corpora as they proved to be useful for many purposes in different areas of expertise. In order to avoid confusion as to what is understood under a parallel corpus in this paper, we define a parallel corpus as a collection of texts in different languages that are translations of each other (cf. Baker 1995:230), as opposed to a comparable multilingual corpus containing texts in different languages on the same topic that are not translations of each other.

Parallel corpora have so far been mostly used either for research or for the development of Human Language Technology applications, but this paper reports on possible usages of parallel corpora for the translation process. Development of CAT tools (Computer-Aided Translation), information and terminology extraction as well as machine translation are connected to translation but they are left out of consideration here since they require a rather complex and costly digital infrastructure and are thus only affordable for big companies and institutions. Nor will the paper deal with the possible use of parallel corpora for training would-be translators as this task mostly pertains to the area of education and not to the translation production as such.

Instead, we will focus on possible usages of parallel corpora that lie within the reach of smaller users, such as translation agencies or even individual translators. Many translators have already discovered that parallel texts can be an excellent aid, some also try to set up corpora of their own, consisting either of multilingual comparable texts or parallel ones. However, the compilation of such a corpus is a time-consuming and technically-demanding task: finding suitable texts and aligning them on different levels. Few translators of non-fictional, specialised texts ever have the luxury of spending time on this since they usually work within a tight schedule. Nor does the average translator have the necessary computer knowledge to turn a collection of texts into a functional corpus.

1.1.2 Parallel corpus types

Further, it is necessary to delimit the parallel corpora that will be discussed. One could roughly divide the existing parallel corpora in three types:

(i) specialised parallel corpora created for corporate clients with a particular usage in mind, e.g. parallel data accumulated by big manufacturing entreprises consisting of manuals and product specifications. Such corpora are normally of good quality but hardly ever accessible to outsiders.

(ii) parallel corpora consisting of readily-available data accumulated and stored anyway, like the Europarl and the UN Parallel Corpus. Such corpora comprise many languages and a huge amount of data but they are only automatically aligned without manual verification, hardly ever annotated and often consist of monotonous data: proceedings, parliamentary documents etc.

(iii) parallel corpora compiled within the framework of scientific projects with a predetermined corpus design and data selection. Such corpora tend to comprise a greater text type and topic variety, they are often annotated, with assured translation quality and are therefore much more interesting for translation purposes. The user's licence is mostly not free for commercial users but still affordable and can prove to be a valuable investment.

This paper reports on the usability of this last type of parallel corpora, which will be in the end illustrated using the example of the Dutch Parallel Corpus, described in more detail in (Macken, Trushkina & Rura, 2007) and (Paulussen, Macken, Trushkina, Desmet & Vandeweghe, 2007).

1.2 Translators and Parallel Corpora

1.2.1 Translator's needs

Parallel corpora are indispensable for translation studies and training translators (Baker 1995:231). Nowadays, they have become just as indispensable for the translation process itself, in fact their importance is so great that they cannot not be neglected in modern translation business. The translator needs resources that can provide an interpretation suggestion of the source text and a translation hypothesis. In practice about 50% of the time necessary for a translation is spent on consulting reference materials (Aston 1999). Electronic corpora can play an important role in improving the quality and speed of the translation process (ibid).

1.2.2 Shortcomings of traditional translator's aids

Parallel corpora come in handy when no good or specialised dictionaries are available, and this is the case for many languages, especially the less widely-used ones. For example, there exist more bidirectional dictionaries for English and French than for Dutch, which forces translators working with this language to rely on other resources such as glossaries and internet search for parallel or comparable texts. Yet, internet search is rather time consuming while glossaries and dictionaries provide no context and make it therefore difficult to decide which word is needed in a particular case.

Dictionaries/ glossaries often fall short when it comes to meaning disambiguation or collocations, e.g. there exist many collocation dictionaries for English but none for French and only a few for Dutch. The weakness of the internet search is the quality issue. "Since the world-wide web is an ever- changing entity of dubious authority whose overall composition is unknown, considerable care must … be exercised in selecting texts" (Aston 1999).

1.3 Advantages of Professionally-Compiled Parallel Corpora

A professionally-compiled parallel corpus has neither of the above-mentioned disadvantages. On the contrary, it provides solutions to many of the above-mentioned problems just at one's fingertips.

Firstly, it consists of aligned parallel texts that provide a “greater certainty as to the equivalence of particular expressions” (Aston 1999). The user can promptly locate all the occurrences of any expression in the other language together with the surrounding context (ibid.), providing thereby clues to the disambiguation of the word, the appropriate register, the needed collocation and the correct grammatical usage in one shot. The concordances in figures 1 and 2 illustrate this.

