# CORPUS-BASED TRANSLATION RESEARCH: ITS DEVELOPMENT AND IMPLICATIONS FOR GENERAL, LITERARY AND BIBLE TRANSLATION

A. Kruger[1]

## ABSTRACT

Corpus-based translation research emerged in the late 1990s as a new area of research in the discipline of translation studies. It is informed by a specific area of linguistics known as corpus linguistics which involves the analysis of large corpora of authentic running text by means of computer software. Within linguistics, this methodology has revolutionised lexicographic practices and methods of language teaching. In translation studies this kind of research involves using computerised corpora to study translated text, not in terms of its equivalence to source texts, but as a valid object of study in its own right. Corpus-based research in translation is concerned with revealing both the universal and the specific features of translation, through the interplay of theoretical constructs and hypotheses, variety of data, novel descriptive categories and a rigorous, flexible methodology, which can be applied to inductive and deductive research, as well as product- and process-oriented studies. In this article an overview is given of the research that has led to the formation of a new subdiscipline in translation studies, called Corpus-based Translation Studies or CTS. I also demonstrate how CTS tools and techniques can be used for the analysis of general and literary translations and therefore also for Bible translations.

## 1. INTRODUCTION

Corpus-based translation research emerged in the late 1990s as a new area of research in the discipline of translation studies. It is informed by a specific area of linguistics known as corpus linguistics which involves the analysis of large corpora of authentic running text by means of computer software. Within linguistics, this methodology has revolutionised lexicographic practices and methods of language teaching. In translation studies this kind of research involves using computerised corpora to study translation as a variety of language behaviour that merits attention in its own right because it is "shaped by its own goals, pressures and context of production" (Baker 1996a:175).

1 Dr. A. Kruger, Department of Linguistics (Translation Studies), University of South Africa, UNISA 0003, South Africa. E-mail: krugea@unisa.ac.za

In this article an overview is given of the research that has led to the formation of Corpus-based Translation Studies or CTS. I also demonstrate how CTS tools and techniques can be used for the analysis of general and literary translations and therefore also for Bible translations.

In order for me to home in on current corpus-based translation research, some remarks on corpus linguistics and the descriptive approach to translation are needed. Such remarks will be brief and of necessity selective as my only purpose will be to highlight the provenance of ideas that influenced computerised studies of translation.

## 2. CORPUS LINGUISTICS

The word *corpus* was originally used for any collection of writings by a specific author (Baker 1995:225). Nowadays, *corpus* means primarily a collection of texts held in electronic form, capable of being analysed automatically or semi-automatically rather than manually. It includes spoken as well as written texts from a variety of sources, on different topics, by many writers and speakers.

According to Leech (1991:10), the beginning of modern corpus linguistics goes back to the early 1960s when the "first-generation" of one-million word computer-readable corpora were first created. The Brown Corpus designed by Francis and Kučera in 1961-1964 was the first and includes samples of written American English. The Lancaster-Oslo-Bergen Corpus contains samples of written British English and was completed in 1978.

Laviosa-Braithwaite (1996b:13ff.) defines the corpus linguistics of the 1980s and 1990s as

> a branch of general linguistics which involves the analysis of large machine-readable corpora of running text, using a variety of software tools designed specifically for textual analysis.

This era of corpus linguistics started with the "second-generation" multi-million word corpora of written and spoken English, namely John Sinclair's Birmingham Collection of English Text and the Longman-Lancaster English Language Corpus, both of which benefited from the availability of the KDEM optical character-recognition device for text inputting.

Any corpus linguistic analysis depends on both the creation of the corpus and the development of software tools to observe, analyse and process it. Owing to the rapid development of text retrieval software, the increased storage and processing power of modern computers, and CD-ROM optical disks it has been possible to create corpora of hundreds of millions of words

such as the British National Corpus (BNC) and the interactive Corpus of Spoken American English.

According to Laviosa-Braithwaite (1996b:14ff.), corpus linguistics is characterised as an independent discipline within general linguistics because it is firmly based on the integration of four elements, namely data, description, theory and methodology. Their interrelationship can be expressed in terms of a continual process involving corpus creation, discovery, hypothesis formation, testing and evaluation. The corpus constitutes the raw linguistic data, assembled and recorded according to specific design criteria, which is examined and processed by means of computerised tools and techniques. The facts discovered about language by this bottom-up approach are accumulated and organised in new descriptions of language behaviour. These feed into linguistic theory, where concepts and language models are created to explain the phenomena empirically observed and hypotheses are put forward for further testing. The new research prompts the expansion of the initial corpus (or the creation of a new one), the refinement of the methodology, the accumulation of more facts and the enhancement of the theory. According to Biber *et al.* (1994:169, 170), there are two major advantages to the use of corpora for linguistic analysis. They provide a large empirical database of natural discourse, so that analyses are based on naturally-occurring structures; and they enable analyses of a scope and reliability not feasible otherwise, allowing researchers to address issues that were previously intractable. In addition, corpus-based analyses demonstrate that, as linguists, we often have strongly held intuitions, but those intuitions frequently prove to be incorrect when they are tested empirically.

Corpus linguistics has had a major impact not only on descriptive linguistics, but also on many branches of applied linguistics (cf. Laviosa-Braithwaite 1996b:20ff.). Corpus-based lexicographic research has had a major impact on compiling dictionaries. For the compilation of the first corpus-based learner's dictionary (the *Collins COBUILD English Language Dictionary*), John Sinclair and his team used concordance lines to observe and record for each word type in the corpus its meaning, collocation, word class, syntactic pattern, register, field of discourse and pragmatic uses. In South Africa, in 1999, the Pan South African Language Board (PANSALB) initiated a project by means of which they will assist the various African languages to set up dictionary units similar to the existing units for Afrikaans and South African English, and help them to become familiar with corpus linguistic tools. Pharos Dictionaries (a division of Nasionale Pers-Boekhandel) in Cape Town, has a main or "balanced" corpus of 100 written texts in contemporary Afrikaans focusing on the core lexicon (3 599 765 words;

comprising 50.27% fiction, 49.27% non-fiction and 0.47% non-categorised), and a text archive, comprising 1 311 written texts (12 966 939 words) (Luther 1998).

Educational linguistics has benefited from corpus-based evidence of language use in the design of pedagogical grammars such as the *Collins COBUILD Basic Grammar* (1995), syllabi, language teaching theory and practice, and computer-assisted language learning (CALL). The corpus-based approach can now also be considered part of mainstream computational linguistics, informing research into a variety of applications, e.g. man-machine interface systems such as speech, handwriting or optical character recognition systems, text-to-speech systems, natural language analysis (e.g. parsers, taggers, automatic alignment of bilingual texts, bilingual concordances, memory-based computer-assisted translation, example-based machine translation). Contrastive interlanguage analysis of an International Corpus of Learner English combines two types of analyses: native English versus English as a foreign language, and the comparison of different English interlanguages (e.g. of French, German, Swedish learners). The aim of this new type of research is to study the over- and under-representation of lexical words, phrases, word categories in learners' language and to identify those features that pertain to interlanguage *per se*, irrespective of the learner's mother tongue (cf. also Granger 1998). Coulthard (in Laviosa-Braithwaite 1996b:20) demonstrates that the norms of general language use, revealed by word frequency statistics, typical collocations and the patterning of lexicogrammatical choices, are crucial in assessing idiosyncratic usage of language in falsified and in authentic texts, or for establishing the authorship of texts claimed to have been produced by different authors.

Corpora are invaluable resources in contemporary linguistics, but without tools and techniques to search, sort, count and display the vast quantities of data they contain, they would be of little practical use. In the next section some of the basic processing techniques that can be used with "raw" corpora are briefly discussed. "Raw" corpora are corpora that are neither tagged nor parsed and may contain minimal annotation indicating structural divisions such as text, paragraph or sentence boundaries (cf. also Sinclair 1991; Biber, Conrad & Reppen 1998; Stubbs 1996). The examples below have been provided by means of *WordSmith Tools*, a Windows-based suite of programmes that offers six tools for the lexical analysis of texts (*WordList*, *Concord*, *Keywords*, *Splitter*, *TextConverter* and *Dual Text Aligner*). *WordSmith Tools* was developed by Mike Scott and is marketed by Oxford University Press (<http://www1.oup.co.uk.elt/catalogu/multimed/ 458946/4589846.html>).

