In pursuit of the 'third code': Using the *ZJU Corpus of Translational Chinese* in translation studies

Richard Xiao, Lianzhen He, and Ming Yue

Zhejiang University

1. Introduction

Since the 1990s, the rapid development of the corpus-based approach in linguistic investigation in general, and the development of multilingual corpora in particular, have brought even more vigor into Descriptive Translation Studies (DTS) (cf. McEnery, Xiao and Tono 2006: 90-95). As Laviosa (1998a: 474) observes, 'the corpus-based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation.' Presently, corpus-based DTS has primarily been concerned with describing translation as a product, by comparing corpora of translated and non-translational native texts in the target language, especially translated and native English. The majority of product-oriented translation studies attempt to uncover evidence to support or reject the so-called translation universal (TU) hypotheses that are concerned with features of translational language as the 'third code' of translation (Frawley 1984), which is supposed to be different from both source and target languages.

As far as the English language is concerned, a large part of product-oriented translation studies

have been based on the *Translational English Corpus* (TEC), which was built by Mona Baker and colleagues at the University of Manchester. The TEC corpus, which was designed specifically for the purposes of studying translated English, consists of contemporary written texts translated into English from a range of source languages. It is constantly expanded with fresh materials, reaching a total of 20 million words by the year 2001. The corpus comprises full texts from four genres (fiction, biography, newspaper articles and in-flight magazines) translated by native speakers of English. Paralinguistic data such as the information of translators, source texts and publishing dates is annotated and stored in the header section of each text. A subcorpus of native English was specifically selected and is being modified from the *British National Corpus* (BNC) to match the TEC in terms of both composition and dates of publication.

The TEC corpus is perhaps the only publicly available corpus of translational English. Most of the pioneering and prominent studies of translational English, which have so far focused on syntactic and lexical features of translated and original texts of English, have been based on this corpus. They have provided evidence to support the hypotheses of translational universals in translated English, most noticeably simplification, explicitation, sanitization, and normalization (see section 2 for further discussion). For example, Laviosa (1998b) studies the distinctive features of translational English in relation to native English (as represented by the BNC corpus), finding that translational language has four core patterns of lexical use: a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, a relatively greater repetition of the most frequent words, and a smaller vocabulary frequently used. This is regarded as the most significant work in support of the simplification hypothesis of translation universals. Olohan and Baker's (2000) comparison of

concordances from the TEC and the BNC corpora shows that the *that*-connective with reporting verbs *say* and *tell* is far more frequent in translational English, and conversely, that the zero-connective is more frequent in native English. 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 hypothesized to be a subliminal phenomenon inherent in the translation process' (Laviosa 2002: 68). Olohan (2004) investigates intensifiers such as *quite*, *rather*, *pretty* and *fairly* in translated versus native English fiction in an attempt to uncover the relationship between collocation and moderation, finding that *pretty* and *rather*, and more marginally *quite*, are considerably less frequent in the TEC-fiction subcorpus; but when they are used, there is usually more variation in usage, and less repetition of common collocates, than in the BNC-fiction corpus.

Similar features have also been reported in the translational variant of a few languages other than English (e.g. Swedish). Nevertheless, research of this area has so far been confined largely to translational English translated from closely related European languages (e.g. Mauranen and Kujamaki 2004). If the features of translational language that have been reported are to be generalized as 'translational universals', the language pairs involved must not be restricted to English and closely related languages. Evidence from 'genetically' distinct language pairs such as English and Chinese is undoubtedly more convincing, if not indispensable. This motivates us to undertake a project that studies the features of translational Chinese.

This article first reviews previous research of the features of translational language (section 2). We will then introduce the newly created *ZJU Corpus of Translational Chinese* (ZCTC), which is

designed with the explicit aim of studying translational Chinese (section 3). Section 4 presents a number of case studies of the lexical and syntactic features of translational Chinese while section 5 concludes the article.

2. Translation universals: a review

An important area of Descriptive Translation Studies is the hypothesis of so-called translation universals (TUs) and its related sub-hypotheses, which are sometimes referred to as the inherent features of translational language, or 'translationese'. It is a well-recognized fact that translations cannot possibly avoid the effect of translationese (cf. Hartmann 1985; Baker 1993: 243-245; Teubert 1996: 247; Gellerstam 1996; Laviosa 1997: 315; McEnery and Wilson 2001: 71-72; McEnery and Xiao 2002, 2007). The concept of TUs is first proposed by Baker (1993), who suggests that all translations are likely to show certain linguistic characteristics simply by virtue of being translations, which are caused in and by the process of translation. The effect of the source language on the translations is strong enough to make the translated language perceptibly different from the target native language. Consequently translational language is at best an unrepresentative special variant of the target language (McEnery and Xiao 2007). The distinctive features of translational language can be identified by comparing translations with comparable native texts, thus throwing new light on the translation process and helping to uncover translation norms, or what Frawley (1984) calls the 'third code' of translation.