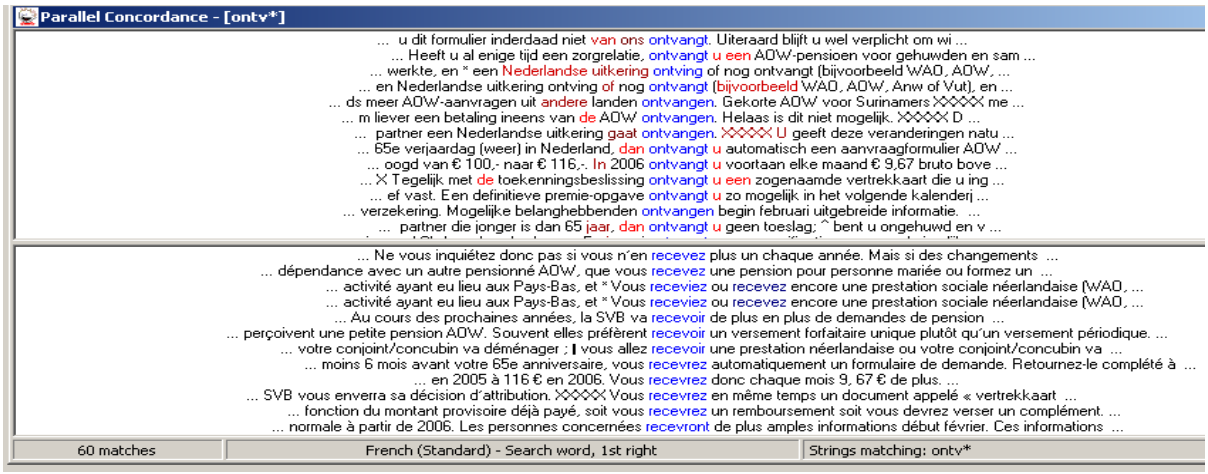


Figure 1. Query Example Verb Phrase (DPC)