## 2.1 Corpus processing techniques

- One very basic type of calculation that any corpus analysis software should be able to carry out is to measure the lexical variation or diversity in a corpus. For this we need the *type-token ratio* of the words in a text: *types* refer to the number of different words in a text; *tokens* refer to the total number of "running words" in the text. For example, the phrase "to be or not to be" contains six tokens, but only four types ("to" and "be"). The type-token ratio for a given text is calculated by dividing the number of tokens by the number of types. The higher the ratio the more varied the vocabulary, i.e. the implication is that there is little repetition. This ratio is sensitive to text or corpus length. The longer a text, the more likely it is that words will be repeated, thus lowering the ratio. For this reason *WordList* standardises type-token ratios by calculating the ratio for consecutive 1,000 word chunks of text, and then takes an average count at the end; allowing one to compare type-token ratios across texts of differing lengths. According to the statistics in the table below, the type-token ratio for the source text is 40,93, for target text 1 it is 38,46, for target text 2 it is 39,00 and for target text 3 it is 35,81. The results show that the source text has the highest ratio, i.e. that the author of this text used a much wider range of vocabulary than the three translators. The third translation has the lowest type-token ratio, i.e. there is a lot of repetition as regards the vocabulary (cf. Kruger 2000:329).

| | Source Text | Target Text 1 | Target Text 2 | Target Text 3 |
|---|---|---|---|---|
| No of tokens | 20 908 | 20 688 | 20 448 | 18 273 |
| Types | 3 278 | 3 052 | 3 211 | 2 562 |
| Type-token ratio/ 100 words | 15,7 | 14,8 | 15,7 | 14,0 |
| *WordSmith Tools* standardised type-token ratio/1000 words | 40,93 | 38,46 | 39,00 | 35,81 |

- We can also obtain *basic statistics* such as the number of sentences and paragraphs in individual texts or whole corpora, average word and sentence lengths, and how many words there are of each length (one-letter words, two-letter words, etc.). The statistics below are for Shakespeare's *The merchant of Venice*.

| Text file | MERCHANT.TXT |
|---|---|
| Bytes | 112,999 |
| Tokens | 20,908 |
| Types | 3,278 |
| Type/Token Ratio | 15,68 |
| Standardised Type/Token Ratio | 40,93 |
| Average word length | 4,01 |
| Sentences | 932 |
| Sentence length | 22,39 |
| sd Sentence Length | 21,22 |
| Paragraphs | 12 |
| Paragragh length | 664,33 |
| sd Paragraph length | 463,81 |
| 1-letter words | 1,099 |
| 2-letter words | 3,862 |
| 3-letter words | 4,614 |
| 4-letter words | 4,827 |
| 5-letter words | 2,390 |
| 6-letter words | 1,556 |
| 7-letter words | 1,153 |
| 8-letter words | 655 |
| 9-letter words | 420 |
| 10-letter words | 200 |
| 11-letter words | 72 |
| 12-letter words | 14 |
| 13-letter words | 5 |

- *Frequency lists* are word lists arranged alphabetically or in order of frequency, enabling one to compare texts lexically. According to Meunier (1998:19)

> this facility is intended to support stylistic comparison, such as comparisons of several versions or translations of the same story, or of texts on the same topic.

*WordList Tool* was used to generate the alphabetical list of the first 10 words in Shakespeare's *The merchant of Venice*. The second and third columns of the list show the frequency and the percentage if the frequency of a word in a file is > 0.01 per cent (cf. Kruger 2000:294).

MERCHANT.LST WORDLIST

| N | WORD | FREQ. | % |
|---|---|---|---|
| 1 | a | 414 | 1.98 |
| 2 | a-bleeding | 1 | |
| 3 | a-brewing | 1 | |
| 4 | a-cap'ring | 1 | |
| 5 | abate | 1 | |
| 6 | abide | 2 | |
| 7 | abject | 1 | |
| 8 | able | 2 | |
| 9 | aboard | 1 | |
| 10 | about | 9 | 0.04 |

- *Concordancing* is perhaps the most familiar corpus processing technique. A concordance is a listing of all occurrences of a selected item in a text or corpus. These occurrences are conventionally displayed in KWIC (key word in context) format, where the software outputs a series of concordance lines, each displaying a single occurrence of the item, along with the words immediately to its left and right in the text or corpus. The *Concord* tool of *WordSmith Tools* generated the following concordance of *hope* in *The merchant of Venice* (cf. Kruger 2000:298):

HOPE: 12 ENTRIES

| N | CONCORDANCE | | WORD NO. |
|---|---|---|---|
| 1 | worst fall that ever fell, I | **hope** I shall make shift to | 2,266 |
| 2 | me as my father (being I | **hope** an old man) shall | 5,540 |
| 3 | that hazard all Do it in | **hope** of fair advantages | 7,725 |
| 4 | tune now To my heart's | **hope**! — gold, silver and | 8,779 |
| 5 | damn'd, there is but one | **hope** in it that can do you | 14,077 |
| 6 | is but a kind of bastard | **hope** neither. And what | 14,094 |
| 7 | hope neither. and what | **hope** is that I pray thee? | 14,098 |
| 8 | Marry, you may partly | **hope** that your father got | 14,108 |
| 9 | were a kind of bastard | **hope** indeed, so the sins of | 14,128 |
| 10 | bond! How shalt thou | **hope** for mercy, rend'ring | 15,453 |
| 11 | bands Which speed (we | **hope**) the better for our | 19,501 |
| 12 | my lord? Not that (I | **hope**) which you receiv'd | 20,124 |

Growing out of corpus linguistics, corpus-based translation research at the same time marks a turn away from prescriptive approaches to translation to descriptive and cultural approaches that were developed by translation scholars and polysystem theorists (Tymoczko 1998:652). In order to show that the principles underlying the descriptive approach can be applied directly to corpus-based translation research, a brief overview of descriptive translation studies is needed.

## 3. DESCRIPTIVE TRANSLATION STUDIES

Modern translation studies research has moved away from the straitjacket of the prescriptive and normative approaches to translation of the late 1970s. "Equivalence" has long since ceased to be the controlling concept it used to be. Gone are the days when the criterion for measuring a translation was located in the original or source text and the translation was inevitably perceived as a mere substitute — a derivative that had to be checked against the original for shortcomings. For this we have to thank two separate approaches to translation which developed independently and almost simultaneously in the 1980s: the functionalist approach and the descriptive approach. The functionalist approach advocates that the function of a translation does not have to be the same as that of the original because translation is "a new communicative act that must be purposeful with respect to the translator's client and readership" (Nord 1997, back cover). (For a detailed discussion of the functionalist approach, see Nord 1997; also Naudé in this volume).

Descriptive translation scholars advocate a descriptive, target-oriented, functional and systemic approach to translation (see Hermans 1985; 1999 and Toury 1995). For descriptive scholars, any text is a translation if it functions as such in the receiving cultural and literary system. Such a view of necessity involves a radically different view of equivalence — even more radical than that of the functionalists. In fact, this view allows Toury (1980; 1995) to dissolve the concept of equivalence: if text A is a translation of text B, then it can be assumed that the relation between them is one of equivalence. In other words, equivalence is merely the name given to the "translational" relation that exists between the two texts, there are no longer any absolute criteria for equivalence. The consequence of this reversal of perspectives is that the researcher no longer has to ask: Do we have a sufficient degree of equivalence (of what kind? at what level?) to call this text a translation? Instead the questions now are: What type of translation relation do we have, and why this type rather than another? (Hermans 1991:158). The answers to these questions, Toury (1980:56) contends, have to do with

norms and conventions reigning in a particular target culture at a particular time. In this way, the concept of norms replaces that of equivalence as the researcher's focus of attention. This is the reason why prescriptive models of translation have been replaced by models that are generally descriptive, historical and socio-cultural (see Lambert & Van Gorp 1985:42-53, Van den Broeck 1985:54-62, Toury 1980:112-121, Heylen 1993:1-25, Van Leuven-Zwart 1989, 1990; for the Göttingen group see Kittel & Frank 1991 and Kittel 1992).

Descriptive translation models can be used to describe real translations and to account for their observed features with reference to the literary, cultural and historical contexts in which they were produced. In other words, they describe (i.e. explain) the specific characteristics of a translated text (or multiple translations of the same original) in terms of constraints or norms reigning in the target culture at a particular time that may have influenced the method of translating and the ensuing product. As Hermans (1999:39) aptly puts it, as a rule, a target-oriented approach is

> a way of asking questions about translations without reducing them to vicarious objects explicable entirely in terms of their derivation.

This "target-orientedness" of DTS remains one of its strongest features and was legitimised in part with reference to Even-Zohar's polysystem theory. Drawing on Russian Formalism, Even-Zohar (1990) argued that literature (including the body of translated literature) is a heterogeneous conglomerate of individual literary systems which are in a constant state of flux. Translated literature, like any literary system, is not inherently peripheral or conservative, but can become central or peripheral, primary or secondary etc. depending on the state of the system.

Polysystem theory has been important in the recent development of translation studies, and corpus-based translation studies, for a number of reasons (cf. Kenny 2001:49ff). In the first place, it reinstates translated literature as a system worthy of study in its own right. Secondly, it ascribes a certain specificity to translated texts that warrants their investigation as a coherent body of texts or corpus. Thirdly, given that translated literature functions as a system in the target culture, it validates the study of such a corpus against the backdrop of non-translated literature in the same language.