Over the past decade, TUs have been an important area of research as well as a target of debate in Descriptive Translation Studies. Some scholars (e.g. Tymoczko 1998) argue that the very idea of making universal claims about translation is inconceivable, while others (e.g. Toury 2004)

advocate that the chief value of general laws of translation lies in their explanatory power; still others (e.g. Chesterman 2004) accept universals as one possible route to high-level generalizations. Chesterman (2004) further differentiates between two types of TUs: one relates to the process from the source to the target text (what he calls 'S-universals'), while the other ('T-universals') compares translations to other target-language texts. Mauranen (2007), in her comprehensive review of TUs, suggests that the discussion of TUs follow the general discussion on 'universals' in language typology.

Recent corpus-based works have proposed a number of TUs, the best known of which include explicitation, simplification, normalization, sanitization and leveling out (or convergence). Other TUs that have been investigated include under-representation, interference and untypical collocations (see Mauranen 2007). While individual studies have sometimes investigated more than one of these features, they are discussed in the following subsections separately for the purpose of this presentation.

2.1 Explicitation

The explicitation hypothesis is formulated by Blum-Kulka (1986) on the basis of evidence from individual sample texts showing that translators tend to make explicit optional cohesive markers in the target text even though they are absent in the source text. It relates to the tendency in translations to 'spell things out rather than leave them implicit' (Baker 1996: 180). Explicitation can be realized syntactically or lexically, for instance, via more frequent use of conjunctions in translated texts than in non-translated texts, additions providing extra information essential for a target culture reader, and resulting in longer text than the non-translated text. For example,

Chen (2006) presents a corpus-based study of connectives, namely conjunctions and sentential adverbials, in a ‘composite corpus’ composed of English source texts and their two Chinese versions independently produced in Taiwan and mainland China, plus a comparable component of native Chinese texts as the reference corpus in the genre of popular science writing. This investigation integrates product- and process-oriented approaches in an attempt to verify the hypothesis of explicitation in translated Chinese. In the product-oriented part of his study, Chen compares translational and native Chinese texts to find out whether connectives are significantly more common in the first type of texts in terms of parameters such as frequency and type-token ratio, as well as statistically defined common connectives and the so-called translationally distinctive connectives (TDCs). He also examines whether syntactic patterning in the translated texts is different from native texts via a case study of five TDCs that are most statistically significant. In the process-oriented part of the study, he compares translated Chinese texts with the English source texts, through a study of the same five TDCs, in an attempt to determine the extent to which connectives in translated Chinese texts are carried over from the English source texts, or in other words, the extent to which connectives are explicitated in translational Chinese. Both parts of his study support the hypothesis of explicitation as a translation universal in the process and product of English-Chinese translation of popular science writing.

Another result of explicitation is increased cohesion in translated text (Øverås 1998). Pym (2005) provides an excellent account of explicitation, locating its origin, discussing its different types, elaborating a model of explicitation within a risk-management framework, and offering a range of explanations of the phenomenon.

In the light of the distinction made above between S- and T-universals (Chesterman 2004), explicitation would seem to fall most naturally into the S-type. Recently, however, explicitation has also been studied as a T-universal. In his corpus-based study of structures involving NP modification (i.e. equivalent of the structure noun + prepositional phrase in English) in English and Hungarian, Váradi (2007) suggests that genuine cases of explicitation must be distinguished from constructions that require expansion in order to meet the requirements of grammar. While explicitation is found at various linguistic levels ranging from lexis to syntax and textual organization, 'there is variation even in these results, which could be explained in terms of the level of language studied, or the genre of the texts' (Mauranen 2007: 39). The question of whether explicitation is a translation universal is yet to be conclusively answered, according to existing evidence which has largely come from translational English and related European languages (see section 4 for further discussion).