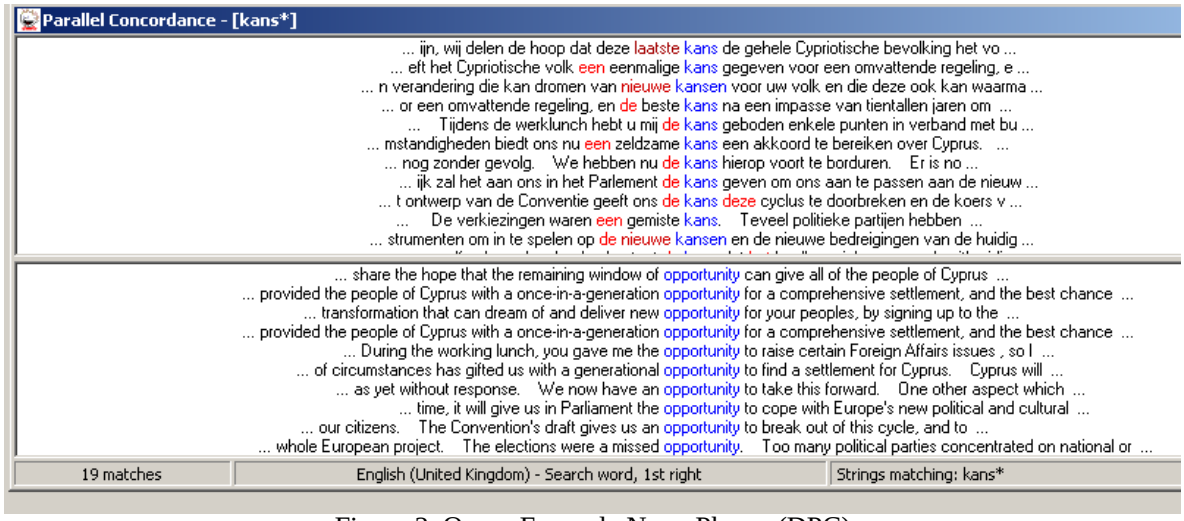
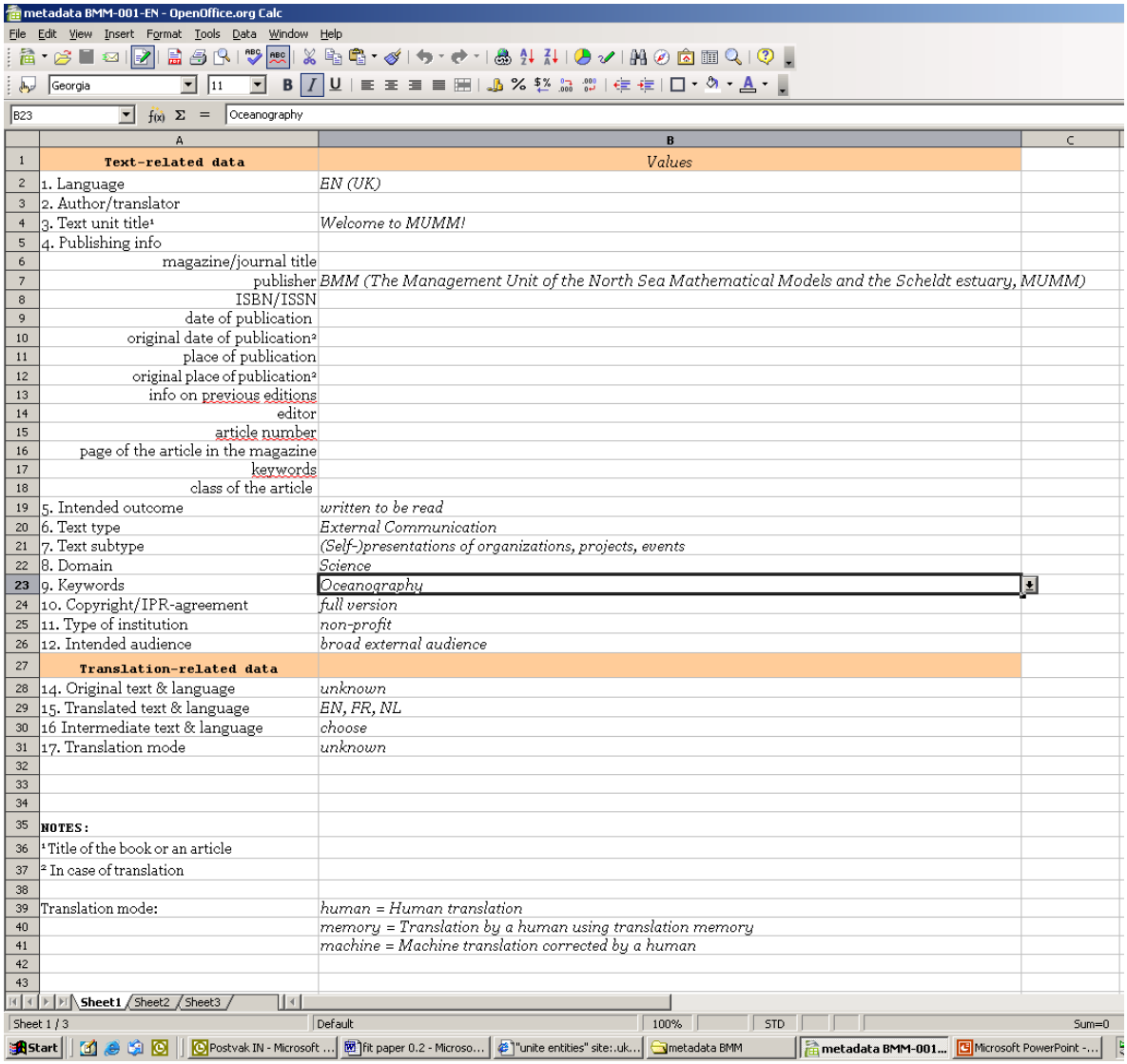


Figure 2. Query Example Noun Phrase (DPC)

Secondly, it contains metadata with information on the source/ target language, translation mode, parallel texts in other languages, the topic, as well as the origins and authorship of the text sample. Besides, the metadata can help the translator to appreciate whether a particular text was written for laymen or specialists and whether the language usage in concrete cases is generally adopted or a regional variety.



metadata BMM-001-EN - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

Georgia 11

B23 Oceanography

	A	B	C
1	**Text-related data**	*Values*	
2	1. Language	*EN (UK)*	
3	2. Author/translator		
4	3. Text unit title[1]	*Welcome to MUMM!*	
5	4. Publishing info		
6	magazine/journal title		
7	publisher	*BMM (The Management Unit of the North Sea Mathematical Models and the Scheldt estuary, MUMM)*	
8	ISBN/ISSN		
9	date of publication		
10	original date of publication[2]		
11	place of publication		
12	original place of publication[2]		
13	info on previous editions		
14	editor		
15	article number		
16	page of the article in the magazine		
17	keywords		
18	class of the article		
19	5. Intended outcome	*written to be read*	
20	6. Text type	*External Communication*	
21	7. Text subtype	*(Self-)presentations of organizations, projects, events*	
22	8. Domain	*Science*	
23	9. Keywords	*Oceanography*	
24	10. Copyright/IPR-agreement	*full version*	
25	11. Type of institution	*non-profit*	
26	12. Intended audience	*broad external audience*	
27	**Translation-related data**		
28	14. Original text & language	*unknown*	
29	15. Translated text & language	*EN, FR, NL*	
30	16 Intermediate text & language	*choose*	
31	17. Translation mode	*unknown*	
32			
33			
34			
35	**NOTES:**		
36	[1] Title of the book or an article		
37	[2] In case of translation		
38			
39	Translation mode:	*human = Human translation*	
40		*memory = Translation by a human using translation memory*	
41		*machine = Machine translation corrected by a human*	
42			
43			