The descriptive approach, however, does not only apply to literary translation. As pointed out in the article by Kruger and Wallmach (1997), which is still the only South African synthesis of theoretical and analytical research frameworks within the scope of DTS, all types of translated texts

can be studied with the purpose of finding out how they have been translated within a specific culture and historical period.

Corpus-based translation research builds upon the studies of scholars working within DTS as well as those of scholars who have worked with corpora that have been manually assembled, examined, and analysed. Already in 1993, Mona Baker (1993:243) predicted that the compilation of various types of corpora of both original and translated texts, together with the development of a corpus-driven methodology, would enable translation scholars to uncover "the nature of translated text as a mediated communicative event" through the investigation of what she then termed "universals" of translation, i.e. linguistic features that occur in translated texts and which are not influenced by the specific language pairs involved in the translation process.

According to Tymoczko (1998:653, 657), corpus-based translation research focuses on both the process of translation and the products of translation, and it takes into account the smallest details of the translated texts as well as the largest cultural patterns both internal and external to the texts. Some of the common themes and commitments between translation studies and corpus-based translation research are, for example, the growing commitment to integrate linguistic approaches and cultural-studies approaches to translation, an awareness of the role of ideology as it affects text, context, translation and the theory, practice and pedagogy of translation, adapting modern technologies to the discipline's needs and purposes. Not only are we now able to study and capture recurrent features ("universals") of translation on a large scale (the norm), and consequently understand translation as a phenomenon in its own right, but we are also able to study creative and individual examples (the exception) (cf. Baker 1996a:179; also Baker 1998). Laviosa-Braithwaite (1996b:47) puts it as follows:

> The corpus-based approach in translation studies emerges as a composite, rich and coherent paradigm, covering many different aspects of the translational phenomenon and concerned with unveiling both the universal and the specific features of translation, through the interplay of theoretical constructs and hypotheses, variety of data, novel descriptive categories and a rigorous, flexible methodology, which can be applied to inductive and deductive research, as well as product- and process-oriented studies.

So, where did it all start? What is corpus-based translation research all about?

# 4. CORPUS-BASED TRANSLATION RESEARCH

## 4.1 The notion of the "third code"

When compared to specific source texts, and to original writing in general, certain features seem to appear only in translated texts, giving them a unique character. One of the earliest references to the idea that the language of translation is distinct from ordinary language can be found in Frawley (1984:168). William Frawley maintains that the confrontation between source language and target language during the translation process results in creating a "third code". In other words, the code (or language) that evolves during translation and in which the target text is expressed is unique. It is a kind of compromise between the norms or patterns of the source language and those of the target language. Concrete examples abound of, for example, borrowings which cause "foreign" lexical patterning in translated texts, i.e. patterning that would not normally occur in the source language nor in the target language.

The notion of the third code provides a useful starting point for explaining some of the concerns of translation scholars who are attempting to apply the techniques of corpus linguistics to investigating the language of translation. Baker (1998) points out that this unique language is not so-called "translationese", a pejorative term that is used when an

> unusual distribution of features is clearly the result of the translator's inexperience or lack of competence in the target language;

on the contrary,

> translation results in the creation of a third code because it is a unique form of communication, not because it is a faulty, deviant or sub-standard form of communication (Baker 1993:248).

The concept of translation as a kind of separate sub-language is therefore not new, "what is new is the non-evaluative view within descriptive translation studies of interlanguage as an *inevitable* aspect of translation" (Øverås 1998:573; her emphasis; cf. also Halverson 1998). Baker (1993: 248) insists that translated texts record "genuine communicative events and in this sense they are different from other communicative events in any language". The nature of this difference, however, needs to be explored and recorded. And in this sense corpus linguistic techniques and tools can be applied in translation studies. First, a brief overview of the early research that led to the observation of these features as being translation "universals" is needed.

## 4.2 Universal features of translation: early research

Based on small-scale studies and casual observation, in the late 1980s and early 1990s, various translation scholars have noted features in translated texts

> which typically occur in translated text rather than original utterances and are not the result of interference from specific linguistic systems (Baker 1993:243).

These translation-specific, rather than language- or culture-specific features were first categorised by Baker (1993:243-247) as universal features of translation as follows:

(i) explicitation, in the form of shifts in cohesion (Blum-Kulka 1986); insertion of additional information in the target text (Baker 1992)

(ii) disambiguation and simplification (Vanderauwera in Baker 1993:243-247)

(iii) textual conventionality in translated novels (Vanderauwera in Baker 1993:243-247); and interpreting (Shlesinger in Baker 1993:243-247)

(iv) a tendency to avoid repetition present in the source text (Shlesinger in Baker 1993:243-247; Toury in Baker 1993:243-247)

(v) a tendency to exaggerate features of the target language (Toury in Baker 1993:243-247; Vanderauwera in Baker 1993:243-247)

(vi) specific distribution of lexical items in translated texts vis-à-vis source texts and original texts in the target language (Shamaa in Baker 1993:243-247).

According to Baker (1993:246), universal features such as these can be seen as

> a product of constraints which are inherent in the translation process itself, and this accounts for the fact that they are universal.

However, pending further research, "they do not vary across cultures".

A brief overview of the early research that led to the observation of these features as being translation "universals" enables me to group some of these features together (cf. Kruger 2000:137) and that leaves us with three distinct categories of "universals", namely a tendency towards explicitation (no. i and v above), a tendency towards disambiguation (no. ii and iv above), and a tendency towards conventionalisation (no. iii and vi above). In my opinion, all the early corpus-based translation research can be incorporated under these three headings.

### 4.2.1 A tendency towards explicitation and addition

Blum-Kulka (1986:19) noticed early on that shifts in the types of cohesion markers used in translated texts raised the level of explicitness in such texts and that such explicitness is "inherent in the process of translation". She subsequently formulated the "explicitation hypothesis" which postulates

> an observed cohesive explicitness from SL [source language] to TL [target language] texts regardless of the increase traceable to differences between the two linguistic and textual systems involved.

Blum-Kulka (1986:21) furthermore shows that explicitation results from "building into [the target text] a semantic redundancy absent in the original". She regards this greater redundancy as the result of the way in which translators interpret source texts and reports that the work of both non-professional and professional translators reveals over-representation of lexical cohesion or repetition (with a non-committant under-representation of reference linkage, e.g. pronominalisation) and expansion of the text. In line with Blum-Kulka's (1986) observations, in her coursebook, Baker (1992) discusses various examples where the translator inserts additional information in the target text in order to fill in a cultural gap.

Toury (1980:130) found that

> binomials composed of synonyms or near-synonyms, which are a common feature of Hebrew writing, tend to occur more frequently in translated than in original Hebrew texts and to replace non-binomials in source texts.

Toury (in Baker 1993:244) accepts that explicitation is "a feature of all kinds of mediated events, including interaction in a foreign language", but wonders whether there are

> any differences in the level and nature of explicitation by, for instance, language learners vs. translators, professional vs. non-professional translators, or in oral vs. written translation.

In 1985, Vanderauwera (in Laviosa-Braithwaite 1996a:154) found in her corpus of 50 English translations of Dutch novels that implicit information was made explicit and more precise, and that ambiguous pronouns were replaced by precise forms of identification. She suggests that translations over-represent or exaggerate features of their host environment in order to make up for the fact that they were not originally meant to function in that environment.

### 4.2.2 A tendency towards disambiguation and simplification

Laviosa-Braithwaite (1997:533) reports that evidence of at least three types of disambiguation or simplification have been found in translated texts, namely lexical, syntactic and stylistic. Reporting on research conducted already in 1983, Blum-Kulka and Levenston (in Laviosa-Braithwaite 1997: 533) found that *lexical simplification*, that is the process and/or result of making do with fewer words, operates according to six microtextual principles (i.e. translation strategies at word level) to deal with various types of non-equivalence. These translation strategies are:

- the use of superordinate terms when equivalent hyponyms are lacking in the target language;
- an approximation of the concept expressed in the original text;
- use of "common" or "familiar" synonyms;
- transfer of all the functions of a source language word to its target language equivalent;
- use of circumlocutions instead of conceptually matching high-level words or expressions — especially with theological, culture-specific or technical terms;
- use of paraphrase where there are cultural gaps between source language and target language.

Baker (1992:26) also provides evidence of the use of superordinates when there are no corresponding hyponyms in the target language and claims that

> this is one of the commonest strategies for dealing with many types of non-equivalence, particularly in the area of propositional meaning.