2.2 Simplification

Explicitation is related to simplification: 'the tendency to simplify the language used in translation' (Baker 1996: 181-182), which means that translational language is supposed to be simpler than native language, lexically, syntactically and / or stylistically (cf. Blum-Kulka and Levenston 1983; Laviosa-Braithwaite 1997). As noted earlier, product-oriented studies such as Laviosa (1998b) and Olohan and Baker (2000) have provided evidence for lexical and syntactic simplification in translational English. Translated texts have also been found to be simplified stylistically. For example, Malmkjaer (1997) notes that in translations, punctuation usually becomes stronger; for example commas are often replaced with semicolons or full stops while semicolons are replaced with full stops. As a result, long and complex sentences in the source text

tend to be broken up into shorter and less complex clauses in translations, thereby reducing structural complexity for easier reading. On the other hand, Laviosa's (1998b: 5) observes that translated language has a significantly greater mean sentence length than non-translated language. Xiao and Yue's (2008) finding that translated Chinese fiction displays a significantly greater mean sentence length than native Chinese fiction is in line with Laviosa's (1998b: 5) observation but goes against Malmkjaer's (1997) expectation that stronger punctuations tend to result in shorter sentences in translated texts. It appears, then, that mean sentence length might not be a translational universal but rather associated with specific languages or genres (see section 4.1 for further discussion).

The simplification hypothesis, however, is controversial. It has been contested by subsequent studies of collocations (Mauranen 2000), lexical use (Jantunen 2001), and syntax (Jantunen 2004). Just as Laviosa-Braithwaite (1996: 534) cautions, evidence produced in early studies that support the simplification hypothesis is patchy and not always coherent. Such studies are based on different datasets and are carried out to address different research questions, and thus cannot be compared.

2.3 Normalization

Normalization, which is also called 'conventionalization' in the literature (e.g. Mauranen 2007), refers to the 'tendency to exaggerate features of the target language and to conform to its typical patterns' (Baker 1996: 183). As a result, translational language appears to be 'more normal' than the target language. Typical manifestations of normalization include overuse of clichés or typical grammatical structures of the target language (but see section 4.4 for counter evidence), adapting

punctuation to the typical usage of the target language, and the treatment of the different dialects used by certain characters in dialogues in the source texts.

Kenny (1998, 1999, 2000, and 2001) presents a series of studies of how unusual and marked compounds and collocations in German literary texts are translated into English, in an attempt to assess whether they are normalized by means of more conventional use. Her research suggests that certain translators may be more inclined to normalize than others, and that normalization may apply in particular to lexis in the source text. Nevalainen (2005, cited in Mauranen 2007: 41) suggests that translated texts show greater proportions of recurrent lexical bundles or word clusters.

Like simplification, normalization is also a debatable hypothesis. According to Toury (1995: 208), it is a 'well-documented fact that in translations, linguistic forms and structures often occur which are rarely, or perhaps even never encountered in utterances originally composed in the target language.' Tirkkonen-Condit's (2002: 216) experiment, which asked subjects to distinguish translations from non-translated texts, also shows that 'translations are not readily distinguishable from original writing on account of their linguistic features.'

2.4 Other translational universals

Kenny (1998) analyzes semantic prosody in translated texts in an attempt to find evidence of sanitization (i.e. reduced connotational meaning). She concludes that translated texts are 'somewhat "sanitized" versions of the original' (Kenny 1998: 515). Another translational universal that has been proposed is the so-called feature of 'leveling out', i.e. 'the tendency of

translated text to gravitate towards the centre of a continuum' (Baker 1996: 184). This is what Laviosa (2002: 72) calls 'convergence', i.e. the 'relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features' that are discussed above.

'Under representation', which is also known as the 'unique items hypothesis', is concerned with the unique items in translation (Mauranen 2007: 41-42). For example, Tirkkonen-Condit (2005) compared frequency and uses of the clitic particle *kin* in translated and original Finnish in five genres (i.e. fiction, children's fiction, popular fiction, academic prose, and popular science), finding that the average frequency of *kin* in original Finnish is 6.1 instances per 1,000 words, whereas its normalized frequency in translated Finnish is 4.6 instances per 1,000 words. Tirkkonen-Condit interprets this result as a case of under representation in translated Finnish. Aijmer's (2007) study of the use of English discourse marker *oh* and its translation in Swedish shows that there is no single lexical equivalent of *oh* in Swedish translation, because direct translation with the standard Swedish equivalent *áh* would result in an unnatural sounding structure in this language.