Sheet1 Sheet2 Sheet3

Sheet 1 / 3 Default 100% STD Sum=0

Start Postvak IN - Microsoft ... fit paper 0.2 - Microso... "unite entities" site:.uk... metadata BMM metadata BMM-001... Microsoft PowerPoint -...

Figure 3. Example of a Corpus Metadata File (DPC)

Thirdly, it normally comprises texts from many different domains, making them useful for translation on different topics and therefore handy for translators and translation agencies working in different areas of expertise. Differentiation according to the text type is also helpful to diminish the “likelihood of encountering other senses of polysemous items” (Aston 1999).

Fourthly, it offers a good quality guarantee, which is imperative in order to make it a reliable source for a translator. Professionally-compiled parallel corpora are mostly created within the framework of well-funded projects giving the compilers the time and opportunity to collect quality-assured data, such as translations produced by big translation agencies and/ or translation divisions that employ native speakers, proofread and revise their translations. Quality in the case of professionally produced translations also assures the idiomaticity of the target text, which is valuable for translating into a foreign language, when translators feel unsure about the choices they make (Aston 1999).

Finally, a parallel corpus can be a nice complement to CAT tools, since it can e.g. be used to quickly fill up the translation memory with high-quality data instead of accumulating text pairs for years.

2. CORPUS DESIGN AS A KEY TO OPTIMISING CORPUS USE

2.1 Typical Difficulties Encountered by Corpus Designers

There is no universal recipe for designing a parallel corpus. The data selection process for parallel corpora is characterised by a number of limitations such as availability of translated data, quality of the translated material and proportional availability of translated material for all targeted languages and translation directions (Olohan, 2004: 25). This means parallel corpora mostly have a design, tailored to the concrete situation of the country/ culture and language(s) concerned.

A parallel corpus can be unidirectional: from language A to language B or bidirectional: both from language A to language B and vice versa. The last type is more valuable for translators as it provides examples for comparison in both languages. However, building a bidirectional parallel corpus poses "additional difficulties since material is seldom translated between two languages in equal quantities… Also the nature and/ or range of translated material may differ: high-brow vs main-stream"(Olohan 2004:25). This is especially true in case of less-widely used languages such as Dutch.

Besides, the more languages the corpus contains, the more difficult it is to find parallel data with a version in every chosen language. "This is one of the reasons why European Union texts are often used", but these have their problems as well: "Texts originating in the institutions of the European Union can be problematic …since it can be difficult to assign the status of 'source text' to one of the language versions; documents may be written in more than one language and, once translations exist there is nothing to distinguish a source text from the other language versions" (Koskinen 2000:55).

2.2 Suggestions for Optimising a Parallel Corpus Design

In order for a parallel corpus to be useful for translators working in different areas of expertise, it has first of all to be well-balanced and compliant with certain principles. Balancing a corpus with respect to the amount of data for every translation direction implies a preliminary survey of existing translated texts before determining the corpus typology.

2.2.1 Text type diversity

Even if corpus typology is initially determined by availability, it does not mean that corpus compilers should limit themselves to readily-available data. For the corpus to be a valuable translation tool, it has to contain various data and especially those for which there exist few other resources such as dictionaries, glossaries, etc. The greater text diversity in the corpus, the more universal its potential use. So the compilers of a multifunctional corpus should also try to obtain data interesting for translators. For instance, it is easier to obtain government data than corporate data, but government institutions produce mostly jargon-ridden administrative texts with outdated vocabulary. An example is shown in figure 4.