In 1985, Vanderauwera (in Baker 1993:247) found that foreign words and dialogue which occur in original texts are either replaced in their translations by target language items, or that they are glossed. Based on a study of a limited corpus of translations of modern, non-literary English texts in a variety of languages, Baker (1992:36) suggests that Japanese seems far more tolerant of the use of loan words in translation than, for instance, Arabic and French. In a corpus of 425 mystery books translated into Hebrew since the early 1960s, Toury (1980:104) identifies a strong norm which he expresses as "the title should never be too complex, witty or sophisticated". This norm manifests itself in two ways. First, sophisticated titles (in terms of lexis) are replaced by simple titles which contain one of a stock of items that include the Hebrew equivalents of "mystery", "murder", "blood",

"death" and so on. Thus, *The case of the ice-cold hands* becomes "The mystery of the murder in the motel".

As regards *syntactic simplification*, Vanderauwera (in Laviosa-Braithwaite 1996b:116) found in 1985 that complex syntax was made easier by replacing non-finite clauses with finite ones and by suppressing suspended periods, potentially ambiguous pronouns are also replaced by forms which allow more precise identification. Vanderauwera (in Laviosa-Braithwaite 1996b:116) also provides evidence for various forms of *stylistic simplification* such as the breaking up of long sentences, replacing elaborate phrases with shorter collocations, reducing or omitting repetitions and redundant information, shortening overlong circumlocutions, and omitting phrases and words. Baker (1996a:182) reports that punctuation was changed in translations to make for easier reading. For example, Malmkjær (in Baker 1996a:182) found that the English translators of Hans Christian Andersen consistently simplified his "unusual" Danish punctuation by turning commas into semicolons or periods and semicolons into periods, thus breaking up long sentences into shorter sections to promote processing ease. Similarly, the Russian and French translators of Virginia Woolf adjusted her "unusual" punctuation in order to make the texts easier to read (May in Baker 1996a:192). According to Baker (1993:244), both Shlesinger and Toury report evidence of a tendency to avoid repetitions occurring in source texts, either by omitting them or rewording them, thereby obtaining a more transparent and fluent style in the translations (cf. also Shlesinger 1998). Toury (in Baker 1993:244) regards this feature as "one of the most persistent, unbending norms in translation in all languages studied so far".

### 4.2.3 A tendency towards conventionalisation and normalisation

According to Baker (1993:244), Shlesinger found that in interpreting the tendency towards conventionalisation manifests itself in an overriding tendency to round off speakers' unfinished sentences, "grammaticise" ungrammatical utterances and omit such things as false starts and self-corrections, even those which are clearly intentional in a courtroom context. Vanderauwera (in Laviosa-Braithwaite 1996a:155) records a similar trend in translation towards "general textual conventionality" as opposed to "textual creativity" in the source texts, which she ascribes to translators' attempts "not to strain the possibilities of target usage", and to the secondary position that translated literature in general, and translations of minority cultures in particular, occupy in target language literary polysystems.

Toury (1980:129) reports that he also found a high level of dependence on a repertory of fixed collocations derived from canonised religious texts

in the corpus of 425 mystery books translated into Hebrew since the early 1960s. The same corpus also reveals that special importance is attached to direct speech: pieces of dialogue are regularly turned into independent paragraphs, indirect speech is replaced by direct speech, and phrases which indicated a move from narration to dialogue are omitted.

In my opinion Baker's (1993:245) sixth "universal" concerning a specific type of distribution of certain features in translated texts vis-à-vis source texts and original writing in the target language, can also be classified as conventionalism. According to Shamaa (in Baker 1993:244)

> common words such as *say* and *day* occur with a significantly higher frequency in English texts translated from Arabic than they do in original English texts. At the same time, their frequency of occurrence in the translated English texts is still considerably lower than the frequency of the equivalent Arabic items in the source texts [...] and leaves a vague impression of being culturally exotic.

This then concludes an overview of early corpus-based translation research, revealing that translated texts tend to display certain features or so-called "universals". In the following section, an overview is given of recent research conducted into the distinctive features of translated texts *per se* with the aid of corpus linguistic tools.

## 4.3 "Universal" features of translation: recent research

Recent and current studies in corpus-based translation research have been shaped by an article published by Mona Baker in 1996 (Baker 1996a). In this article she discusses three fundamental aspects of corpus-based translation studies: its theoretical links with target-oriented approaches (advocated by the DTS theorists), the unique methodology it employs, and the potential of this methodology for investigating "the distinctive nature of translation as a communicative event, shaped by its own goals, pressures and context of production" (Baker 1996a:175).

As regards theoretical assumptions, corpus-based studies recognise that

> a translation, like any kind of text production, develops in response to the pressures of its own immediate context and draws on a distinct repertoire of textual patterns (Baker 1996a:175).

According to Laviosa-Braithwaite (1996b:40), such studies represent a further development of the general trend towards greater and greater autonomy of the translated text vis-à-vis the source text. In terms of methodology, Baker (1996a:178) emphasises the importance of elaborating corpus design criteria and hypotheses which are specific to the needs of descriptive

research in translation studies. Baker (1996a:178-180) argues that the specificity of translational text production, expressed in the literature in terms of the so-called "universals" of translation, can be fruitfully investigated, provided that at least two conditions are met. The first is the elaboration of explicit criteria and procedures for the selection, acquisition and annotation of the texts to be included in the corpus. The second is the precise definition of the linguistic features which are considered concrete manifestations of the "universals" of translation such as simplification and explicitation in order to render these global and abstract constructs operational and verifiable.

One issue that deserves brief attention at this point and which links up with the first condition mentioned above by Baker (1996a), is the types of corpora and the accompanying terminology specifically used in corpus-based translation research.

### 4.3.1 Types of corpora

Elaborate typologies of corpora in use in translation studies have been established by Baker (1995) and Laviosa (1997, 2001; cf. also Ulrych 1997). Central to such typologies are three basic questions (Kenny 2001:57):

- How many languages are represented in the corpus in question?
- In the case of monolingual corpora, do all texts originate in the language of the corpus, or are some, or all of them, translations?
- In the case of bilingual and multilingual corpora, is there a relationship of translation between the different language sections of the corpus?

*Monolingual corpora* such as the British National Corpus (BNC) consist of texts in one language and is useful in translation pedagogy to reinforce students' knowledge of normal target language patterns and to improve translation quality (Bowker 1998, Pearson 1999); as an aid in translation quality assessment (Bowker 1999); and in terminology extraction (Pearson 1998). The BNC and the Cobuild Bank of English which consist of 100 million and 200 million words respectively, can also be used as "controls" in descriptive studies of translation, allowing patterns observed in a source or target text to be set off against what is known about the language in general (Munday 1998; Kenny 2001).

*Multilingual* or *bilingual corpora* refer to

> sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria (Baker 1995:232),

e.g. newspaper articles covering a particular period in original English and original Spanish. In other words, this type of corpus does not necessarily contain texts related to each other through translation, rather their component texts may be comparable on the basis of similarity of their content, domain and communicative function (Zanettin 1998:617). Some of the advantages of this type of corpus from the translation researcher's point of view is that no alignment software is needed and that one deals with authentic texts in a natural environment (e.g. *day* and *jour* patterns: do they always behave the same?). These corpora are mainly used in contrastive linguistics and lexicography, but their use is limited in descriptive translation studies because the design criteria for these corpora usually do not include information about authors, translators or sources.

*Multilingual* or *bilingual parallel corpora* consist of original, source-language texts in language A and their translated version in language B, e.g. the Canadian Hansards or the German-English Parallel Corpus of Literary Texts (GEPCOLT) compiled by Dorothy Kenny under the supervision of Mona Baker at UMIST in co-operation with Dublin City University. Most parallel corpora are bilingual and allow the translation researcher to focus on a specific language pair, but this type of corpus can contain translations into several target languages of the same source-language texts, in this case such a corpus is called a multilingual parallel corpus. According to Malmkjaer (1998), amongst other things, parallel corpora can reveal characteristics of translated texts, such as tendencies towards explicitness and avoidance of repetition. Some of the disadvantages of using this type of corpus from the translation researcher's point of view are that copyright permission is needed for both the source text and its translation(s) (i.e. if one cannot get copyright permission for a particular translation the source text can also not be used), and that alignment software is needed to provide links between words or sentences.

A *monolingual comparable corpus* consists of two single monolingual corpora (i.e. two separate collections of texts in the same language), one *non-translational corpus*, which comprises original texts in the language in question and one *translational corpus*, consisting of translations in that language from a given source language or languages. One of the key features of a comparable corpus is that the two collections of texts cover a similar domain, variety of language, time span and length and are representative in terms of the range of original authors and of translators. This is

> to ensure that any linguistic differences found between them can be reliably attributed to their different status as translation vs. non-translation, rather than to confounding variables (Laviosa 1997:290).