3. The ZJU Corpus of Translational Chinese

As can be seen in the discussion above, while we have followed the convention of using the term 'translation universal', the term is highly debatable in the literature. Since the translational universals that have been proposed so far are identified on the basis of translational English – mostly translated from closely related European languages, there is a possibility that such linguistic features are not 'universal' but rather specific to English and / or genetically related

languages that have been investigated. For example, Cheong's (2006) study of English-Korean translation contradicts even the least controversial explicitation hypothesis.

We noted in section 2.1 that the explicitation hypothesis is supported by Chen's (2006) study of connectives in English-Chinese translation of popular science books. Nevertheless, as Biber (1995: 278) observes, language may vary across genres even more markedly than across languages. Xiao (2008) also demonstrates that the genre of scientific writing is the least diversified of all genres across various varieties of English. The implication is that the similarity reported in Chen (2006) might be a result of similar genre instead of language pair. Ideally, what is required to verify the English-based translation universals is a detailed account of the features of translational Chinese based on balanced comparable corpora of translational and native Chinese. This is the aim of our ongoing project 'A corpus-based quantitative study of translational Chinese in English-Chinese translation', which is funded by the China National Foundation of Social Sciences.

The project has two major parts. The first part aims to develop a translational counterpart of the *Lancaster Corpus of Mandarin Chinese* (LCMC), a one-million-word balanced corpus of native Chinese, while the second part undertakes a quantitative study of translational Chinese using a composite approach that integrates monolingual comparable corpus analysis and parallel corpus analysis as advocated in McEnery and Xiao (2002). The monolingual comparable corpus approach compares comparable corpora of translated language with the native target language in an attempt to uncover salient features of translations, while the parallel corpus approach compares source and target languages to determine to what extent the features of translated texts are transferred from the

source texts.

We have so far completed the first part of the project. The reminder of this section introduces the *ZJU Corpus of Translational Chinese* (ZCTC), while section 4 will present a number of case studies based on this corpus.

3.1. Corpus design

The *ZJU Corpus of Translational Chinese* (ZCTC) is created with the explicit aim of studying the features of translated Chinese in relation to non-translated native Chinese. It has modeled the *Lancaster Corpus of Mandarin Chinese* (LCMC), which is a one-million-word balanced corpus designed to represent native Mandarin Chinese (McEnery and Xiao 2004). Both LCMC and ZCTC corpora have sampled five hundred 2,000-word text chunks from fifteen written text categories published in China, with each amounting to one million words. Table 1 shows the text categories covered in the two corpora, together with their respective proportions.

Since the LCMC corpus was designed as a Chinese match for the FLOB corpus of British English (Hundt, Sand and Siemund 1998) and the Frown corpus of American English (Hunt, Sand and Skandera 1999), with the specific aim of comparing and contrasting English and Chinese, it has also followed the sampling period of FLOB / Frown and sampled written Mandarin Chinese within three years around 1991. While it was relatively easy to find texts of native Chinese published in this sampling period, it would be much more difficult to get access to translated Chinese texts of some genres - especially in electronic format - published within this time frame. This pragmatic consideration of data collection has forced us to modify the LCMC model slightly by extending the

sampling period by a decade, i.e. to 2001, when we built the ZCTC corpus. This extension has been particularly useful because the popularization of the Internet and online publication in the 1990s have made it possible and easier to access a large amount of digitalized texts.[1]

Table 1. The genres covered in LCMC and ZCTC

Code	Genre	Number of samples	Proportion
A	Press reportage	44	8.8%
B	Press editorials	27	5.4%
C	Press reviews	17	3.4%
D	Religious writing	17	3.4%
E	Skills, trades and hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies and essays	77	15.4%
H	Miscellaneous (reports, official documents)	30	6%
J	Science (academic prose)	80	16%
K	General fiction	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Adventure fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humor	9	1.8%
Total		500	100%

While English is the source language of the vast majority of the text samples included in the ZCTC corpus, we have also included a small number of texts translated from other languages to mirror the reality of the world of translations in China.

Table 2. A comparison of ZCTC and LCMC corpora

Genre	ZCTC	Proportion	LCMC	Proportion
A	88,196	8.67	89,367	8.73
B	54,171	5.32	54,595	5.33
C	34,100	3.35	34,518	3.37
D	35,139	3.45	35,365	3.46
E	76,681	7.54	77,641	7.59
F	89,675	8.81	89,967	8.79
G	155,601	15.29	156,564	15.30
H	60,352	5.93	61,140	5.97
J	164,602	16.18	163,006	15.93
K	60,540	5.95	60,357	5.90
L	48,924	4.81	49,434	4.83
M	12,267	1.21	12,539	1.23
N	59,042	5.80	60,398	5.90
P	59,033	5.80	59,851	5.85
R	19,072	1.87	18,645	1.82
Total	1,017,395	100.00	1,023,387	100.00