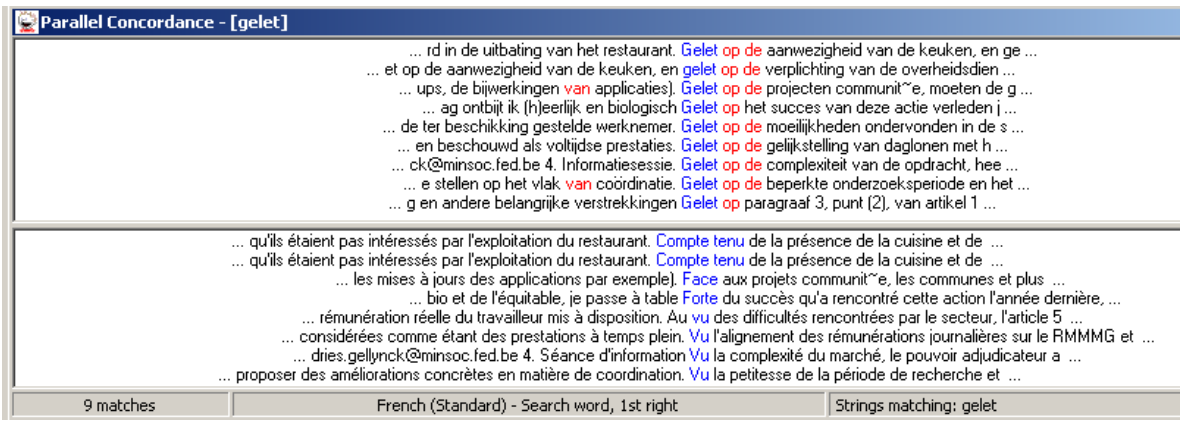


Figure 4. Example of an Administrative Text (DPC)

Moreover, government organisations tend to translate their texts in one direction only: from the official country language into major foreign languages such as English and French. Thus, filling the corpus up with mainly institutional texts is undesirable.

2.2.2 Corpus samples

A functional corpus should consist of adequate samples that are big and representative enough to provide a reliable translation suggestion. Ideally, the samples should be full texts with sufficient lexical and grammatical diversity (Olohan 2004:57), but for some text types it is problematic, especially for texts originating from commercial publishers. They are often afraid that a database like a corpus may intentionally or unintentionally serve as a cheap or even free distribution channel for the texts they themselves sell for money. There is, however, a technical solution to this problem that may ease their worries: the corpus interface can be equipped with a safeguard mechanism allowing the user to view only a small fragment of the same text at a time.

In cases when only fragments can be obtained, the corpus compilers might want to allow for some variation in the sample type so that the corpus data will provide not only examples of the beginnings of texts but also of the central part and the end.

2.2.3 Corpus structure

Seminal works on the subject mention three major steps for designing text typology in corpora: (i) delimiting the target population of texts, (ii) determining text categories (e.g. text types, genres, topics, etc.) and (iii) finding a way of organising the typology, i.e. designing a corpus taxonomy.

(i) The first step, the selection of suitable and available data, has already been discussed above in 2.1 en 2.2.

(ii) The second step, choosing suitable categories for text typology has far-reaching consequences for the corpus transparency. In order to delimit text categories one needs to have criteria that would define them. However, many criteria normally used for text corpora are not expedient in a parallel corpus since they ‘in general have been developed by linguists, and on the basis of monolingual corpora only’ (Baker 1995: 229-230)

The existing criteria can be roughly divided into two main types, on which there is an international consensus: external (situational) criteria and internal (linguistic) criteria (EAGLES, 1996: 4, 8, 16). Both types have their merits and drawbacks, but the internal ones are deemed more objective. EAGLES states, however, that it is impossible to have a classification based exclusively on internal criteria since the technology still lacks the necessary level of sophistication and because a classification based purely on internal criteria may obscure the relationship between the linguistic and non-linguistic criteria (EAGLES, 1996: 7, 21). Besides, many linguistically-defined categories prove in practice to be vague and overbroad resulting in text collections with heterogeneous material, or differentiation of similar texts. Reiss for instance, provided a typology with three ‘pure’ text types: informative, expressive and operative (Reiss 1981:124). However, Reiss herself points out that texts rarely, if ever, exist in their ‘pure’ forms, and that most texts consist of mixes of the above types. Another example of a linguistically-defined typology comes from Longacre (1983) “who distinguished four text types: narration, procedural discourse, behavioural discourse, and expository discourse” (Fludernik 2000).

Internal categories would be of little help for a translator who wants to search the corpus in order to find a particular kind of document because it can be classified into several of the above-mentioned categories. Categories do not only have to be theoretically defendable but also practical. A parallel corpus used for translation purposes should be easy to navigate. Therefore, most corpora rely on external criteria.

A possible external category is the one defined by text topic. However, topic classification used “so far in many corpus designs has been to erect a make-shift and broad-mesh framework, within which the texts are disposed into undefined or inadequately defined categories”, which makes topic (or subject) typology

unadvisable and unmanageable (EAGLES, 1996: 5,6). Instead, EAGLES recommends reusing existing and established categories: "In establishing a typology for texts in computer corpora, we should not forget categories of classification that are already in use"(EAGLES, 1996: 8). Also we can use the fact that a society institutionalises a number of topic-related classifications; of particular value to text typology are lists of recognised professions and educational courses (EAGLES, 1996: 6).