Apart from the fact that the researcher does not have access to original texts and that it may be difficult to identify some translations, other disadvantages of a comparable corpus is the problem of comparability (e.g. only male or female translators), and the fact that to an extent the methodology of such corpora is still under-developed. Another possibility is simply to examine the features of a *single translational corpus*, in order to study translated text as a variety of language in its own right, rather than comparing a translational corpus to a non-translational corpus as part of a larger *comparable corpus*. According to Kenny (2001:58)

> such studies are motivated by a belief in the specificity of translation, by a conviction that there are features that occur in translated text but not in original text (or at least not to the same extent), and that can be explained not with reference to "interference" from a source language, but rather in terms of the nature and pressures of the translation process itself.

This type of corpus supports theoretical studies rather than merely pedagogical applications and is therefore helping translation studies to develop as an independent discipline.

Sara Laviosa (Laviosa-Braithwaite 1996b; Laviosa 1997; 1998a; 1998b; 2001) designed and compiled the English Comparable Corpus (ECC), the first of its kind, in 1996 at the Centre for Translation Studies, UMIST (University of Manchester Institute of Science and Technology), under the supervision of Mona Baker. The ECC is a computerised corpus made up of two separate collections of texts in English. One collection (the Translational English Corpus or TEC), contains texts translated into English from a variety of source languages, the other (the Non-Translational English corpus or NON-TEC) includes texts originally produced in English which are comparable to the TEC for text genre, time of publication (1983-1993), distribution of female and male authors, distribution of single and team authorship, overall size, and target-audience age, gender and level.

The best known single translational corpus is the Translational English Corpus (TEC), mentioned above, that was first started by Sara Laviosa. It is a monolingual corpus consisting of 6.6 million words (at the time of writing) of English texts translated from both European and non-European languages. It consists of four subcorpora: fiction, biography, newspaper articles and inflight magazines. The corpus can be accessed freely via the web, using a suite of basic software tools provided on the relevant site (http://tec.

ccl.umist.ac.uk/tec/). According to Baker (1999:284), researchers are able to select

- only texts translated from a specific source language (i.e. the influence of a particular source language on the patterning of translated English can be examined);
- the texts translated by one or several distinguished British and American literary translators;
- any one or more of the subcorpora, e.g. fiction only, inflight magazines only, etc.

In the following section, an overview is given of recent research conducted into the distinctive features of translated texts *per se* with the aid of corpus linguistic tools.

### 4.3.2 Explicitation as a universal of translation

Baker (1996a:180) states that she takes explicitation to mean that "there is an overall tendency to spell things out rather than leave them implicit in translation". The evidence for this tendency is found in the fact that translations are usually longer than their originals, irrespective of the languages concerned. Lexically the tendency to make things explicit in translation may be expressed through the use or overuse of "explanatory vocabulary" and conjunctions (Baker 1996a:181) that are added to the target text. Addition is listed as one of the five general transformation categories used already centuries ago by the ancient rhetoricians (i.e. substitution, repetition, deletion, addition and permutation; cf. Delabastita 1993:33-39). According to Delabastita (1993:36), addition as translation strategy (i.e. the insertion of information in the translation that is absent in the original text) can partly be ascribed to translators'

> understandable concern for clarity and coherence, which prompts them to disentangle complicated passages, provide missing links, lay bare unspoken assumptions, and generally give the text a fuller wording.

In other cases,

> the additions are due to conscious, intentional interventions of the translator, who may believe, for example, that s/he can enhance the aesthetic qualities of his/her translation by adding rhyme to an unrhymed ST [source text], by using a more strongly metaphorical language, by adding to the exotic flavour of the text, and so forth.

There are various reasons why these items are inserted in the translation. The complex structure of the original text may be an important constraint. Delabastita (1993:37) comments that addition and deletion often go hand in hand, especially if the translator wants to retain the macrostructural properties of the source text or the same volume of text. Very often, however, additions will be found that have to be explained by other principles. For instance, it is well known that translators show a tendency to expand the translation. This is partly due to their understandable concern with clarity and coherence, which prompts them to "explain" complicated passages, provide missing links, lay bare implicit meanings and generally elaborate on the original. In other cases, additions are due to intentional interventions of the translator, who may believe that s/he can enhance the aesthetic qualities of the translation by adding rhyme, or stronger metaphorical constructions, for instance.

Following Blum-Kulka's research in 1986, Linn Øverås (1998) investigated explicitation, expressed in terms of a rise in the level of cohesion, in an English-Norwegian parallel corpus. She reports shifts in conjunctive and reference cohesion through addition and expansion in the specification of nouns by way of e.g. determiners, or substitution (i.e. the replacement of one grammatical device by another), and shifts in lexical cohesion through addition and lexical specification (e.g. source text items being replaced by more specific target language items). According to Laviosa (2001:64),

> these findings therefore confirm the explicitation hypothesis and reveal an increase in the number of grammatical and lexical devices in both translated English and translated Norwegian.

Laviosa praises Øverås by stating that her ability to unravel the differing elements bearing on the phenomenon of explicitation is no mean feat — what Øverås shows is the possibility, thanks to the availability of corpus data, to analyse in greater depth than ever before and from a socio-cultural perspective, the compositeness and complexity of notions such as those of "universals".

Maeve Olohan and Mona Baker (Olohan & Baker 2000) provide evidence of syntactic explicitation in translational English through an analysis of patterns of inclusion and omission of the optional *that* in reported speech. The authors discover a striking preference for the use of *that* with the various forms of the reporting verbs *say* and *tell* in translated versus non-translated English. On the basis of a comparison of concordance data from the BNC (British National Corpus) and the TEC (Translational English Corpus), the quantitative results show that the *that*-connective is far more

frequent in TEC than in BNC, and conversely, that the *zero*-connective is more frequent for all forms of both verbs in the BNC than in TEC. These results provide strong evidence for syntactic explicitation in translated English, which, unlike the addition of explanatory information used to fill in knowledge gaps between source text and target text readers, is hypothesised to be a subliminal phenomenon inherent in the translation process.

### 4.3.3 Simplification as a universal of translation

So far, more evidence of tendencies towards disambiguation and simplification has been recorded than tendencies towards explicitation. Baker (1996a:181,182) defines simplification tentatively as "the tendency to simplify the language used in translation", in other words, the translator attempts to make things "easier for the reader (but not necessarily more explicit)". Toury (1995:270) states that if the target text has a lower information load than the source text it is because ambiguous information in the original has been disambiguated (spelled out or made simpler), in the translation process. To the list of translation strategies that Blum-Kulka and Levenston (in Laviosa-Braithwaite 1997:533) put forward and which was mentioned above, should be added deletion or omission ("pruning or trimming" of the original — Delabastita 1993:35). Baker (1992:40) states that translators can and often simply omit translating a word or expression if the meaning conveyed by such a word or expression "is not vital enough to the development of the text to justify distracting the reader with lengthy explanations". Omitting words, phrases, sentences or sections of the original text is the most direct way of simplifying a translation.

The most significant research into simplification as translation universal to date was conducted by Sara Laviosa (Laviosa-Braithwaite 1996a; 1996b; 1997; Laviosa 1998a; 1998b; 2001) who cautions that early evidence supporting simplification in translation is patchy, not always coherent and cannot easily be compared because the studies have been carried out for different purposes, have asked different types of questions and have made use of different sets of data (Laviosa-Braithwaite 1996b:534). As shown above, these early analyses have all been carried out manually on parallel corpora. Strategies have been limited to particular language combinations and consequently plausible suggestions as to whether simplification can be considered the result of the confrontation of two languages or a phenomenon linked to the nature of the translation process itself have been prevented. The object of analysis consisted mainly of shifts that occur during the translation process at sentence level and the impact of simplification strategies over entire texts has also not been directly assessed. In contrast, recent research

into simplification as a hypothesised universal of translation attempts to provide a consistent methodology for investigations of this kind.

An investigation of translated versus non-translated newspaper articles in the ECC (Laviosa-Braithwaite 1996b:116-118; 1997:538) revealed *inter alia* that, in the British newspapers *The Guardian* and *The European*, the translated articles have a relatively lower proportion of lexical or content words versus grammatical words (i.e. the range of vocabulary used is more limited than in the original articles), independently of the source language, as well as a higher proportion of frequent words versus less frequent words. Moreover, the 108 most frequent words (or list head) are repeated more often, the nucleus of the words most frequently used is less varied, and the average sentence length is lower. In both newspapers, the translations use the present tense of the auxiliary verbs *to be* and *to have* more frequently. A more recent investigation of the translated versus non-translated narrative prose section of the ECC (Laviosa 1998a; 1998b:565) revealed four "core patterns of lexical use":

- Translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower);
- The proportion of high frequency words versus low frequency words is relatively higher in translated texts;
- The list head of a corpus of translated texts accounts for a larger area of the corpus (i.e. the most frequent words are repeated more often);
- The list head of translated texts contain fewer lemmas.