As Chinese is written as running strings of characters without white spaces delimiting words, it is only possible to know the number of tokens in a text when the text has been tokenized (see section

3.2). As such, the text chunks were collected at the initial stage by using our best estimate (1:1.67) between the number of characters and number of words based on our previous experience (McEnery, Xiao and Mo 2003). Only textual data was included, with graphs and tables in the original texts replaced by placeholders. A text chunk included in the corpus can be a sample from a large text (e.g. an article and book chapter) or an assembly of several small texts (e.g. for the press categories and humors). When parts of large texts are selected, an attempt has been made to achieve a balance between initial, medial and ending samples. When the texts are tokenized, a computer program was used to cut large texts to approximately 2,000 tokens while keeping the final sentence complete. As a result, while some text samples may be slightly longer than others, they are typically around 2,000 words. Table 2 compares the actual numbers of tokens in different genres as well as their corresponding proportions in the ZCTC and LCMC corpora.[2] As can be seen, the two corpora are roughly comparable in terms of both overall size and proportions for different genres.

3.2. Corpus annotation

The ZCTC corpus is annotated using ICTCLAS2008, the latest release of the *Chinese Lexical Analysis System* developed by the Institute of Computing Technology, the Chinese Academy of Sciences. This annotation tool, which relies on a large lexicon and the Hierarchical Hidden Markov Model (HMM), integrates word tokenization, named entity identification, unknown word recognition, as well as part-of-speech (POS) tagging. The ICTCLAS part-of-speech tagset distinguishes between 22 level 1 part-of-speech categories (see Table 3), which expand into over 80 levels 2 and 3 categories for word tokens in addition more than a dozen categories for symbols and punctuations.[3] The ICTCLAS2008 tagger has been reported to achieve a precision rate of

98.54% for word tokenization. Latest open tests have also given encouraging results, with a precision rate of 98.13% for tokenization and 94.63% for part-of-speech tagging.[4]

Table 3. Level 1 part-of-speech categories

Level 1 POS category	Explanation
a	Adjective
b	Non-predicate noun modifier
c	Conjunction
d	Adverb
e	Interjection
f	Space word
h	Prefix
k	Suffix
m	Numeral and quantifier
n	Noun
o	Onomatopoeia
p	Preposition
q	Classifier
r	Pronoun
s	Place word
t	Time word
u	Auxiliary
v	Verb

w	Symbol and punctuation
x	Non-word character string
y	Particle
z	Descriptive adjective

3.3. Corpus markup

The ZCTC corpus is marked up in Extensible Markup Language (XML) which is complaint with the Corpus Encoding Standards (CES). Each of the 500 data files has two parts: a corpus header and a body. The *cesHeader* gives general information about the corpus (*publicationStmt*) as well as specific attributes of the text sample (*fileDesc*). Details in the *publicationStmt* element include the name of the corpus in English and Chinese, authors, distributor, availability, publication date, and history. The *fileDesc* element shows the original title(s) of the text(s) from which the sample was taken, individuals responsible for sampling and corpus processing, the project that creates the corpus file, date of creation, language usage, writing system, character encoding, and mode of channel.

The body part of the corpus file contains the textual data, which is marked up for structural organization such as paragraphs (*p*) and sentences (*s*). Sentences are consecutively numbered for easy reference. Part-of-speech annotation is also given in XML, with the POS attribute of the *w* element indicating its part-of-speech category.

The XML markup of the ZCTC corpus is perfectly well-formed and has been validated using Altova XMLSpy 2008, a comprehensive editing tool for XML documents. The XML elements of

the corpus are defined in the accompanying Document Type Definition. The ZCTC corpus is encoded in Unicode, applying the Unicode Transformation Format 8-Bit (UTF-8), which is a lossless encoding for Chinese while keeping the XML files at a minimum size. The combination of Unicode and XML is a general trend and standard 'configuration' in corpus development, especially when corpora involve languages other than English (cf. Xiao, McEnery, Baker and Hardie 2004).

4. Some lexical and syntactic features of translational Chinese

This section presents four case studies of lexical and syntactic features of translational Chinese as represented in the new ZCTC corpus in comparison with the retagged edition of the LCMC corpus (see note 2). We will first verify Laviosa's (1998b) core features of lexical use in translational Chinese (sections 4.1 and 4.2), and then compare the use of connectives and passives in translated and native Chinese (sections 4.3 and 4.4).