Another way of determining text categories may be offered by the prototype approach to category delimitation, which has been advocated by some authors writing in the cognitive linguistics vein (Halverson, 1998; Lee, 2001). This approach builds on cognitive saliency in delimiting the categories. The prototype categories otherwise known as basis-level categories "represent the preferred cognitive perspective. They seem to meet 'basic' cognitive needs because they pinpoint where the focus of human interest lies" (Ungerer & Schmid 1996: 61-62). They differ from "classic scientific taxonomies (Linnaeus), which are too complex and rigid, leaving no room for change of perception" (ibid). Besides, "…a prototype category structure provides us with a means of addressing the relativity of definitions, ….It also provides us with a means of coping with the quite obvious asymmetries in the category members" (Halverson 1998:18). Another way of looking at prototypes, is seeing them as "'best examples', which are seen as cognitive reference points" (Olohan 2004:17).

Since basic-level categories represent best examples of a category, they do not have to be definite, they just unite entities through typical attributes. Therefore, they allow for permeable boundaries, as one entity can belong to different categories through different attributes and yet remain cognitively salient both to the corpus compilers and the corpus users.

(iii) The third step in designing a corpus implies structuring the chosen categories in a corpus taxonomy, which is important for the corpus transparency, navigability and efficient data retrieval.

Some corpora have several categories existing independently of each other, which forces the user to find the crossing point between the categories by himself. However, such a taxonomy implies an additional effort and is not always expedient in a corpus used by translators who would be unwilling to invest time in such a search.

Another possibility of organising a corpus is the introduction of several levels of categorisation, thereby creating a subtypology. Lee (2001:48) proposes a categorisation with the basic level as the building stone of his three levels typology and the superordinate and subordinate levels built around it.

Table 1. Example of Text Typology Based on Basic-Level Categories by D.Y. Lee

SUPERORDINATE	Mammal	Literature	Advertising
BASIC-LEVEL	Dog/Cat	Novel/ Poem/ Drama	Advertisement
SUBORDINATE	Cocker Spaniel/ Siamese	Western/ Romance/ Adventure	Print ad, Radio ad, TV ad, Tshirt ad

If the basic-level categories can be described as focal since they highlight central attributes of an entity, superordinate categories unite basic-level categories by highlighting "salient general attributes" (Ungerer & Schmid 1996: 78-79). The subordinates, which are situated below the basic-level categories, do not unite but make a distinction between categories through specific attributes that "are not normally shared by other categories" (Ungerer & Schmid 1996:88).

The subtype labels obtained in this way, refer to categories which are known from everyday experience - cf. the 'Primärtexte' in the taxonomy by Göpferich (1995) -, that neither require abstract thinking nor specialised knowledge in order to be grasped. This leads to a workable and transparent corpus structure, where all levels are easily identifiable through cognitively recognisable attributes.

2.2.4 Metadata criteria

Metadata represent an important tool for corpus navigability, they contain additional information that would allow the user to make a precise selection. Here are some examples of metadata potentially relevant to translators.

2.2.4.1 Source language, translation direction

Metadata on source language and translation direction allow to appreciate the reliability of the proposed translation. Mona Baker (1996) described four universals that are prone to surface in translations, representing certain features that distinguish translated texts from the original ones and are the result and product of the translation process. They are (i) explicitation, (ii) simplification (iii) normalisation and (iv) levelling out, and together they usually result in translations being less complex than the source texts or than texts, written originally in the source language in general (Baker 1996:176-7).

(i) *Explicitation* basically refers to the translator's tendency to 'spell things out rather than leave them implicit in translation' (Baker 1996:180). Explicitations can be syntactical or lexical. The former may result in a translated text having more conjunctions than the original. The latter refers to additional information not present in the original inserted in order to explain the source language information to a target culture reader.

(ii) *Simplification* means that translated texts contain simpler language than the original ones at syntactical and lexical level. The former means that long sentences are often divided into several shorter ones thereby causing changes in punctuation by making it stronger (Malmkjaer 1997). The latter can result in less varied vocabulary or a "lower lexical density", e.g. the translation having more function words or grammatical words than the original. "Lexical words contain more information than grammatical words, and using fewer lexical words means that the reader will have to keep track of less information" (Helgegren 2005:20).

(iii) *Normalisation* is the "tendency to exaggerate features of the target language and to conform to its typical patterns" (Baker 1996: 183). "It results in the translator using many clichés or typical grammatical structures of the target language, often grammaticising elements of texts that are ungrammatical in the source" (Helgegren 2005:21).