Laviosa (1998b:565) cautions that an ECC-based methodology cannot tell us why certain patterns occur and how they come about:

> the corpus design and methods of analysis adopted in this research focus on the character of the final product of translation, rather than the processes underlying it.

She proposes that in future studies these core patterns be used as sources of hypotheses to test on a variety of translational text genres and also interpreted texts. If these hypotheses were confirmed,

> we would be in a position to suggest quite strongly that translation is a type of linguistic behaviour characterised by distinctive patterns of simplification that set it clearly apart from original text production (Laviosa-Braithwaite 1997:539).

In South Africa, Afrikaans and African language translators in particular have to take great care to "reformulate" rather than just translate health

brochures on AIDS, TB, Hepatitis B and so forth, since these brochures must be made as accessible as possible for semi-literate readers. Two doctoral students (Victor Ndlovu and Theo Rodrigues) are currently conducting research in this area under my supervision. Another interesting doctoral study currently being undertaken by Koliswa Moropa, also under my supervision, involves a corpus-based investigation into simplification and explicitation in translated Xhosa government texts. Moropa's objective is to establish an understanding of these "universal features" as they pertain to Xhosa, in the context of the necessity for Xhosa translators to develop translation and term-formation strategies to overcome the lack of standardised terminology in this language.

### 4.3.4 Normalisation as a universal of translation

Baker (1996a:183) regards normalisation or conservatism, that is "a tendency to exaggerate features of the target language and to conform to its typical patterns", as a third feature of translated texts. She claims that this tendency is quite possibly influenced by the status of the source text and the source language, so that the higher the status of the source text and language, the less the tendency to normalise. Normalisation is most evident in the use of typical grammatical structures, punctuation and collocational patterns.

Dorothy Kenny's (1998; 1999; 2000; 2001) research in this regard has a dual focus: to investigate how unusual and marked compounds and collocations in German literary source texts are rendered in English translation in order to assess whether they are retained or neutralised (i.e. normalised) by means of more habitual ones, and to investigate "sanitisation" in translated texts through the analysis of semantic prosody ("a consistent aura of meaning with which a form is imbued by its collocates" — Louw in Kenny 1998:520). Novel word forms are first of all identified in the German corpus of GEPCOLT (German-English Corpus of Literary Texts) through the retrieval and analysis of potentially creative *hapax legomena* (word forms that occur only once in the corpus). Their creative use is then verified by consulting lexicographical sources, native speakers and the German Mannheim Corpora. With the aid of a bilingual concordancer, the corresponding English translations of these creative lexical items are isolated and their creativity assessed using the information provided by dictionaries, native speaker judgements and the British National Corpus (BNC). Her research suggests that

> certain translators may be more inclined to normalise than others, and that normalisation may apply in particular to source text lexi-

> cal features that draw on the more systematic processes of word formation in German — derivation and conversion to verbal nouns - and creative compounds and collocations that represent exploitations of more habitual lexical combinations (Kenny 2001:211).

The next section deals with other applications of corpus-based translation research that might be of interest to Bible translators.

# 5. OTHER APPLICATIONS OF CORPUS-BASED TRANSLATION RESEARCH

Unlike the previous studies which involved some form of comparative analysis between two or more corpora of different types, the following studies demonstrate that corpus linguistic tools can be borrowed from another discipline and shaped to execute a specific task; also, that translation research can be based on the examination of a single resource, namely a monolingual corpus of translational English, the TEC.

## 5.1 Analysing linguistic features signalling involvement in drama translation

On the assumption that different registers of translated drama have different functions and that they therefore present information differently, in Kruger (2000) my aim was to examine whether the Afrikaans "stage translation" (cf. Kruger 1998) of *The merchant of Venice* reveals more spoken language features signalling involvement and interaction between the characters than a "page translation". I therefore required an analytical tool that would not only enable me to quantify linguistic features of involvement in four Shakespeare texts (the original and three translations, totalling 80 317 words), but also provide a "norm" of the occurrence of such features in authentic spoken English.

Douglas Biber's (1988; 1994; 1995a; 1995b; 1996) multi-dimensional approach to register variation was adapted to suit my purposes. Methodologically, his approach uses huge computerised text corpora, computational tools and multivariate statistical techniques to analyse the linguistic characteristics of spoken and written registers in English. In terms of its situational characteristics, Biber (1988:37) found that typical speech is interactive and involved, and dependent on shared space, time and background knowledge; typical writing has the opposite characteristics, e.g. being "informational". In terms of its linguistic characteristics, typical speech is structurally simple, fragmented, concrete, and dependent on exophoric re-

ference; again, typical writing has the opposite characteristics. Biber (1988:21; 43) also found that linguistic variation must be analysed in terms of sets of co-occurring dimensions because they work together to mark some common underlying function. This is why he calls his approach "multi-dimensional". Each dimension comprises an independent group of co-occurring linguistic features, and each co-occurrence pattern can be interpreted in functional terms such as "involved", "informational", etc.

The following twelve features were analysed from my corpus: private verbs, contractions, second person pronouns, analytic negation, demonstratives, emphatics, first person pronouns, causative subordination, discourse particles, amplifiers and questions. The reason for doing so is because all of them can be characterised as verbal, interactional, affective, fragmented, reduced in form and generalised in content (Biber 1988:104) — exactly the kind of features that are found in the dramatic dialogue of drama texts. The overall finding was that the stage translation by Potgieter (1991) displayed more involvement than the page translation by Malherbe (1949), to a statistically highly significant extent. The features analysed cluster together sufficiently to reveal that in comparison with an older page translation, a recent stage translation displays a definite tendency towards a more oral, more involved and more situated style, reflecting no doubt a general modern trend towards creating more appropriate and accessible texts. The dialogue in a Shakespearean stage translation is more speakable than that of a page translation precisely because it comprises more spoken language features.

Researchers into Bible translation might well be able to do a similar type of study, comparing source text-oriented Bible translations to target-oriented Bible translations. For instance, do translations such as the 1983 *Nuwe Afrikaanse Vertaling* (New Afrikaans Translation) or the 1986 New Testament and Psalms translation into Zulu, which aim at making the Bible more accessible to the modern reader, really fulfil this aim in comparison to previous, more literal translations?

## 5.2 Analysing the style of professional literary translators

In Baker (2000) the question is asked whether individual literary translators can plausibly be assumed to use distinctive styles of their own such as a preference for specific lexical items, syntactic patterns, cohesive devices or even style of punctuation. If so, then we need to address a number of questions, such as (a) Is a translator's preference for specific linguistic options independent of the style of the original author? (b) Is it independent of general preferences of the source language and possibly the norms or poetics of a given sociolect? (c) If the answer is yes to both, is it possible to explain

those preferences in terms of the social, cultural or ideological positioning of the individual translator?

Baker (2000) made use of the fictional subcorpus of the TEC to examine aspects of linguistic patterning in the works of two British literary translators, namely Peter Bush (five translated novels or a total of 296,146 words in the corpus), and Peter Clark (three translated novels or a total of 173,932 words in the corpus). Peter Bush is found to prefer works written in Spanish and Brazilian Portuguese with an elaborate narrative which creates a world of intellectually sophisticated characters who speak largely through the narrator's voice. These works assume a highly educated readership. Peter Clark, in contrast, translates Arabic texts with an ordinary narrative which convey a social message accessible to a wider lay readership. In these stories everyday people interact with one another and focus mainly on emotions. Peter Bush's translations have a higher average sentence length and a higher type-token ratio. Moreover, the analysis of the reporting verb *say* reveals a tendency for Peter Clark to use the simple past *said* more often than any other form and in direct speech, while Peter Bush tends to use it in indirect speech in the typical structure as *someone said*, he also uses *says* more often and in indirect speech. While indirect speech creates a world with unclear boundaries where the reader is encouraged to identify with the fictional or autobiographical world, direct speech clearly defines the beginning and end of the characters' utterances and thoughts which are directly and unambiguously conveyed to the reader. The tendency to use direct speech in Peter Clark's translated narrative may be tentatively explained by a subconscious attempt to render the source text, which belongs to a distant and alien culture, more accessible to the English readership. Baker concludes that, however methodologically difficult, it is possible in principle to identify patterns of choice which together form a particular thumb-print or style of an individual literary translator. It is also possible to use the description which emerges from a study of this type to elaborate the kind of text world that each translator has chosen to recreate in the target language.

A similar study, this time in the context of South African Bible translation, is being undertaken by Rose Masubelele, who is conducting doctoral research into the role played by Bible translation in the growth and development of written Zulu (under the supervision of my colleague Dr. Kim Wallmach and co-supervision of Dr. Eric Hermanson of the Bible Society of South Africa). She aims to examine the changes in orthography, phonology, morphology, syntax, lexis and register of nine Zulu translations of the book of Matthew (1848-1997), which form a monolingual translational corpus.