(iv) *Levelling out* means that "translated texts steer the middle course between two extremes, converging towards the centre" (Baker 1996:184) The translated texts may show less textual variance and the unusual textual features of the source text will be flattened out in the translation, thus enhancing its conformity to the standard.

2.2.4.2 Translation mode

Translation mode distinguishes between a human translator, translation memory and machine translation. The mode is important for appreciating the translation suggestion offered by the corpus. The less human participation, the more uniform the grammar, the vocabulary and translation suggestions. It is known that even the use of translation memory limits variability as the translator is prone to chose a readily-offered solution rather than to spend time searching for an alternative.

2.2.4.3 Text type, topic, keyword

Text type yields information on the text genre and style, the topic characterises the domain from which the text originates and the keyword reunites similar texts across different areas of expertise. These types of metadata can also give the translator a clue as to whether the text was written for laymen or experts.

2.2.4.4 Language variety

Metadata can help to distinguish between generally accepted usage and a regional variety. This is important for many languages, e.g. American English differs from the British, Australian, and other kinds of English.

The same is true for French, spoken in France, Belgium, Canada, Switzerland and many African countries. Even a much less-widely used language, like Dutch covers three countries: Belgium, the Netherlands and Surinam.

3. **WHAT THE DPC HAS TO OFFER TO A TRANSLATOR**

The Dutch Parallel Corpus (DPC) is a 10-million-word parallel corpus under construction comprising texts in Dutch, English and French with Dutch as a pivotal language. This corpus represents a good example of the parallel corpora discussed above.

3.1 Quality and Directionality

The corpus contains only quality-assured data and is bidirectional (Dutch as a source and a target language). A part of the corpus is trilingual, consisting of Dutch texts translated into English and French. The corpus is balanced: (i) with respect to the amount of data in every language and translation direction (with a minimum of two million words per translation direction) and (ii) with respect to the text type.

3.2 Typology and Taxonomy

Firstly, the existing translated texts were surveyed to make sure there is enough available data for a bidirectional corpus. Then the data was divided according to the data provider: (i) commercial publishers producing informative-recreational texts and (ii) institutions producing only texts for various practical purposes. This division was used to separate the data into two main groups.

Subsequently, each group was divided into several text types but the criteria for this division are not of the same nature. Informative-recreational texts were divided into established genres: fictional literature (imaginative), non-fictional literature (non-imaginative) and journalistic texts. Institutional texts were divided according to their function and purpose: instructive texts, administrative texts and texts meant for external communication.

Since the adopted six text types still turned out to be too broad, a prototype approach, as discussed in 2.2.3, was applied to create a subtypology. The project team opted for a two-level typology. The introduction of subtypes however has no implications for the balancing of the corpus, it is merely a way of mapping the actual landscape within each text type and assigning accurate labels to the data in order to enable the user to correctly select documents. The labels for the subtypes were chosen from cognitively tangible categories that are easily identifiable and recognisable both to the corpus compilers and the corpus users.

Table 2: The DPC typology

[SUPERORDINATE]	[BASIC LEVEL]
1. Fictional literature	1.1 Novels
	1.2 Short stories
2. Non-fictional literature	2.1 Essayistic texts
	2.2 (Auto)biographies
	2.3 Expository non-fictional literature
3. Journalistic texts	3.1 News reporting articles
	3.2 Comment articles (background articles, columns, editorials)
4. Instructive texts	4.1 Manuals
	4.2 Internal legal documents
	4.3 Procedure descriptions
5. Administrative texts	5.1 Legislation
	5.2 Proceedings of parliamentary debates
	5.3 Minutes of meetings
	5.4 Yearly reports
	5.5 Correspondence
	5.6 Official speeches

6. External communication	6.1 (Self-)presentations of organisations, projects, events
	6.2 Informative documents of a general nature
	6.3 Promotion and advertising material
	6.4 Press releases and newsletters
	6.5 Scientific texts

3.3 Metadata

The DPC metadata files (cf. figure 3) contain information concerning the source language, the translation direction, translations in other languages included in the corpus, the intended audience (broad external audience, limited internal audience, specialists), the type of text provider (profit vs non-profit) and domains with keywords indicating the field of human activities, to which a particular document belongs. This information is meant as an additional navigation tool.

3.4 Alignment and annotation

The corpus data will be aligned on the sentence level with manual verification and enriched with annotations such as lemmatization and PoS-tagging. A small part of the corpus will be aligned on a sub-sentential level, syntactically annotated and manually verified at every step of processing.

3.5 Availability

The corpus will be made available for various commercial and non-commercial purposes. Therefore, permission clearances were obtained for every text in the corpus. The corpus will be distributed through the Dutch Language Union, more particularly through the HLT Agency (the original Dutch name is 'TST-centrale', which is a Dutch-Flemish organisation responsible for the management, maintenance and distribution of Dutch digital language resources) .

4. CONCLUSION

The conclusion is that the DPC will possess all the desirable features of a professionally-compiled parallel corpus as discussed in 1.3. (i) It is bi-directional allowing not only for comparison between the source and target language but sometimes between translations into two languages. (ii) The text samples are mostly full texts providing enough lexical and grammatical clues for a plausible translation suggestion. (iii) It contains only data of assured quality making it a reliable source. (iv) It has a diverse typology with texts from different domains including areas for which no of few other resources such as dictionaries exist, such as software (IBM data). (v) The data is complemented with a set of metadata relevant for translation purposes. (vi) It will be aligned on sentence level allowing for prompt retrieval of sentence pairs. (vii) It will be enriched with syntactic annotations allowing for meaning disambiguation and providing the sentence structure. (vii) It will be equipped with a user-friendly web interface allowing for efficient and fine-tuned search. The DPC project team aims at creating not only a useful resource for research but also in a functional tool for translators.

ACKNOWLEDGEMENT

The DPC project is carried out within the STEVIN program, which is funded by the Dutch and Flemish governments. The entire DPC team includes Willy Vandeweghe, Lieve Macken, Lidia Rura (Ghent University College) and Piet Desmet, Hans Paulussen, Maribel Montero Perez (KU Leuven – Campus Kortrijk).

REFERENCES

[1] Aston, G. (1999) "Corpus use and learning to translate", *Textus* 12: 289-314, available online, http://www.sslmit.unibo.it/~guy/textus.htm

[2] Baker, M. (1995) “Corpora in translation studies: An overview and some suggestions for future research”, *Target* 7(2): 223–243.

[3] Baker, M. (1996) “Corpus-based translation studies: The challenges that lie Ahead”, in H.Somers, ed., *Terminology, LSP and Translation*, Amsterdam: Benjamins.

[4] EAGLES. (1996) *Preliminary recommendations on text typology: Expert Advisory Group on Language Engineering Standards*, available online, http://www.ilc.cnr.it/EAGLES96/browse.html

[5] Fludernik, M. (2000) “Genres, Text Types, or Discourse Modes? Narrative Modalities and Generic Categorization”, *Style* 34 (2), Summer 2000: 274-292, available online, http://www.encyclopedia.com/doc/1G1-68279076.html

[6] Göpferich, S. (1995) *Textsorten in Naturwissenschaften und Technik: Pragmatische Typologie – Kontrastierung – Translation*. Tübingen: G. Narr.

[7] Halverson, S. (1998) “Translation studies and representative corpora: establishing links between translation corpora, theoretical/descriptive categories and a conception of the object study”. *META* 43(4): 494-514.

[8] Helgegren, S. (2005) *Tracing Translation Universals and Translator Development by Word Aligning a Harry Potter Corpus.* Diss., available online, http://www.diva-portal.org/diva/getDocument?urn_nbn_se_liu_diva-4579-1__fulltext.pdf

[9] Koskinen, K. (2000) “Institutional Illusions : Translating in the EU Commission”, *The Translator* 6: 49-65.

[10] Lee, D. Y. (2001) “Genres, Registrers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle”. *Language Learning & Technology* 5(3): 37-72.

[11] Longacre, R. E. (1983) *The Grammar of Discourse*. New York: Plenum Press.

[12] Macken, L., Trushkina, J. and Rura, L. (2007) “Dutch Parallel Corpus: MT Corpus and translator's aid.” *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 313-320.

[13] Malmkjær, K. (1997) “Punctuation in Hans Christian Andersen’s stories and their translations into English” in F.Poyatos, ed., *Nonverbal communication and translation : new perspectives and challenges in literature, interpretation and the media*, Benjamins, Amsterdam.

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

[15Paulussen, H., Macken, Trushkina, J. Desmet, P., & Vandeweghe, W. (to appear) “Dutch Parallel Corpus: a multifunctional and multilingual corpus.” *Cahiers de l'Institut de Linguistique de Louvain*, CILL, Louvain-La-Neuve, 32.1-4 [2006 issue], 269-285.

[16] Reiss, K. (1981) "Type, Kind and Individuality of Text. Decision-making in Translation", *Translation Theory and Intercultural Relations special issue of Poetics Today* 2 (4), Even-Zohar, I., and Toury, G.,(Eds), 121-131.

[17] Ungerer, F., & Schmid, H.J. (1996) *An Introduction to Cognitive Linguistics*. London and New York: Longman.