## 5.3 Studying what is "in" and "of" translational English

Laviosa (2000) also makes use of the TEC as a single resource representing translational English for the lexicogrammatical analysis of five semantically related words which are frequently used in translated newspaper articles (*The Guardian* and *The European*) and can be considered to be "key words" or words that are important from a sociological point of view as they embody social values and transmit culture. The words selected are: *Europe*, *European*, *European Union*, *Union*, and *EU*. The concordance findings suggest that *Europe* appears to be viewed as a political project or a political reality. In contrast, *European* is associated with lexis that refers to the institutions that are part of the EU, as well as its political, military, economic, and diplomatic activities, and its membership. None of the collocates reveal any distinctive positive or negative semantic prosodies, they are neutral and factual rather than evaluative. Collocates not only disambiguate the meaning of polysemous words such as *Europe* as a continent and *Europe* as *EU*, but they reveal something about the cultural message subliminally conveyed by the typical use of these words. The image of Europe that seems to be portrayed by the translated articles in *The Guardian* and *The European* is that of a political reality whose activities, ideas, projects, and ideals are the object of reasonably well balanced debates and discussions, and which are reported on in a seemingly detached and objective manner. Laviosa (2000) suggests that this conclusion can become a new hypothesis that can be tested by investigating other text genres in the TEC by means of the same analytical techniques. Moreover, given that the TEC is a multi-source-language corpus of translational English, it is possible, providing the corpus is large enough, to examine the extent to which specific patterns are associated with specific source languages. Comparative analyses could also be carried out between *Europe* and other lemmas of cultural keywords such as *Britain* and *British*, *France* and *French*, and *Italy* and *Italian*, etc.

Because the results reveal descriptive features of the particular type of translational language found in newspapers, they tell us primarily something about what is "in" the language itself, while "of" refers to the text in its entirety, its overall impact in the target language and culture in terms of ideology. Therefore, although this small-scale study does not allow Laviosa (2000:172) to generalise about the semantic prosodies of the lemma *Europe* in translational English as a whole, she managed to show that it is possible to develop

> a corpus-based methodology or, more specifically, a TEC-based methodology, through which the ideological impact of translated texts can be investigated in a truly target-oriented environment,

> where the language of translation is investigated *per se*, as a specific variety of the target language, without necessarily referring to other corpora, either comparable or parallel, in order to elaborate hypotheses on the specificity of the language of translation.

In this article an overview was given of the research that has led to the formation of Corpus-based Translation Studies or CTS. I also demonstrated how CTS tools and techniques can be used for the analysis of general and literary texts, the style of professional literary translators as well as what is "in" and "of" translational English. The question which inevitably springs to mind at this point is what are the implications of corpus-based translation research for Bible translation?

## 6. ANY IMPLICATIONS FOR BIBLE TRANSLATION?

In order to discuss the implications of corpus-based research into the "universal" features of the language of translation for Bible translation it is necessary to mention its strengths and weaknesses. Laviosa (2001:74-75) points out that (1) it has developed the initial intuitive and somewhat vague notion of "universal" into clear, detailed operational research hypotheses; (2) it has progressed from small scale, manual, language pair and text genre-specific studies to large scale, systematic, comparable, and target-oriented research. (3) From somewhat inconclusive findings, the discipline has progressed to more consistent evidence which takes into account both trends and exceptions; and (4), from theoretical elaborations of the notion of "universal", mainly rooted in the linguistic tradition, corpus-based translation studies are beginning to take into account a wider range of factors involving socio-cultural elements such as the relative status of a language and the position of a literary genre within the general polysystem of literary production. This development represents an important shift from description to causal explanation of the phenomena being studied.

However, the notion of "universal", together with the search for general laws of translation, has been questioned by some scholars on theoretical and empirical grounds. From a theoretical stance, universals have been criticised as being

> anachronistic constructs, a heritage of the intellectual doctrine of positivism which aspires to objectivity and undermines the value of intuition and interpretation as means for making sense of reality, a position that is no longer tenable in the light of modern thought which, from the early decades of the twentieth century, has challenged, in the natural and the social sciences, the rigid division

> between the object of study, the act of observing it, and the observer (Tymoczko in Laviosa 2001:75).

The very process of collecting data and the tools used to analyse them are inevitably influenced by the researcher's perspective; therefore, if scholars put forward hypotheses supported by empirical evidence to suggest that the language of translation has some universal features their claim is

> basically authoritarian to the extent that it not only attempts its own local, historically and ideologically determined conclusions into general neutral "laws" but also disregards other (inescapably) local statements or hypotheses (Arrojo in Laviosa 2001:76).

According to Laviosa (2001:76), from an empirical point of view, Kenny's findings concerning lexical creativity in translations, for example, may cast doubt on the validity of normalisation as a universal feature of translation: The fact that different translators respond differently, or the same translator responds differently on different occasions, to what is essentially the same problem posed by a source text,

> lends weight to the hypothesis that normalisation is norm-governed behaviour; it represents a tendency ... rather than any absolute necessity to do so, which one might expect if normalisation were approached as a manifestation of a translation universal (Kenny in Laviosa 2001:76).

At the European Society for Translation Studies (EST) conference in Denmark in September 2001, Mona Baker (personal communication) herself retracted her label of "universals" and opted instead to call these features of translated text simply "translational patterns and regularities". Laviosa (2001:78) agrees. Provided we do not consider it as "a static, absolute category, capable of explaining the translator's choices in every circumstance", but as a "descriptive construct, an open-ended working hypothesis", she argues, the notion of universal can still be fruitfully exploited to reveal the state of the art in Descriptive Translation Studies:

> What universals-based studies intend and hope to show is not the existence of all-or-none phenomena, but tendencies, trends, regularities which do not occur in an aseptic, dull environment devoid of singular behaviours, but emerge from a rich, intricate, dynamic world of diversity and contrasts.

In my opinion, Bible translators should become acquainted with the tools and techniques used by scholars who pursue corpus-based translation research to *describe and explain in functional and systemic terms* what existing translations actually look like, and why. For instance, they may investigate

which translation strategies have been used to solve problems of non-equivalence at word level and above word level), taking into account the fact that no translation is ever produced in a vacuum, and is always shaped by the historical, cultural, linguistic, literary and religious system of the target language.

As shown above, small scale and manual investigations can be expanded by large scale, systematic, comparable, and target-oriented research of existing source texts and translations with a view to examining, and ultimately improving the quality of modern Bible translations, enhancing their accessibility, and producing new translations for different purposes and with different skopoi.

This new approach could assist researchers and practitioners by shedding light on the linguistic and textual features of different translations and/or revisions of the same Bible in respect of

- consistency of, for instance, terminology, orthography and register;
- changes in orthography, terminology and register over time;
- the effect of regionalisms or dialectal variation;
- interference of the source language (e.g. English, or Xhosa in the case of some of the translations into Zulu);
- interference of the translator's mother tongue;
- regularities and recurring patterns in certain translations but not in others;
- whether the profile of the publishing house or Bible Society that is responsible for the translation(s) impacts upon the linguistic make-up of a translation;
- institutional or team work as opposed to the work of individual translators;
- whether certain strategies or features are more typical of certain target languages than others;
- the extent to which differences in the source texts are actually reflected in the translation(s), so as to obtain some indication of whether the use of alternative source texts is theologically significant.

In conclusion, I fully agree with Tymoczko (1998:652), who, incidentally, is not personally involved in corpus-based research!, that "like large databases in the sciences, corpora will become a legacy of the present to the future, enabling future research to build upon that of the present".

# BIBLIOGRAPHY

AIJMER K & ALTENBERG B (eds.)
1991. *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman.

BAKER M
1992. *In other words: a coursebook on translation*. London: Routledge.
1993. Corpus linguistics and translation studies: implications and applications. In: M. Baker, G. Francis & E. Tognini-Bonelli (eds.):233-250.
1995. Corpora in translation studies: an overview and some suggestions for future research. *Target* 7(2):223-243.
1996a. Corpus-based translation studies: the challenges that lie ahead. In: H. Somers (ed.):175-186.
1996b. Linguistics and cultural studies: complementary or competing paradigms in translation studies? In: A. Lauer, H. Gerzymisch-Arbogast, J. Haller & E. Steiner (eds.).
1998. Réexplorer la langue de la traduction: une approche par corpus (Investigating the language of translation: a corpus-based approach). *Meta* 43(4):480-485.
1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4(2):281-298.
2000. Towards a methodology for investigating the style of a literary translator. *Target* 12(2):241-266.

BAKER M, FRANCIS G & TOGNINI-BONELLI E (eds.)
1993. *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins.

BEYLARD-OZEROFF A, KRÁLOVÁ J & MOSER-MERCER B
1998. *Translators' strategies and creativity*. Amsterdam: John Benjamins.

BIBER D
1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
1994. An analytical framework for register studies. In: D. Biber & E. Finegan (eds.):31-56.
1995a. *Dimensions of register variation*. Cambridge: Cambridge University Press.
1995b. On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: a reply to Watson. *Text* 15(3): 341-370.
1996. Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics* 1(2):171-197.

BIBER D & FINEGAN E (eds.)
1994. *Sociolinguistic perspectives on register.* Oxford: Oxford University Press.

BIBER D, CONRAD S & REPPEN R
1994. Corpus-based approaches to issues in Applied Linguistics. *Applied Linguistics* 15(2):169-189.

1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

BLUM-KULKA S
1986. Shifts of cohesion and coherence in translation. In: J. House & S. Blum-Kulka (eds.):17-35.

BOWKER L
1998. Using specialized monolingual native-language corpora as a translation resource: a pilot study. *Meta* 43(4):631-651.
1999. The design and development of a corpus-based aid for assessing translations. *Teanga* 18:11-24.

BOWKER L, CRONIN M, KENNY D & PEARSON J (eds.)
1998. *Unity in diversity? Current trends in translation studies*. Manchester: St. Jerome Publishing.

DELABASTITA D
1993. *There's a double tongue: an investigation into the translation of Shakespeare's wordplay, with special reference to 'Hamlet'*. Rodopi: Amsterdam.

DODD B (ed.)
2000. *Working with German corpora*. University of Birmingham: University Press.

EVEN-ZOHAR I
1990. Polysystem studies. *Poetics Today* 11(1):1-94.

FRAWLEY W
1984. *Translation: literary, linguistic and philosophical perspectives*. Newark: University of Delaware Press.

GRANGER S (ed.)
1998. *Learner English on computer*. London: Longman.

HALVERSON S
1998. Translation studies and representative corpora: establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta* 43(4):494-514.

HERMANS T
1985. Translation studies and a new paradigm. In: T. Hermans (ed.):7-15.
1991. Translational norms and correct translation. In: K.M. van Leuven-Zwart & T. Naaijkens (eds.):155-169.
1999. *Translation in systems: descriptive and systemic approaches explained*. Manchester: St. Jerome.

HERMANS T (ed.)
1985. *The manipulation of literature: studies in literary translation*. London: Croom Helm.

HEYLEN R
1993. *Translation, poetics and the stage: six French 'Hamlets'*. London: Routledge.

HOUSE J & BLUM-KULKA S (eds.)
1986. *Interlingual and intercultural communication: discourse and cognition in translation and second language acquisition studies*. Tübingen: Gunter Narr.

KLAUDY K & KOHN J (eds.)
1997. *Transferre necesse est*. Proceedings of the Second International Conference on current trends in studies of translation and interpreting, 5-7 September 1996, Budapest, Hungary. Budapest: Scholastica.

KENNY D
1998. Creatures of habit? What translators usually do with words. *Meta* 43(4):515-523.
1999. The German-English parallel corpus of literary texts (GEPCOLT): a resource for translation scholars. *Teanga* 18:25-42.
2000. Translators at play: exploitations of collocational norms in German-English translation. In: B. Dodd (ed.):143-160.
2001. *Lexis and creativity in translation: a corpus-based study*. Manchester: St. Jerome.

KITTEL H (ed.)
1992. *Geschichte, System, Literarische Übersetzung/Histories, systems, literary translations*. Berlin: Erich Schmidt Verlag.

KITTEL H & FRANK A P (eds.)
1991. *Interculturality and the historical study of literary translations*. Berlin: Erich Schmidt Verlag.

KRUGER A
1998. Shakespeare translations in South Africa: a history. In: A. Beylard-Ozeroff, J. Králová & B. Moser-Mercer:107-115.
2000. Lexical cohesion and register variation in translation: *the merchant of Venice* in Afrikaans. Unpublished D. Litt. et Phil. thesis, Pretoria: University of South Africa.

KRUGER A & WALLMACH K
1997. Research methodology for the description of a source text and its translation(s) — a South African perspective. *SA Journal of African Languages* 12(4):119-126.

LAMBERT J & VAN GORP H
1985. On describing translations. In: T. Hermans (ed.):42-53.

LAUER A, GERZYMISCH-ARBOGAST H, HALLER J & STEINER E (eds.)
1996. *Übersetzungswissenschaft im Umbruch: Festschrift für Wolfram Wilss zum 70. Geburtstag*. Tübingen: Gunter Narr.

LAVIOSA S
1997. How comparable can 'comparable corpora' be? *Target* 9(2):289-319.
1998a. The English Comparable Corpus: a resource and a methodology. In: L. Bowker, M. Cronin, D. Kenny & J. Pearson (eds.):101-112.
1998b. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4):557-570.
2000. TEC: a resource for studying what is "in" and "of" translational English. *Across Languages and Cultures* 1(2):159-177.
2001. *Corpus-based translation studies: theory, findings, applications*. Amsterdam: Rodopi.

LAVIOSA-BRAITHWAITE S
1996a. Comparable corpora: towards a corpus linguistic methodology for the empirical study of translation. In: M. Thelen & B. Lewandowska-Tomaszczyk (eds.):153-163.
1996b. The English Comparable Corpus (ECC): a resource and a methodology for the empirical study of translation. Unpublished Ph. D. thesis, Dept of Language Engineering. Manchester: UMIST.
1997. Investigating simplification in an English Comparable Corpus of newspaper articles. In: K. Klaudy & J. Kohn (eds.):531-540.

LEECH G
1991. The state of the art in corpus linguistics. In: K. Aijmer & B. Altenberg (eds.):8-29.

LEWANDOWSKA-TOMASZCZYK B & MELIA P J (eds.)
1997. *Practical applications in language corpora*. Lódź: Lódź University Press.

LUTHER JANA
1998. Personal communication and correspondence. Senior Editor (Afrikaans) of Pharos Dictionaries, a division of Nasionale Pers-Boekhandel in Cape Town.

MALHERBE D F
1949. *Die koopman van Venesië*. Johannesburg: Afrikaanse Pers-Boekhandel.

MALMKJÆR K
1998. Love thy neighbour: will parallel corpora endear linguists to translators? *Meta* 43(4):534-541.

MEUNIER F
1998. Computer tools for the analysis of learner corpora. In: S. Granger (ed.):19-37.

MUNDAY J
1998. A computer-assisted approach to the analysis of translation shifts. *Meta* 43(4):542-556.

NORD C
1997. *Translating as a purposeful activity: functionalist approaches explained*. Manchester: St. Jerome.

OLOHAN M & BAKER M
2000. Reporting that in translated English: evidence for subconscious processes of explicitation. *Across Languages and Cultures* 1(2):141-158.

ØVERÅS L
1998. In search of the third code: an investigation of norms in literary translation. *Meta* 43(4):571-588.

PEARSON J
1998. *Terms in context*. Amsterdam: John Benjamins.
1999. Genes go wild in the countryside: using corpora to improve translation quality. *Teanga* 18:71-83.

POTGIETER T
1991. Die sakeman van Venesië. Unpublished manuscript. (PACT; the translator).

SHLESINGER M
1998. Corpus-based interpreting as an offshoot of corpus-based translation studies. *Meta* 43(4):487-493.

SINCLAIR J
1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SOMERS H (ed.)
1996. *Terminology, LSP and translaton: studies in language engineering in honour of Juan C. Sager*. Amsterdam: John Benjamins.

STUBBS M
1996. *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell.

THELEN M & LEWANDOWSKA-TOMASZCZYK B (eds.)
1996. *Translation and meaning. Part 3*. Maastricht: Hogeschool Maastricht.

TOURY G
1980. *In search of a theory of translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.
1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

TYMOCZKO M
1998. Computerized corpora and the future of translation studies. *Meta* 43(4):652-660.

ULRYCH M
1997. The impact of multilingual parallel concordancing on translation. In: B. Lewandowska-Tomaszczyk & P.J. Melia (eds.):421-435.

VAN DEN BROECK R
1985. Second thoughts on translation criticism: a model of its analytic function. In: T. Hermans (ed.):54-62.

VAN LEUVEN-ZWART K
1989. Translation and original: similarities and dissimilarities, part 1. *Target* 1(2):151-181.
1990. Translation and original: similarities and dissimilarities, part 2. *Target* 2(1):69-95.

VAN LEUVEN-ZWART K M & NAAIJKENS T (eds.)
1991. Translation studies: the state of the art. Proceedings of the First James S. Holmes Symposium on Translation Studies. Amsterdam: Rodopi.

ZANETTIN F
1998. Bilingual comparable corpora and the training of translators. *Meta* 43(4):616-630.

*Keywords*

Translation

Corpus-based translation

General translation

Literary translation

Bible translation

Descriptive translation studies

Corpus linguistics