4.1. Lexical density and mean sentence length

This section discusses the parameters used in Laviosa (1998b) in an attempt to find out whether the core patterns of lexical use that Laviosa observes in translational English also apply in translated Chinese. We will first compare lexical density and mean sentence length in native and translated Chinese, and then examine the frequency profiles of the two corpora in the following section.

There are two common measures of lexical density. Stubbs (1986: 33; 1996: 172) defines lexical density as the ratio between the number of lexical words (i.e. content words) and the total number of words. This approach is taken in Laviosa (1998b). As our corpora are part-of-speech tagged,

frequencies of different POS categories are readily available.

The other approach commonly used in corpus linguistics is the type-token ratio (TTR), i.e. the ratio between the number of types (i.e. unique words) and the number of tokens (i.e. running words). However, since the TTR is seriously affected by text length, it is reliable only when texts of equal or similar length are compared. To remedy this issue, Scott (2004) proposes a different strategy, namely, using a standardized type-token ratio (STTR), which is computed every n (the default setting is 1,000 in the WordSmith Tools) words as the Wordlist application of WordSmith goes through each text file in a corpus. The STTR is the average type-token ratio based on consecutive 1,000-word chunks of text (Scott 2004: 130). It appears that lexical density defined by Stubbs (1986, 1996) measures informational load whereas the STTR is a measure of lexical variability, as reflected by the different ways they are computed.

Let us first examine the Stubbs-style lexical density in native and translational Chinese. Xiao and Yue (2008) find that the lexical density in translated Chinese fiction (58.69%) is significantly lower than that in native Chinese fiction (63.19%). Does this result also hold for other genres or for Mandarin Chinese in general as represented in the two balanced corpora in the present study? Figure 1 shows the scores of lexical density in the fifteen genres covered in the ZCTC and LCMC corpora as well as their mean scores. As can be seen, the mean lexical density in LCMC (66.93%) is considerably greater than that in ZCTC (61.59%). This mean difference -5.34 is statistically significant ($t = -4.94$ for 28 d.f., $p<0.001$). It is also clear from the figure that all of the fifteen genres have a greater lexical density in native than translated Chinese, which is statistically significant for nearly all genres (barring M, i.e. science fiction), as indicated by the statistic tests in

Table 4. These findings are in line with Laviosa's (1998b) observations of lexical density in translational English.

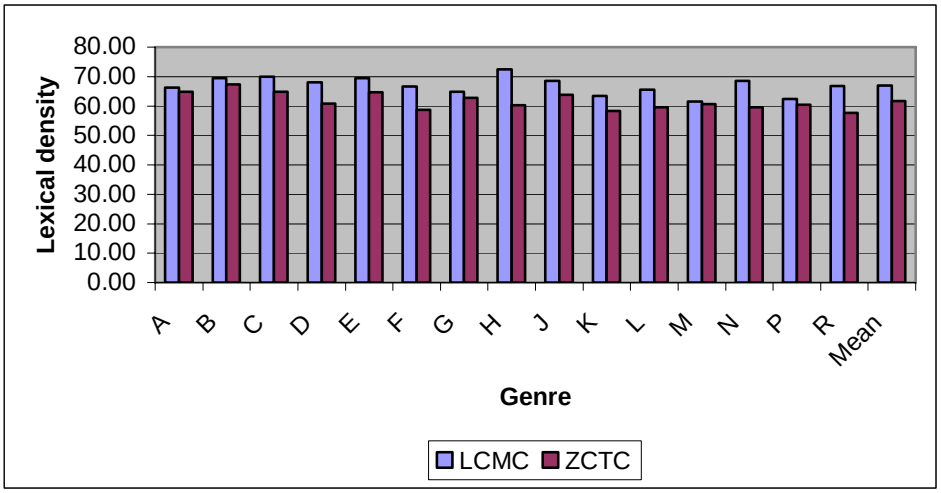

Figure 1. Lexical density in ZCTC and LCMC

Table 4. Mean differences in lexical density across genres

Genre	t score	Degree of freedom	Significance level	Mean difference
A	-2.43	86 d.f.	p=0.017	-1.446
B	-3.35	52 d.f.	p=0.002	-2.180
C	-6.96	32 d.f.	p<0.001	-5.144
D	-8.07	32 d.f.	p<0.001	-7.307
E	-4.93	74 d.f.	p<0.001	-4.703
F	-9.79	86 d.f.	p<0.001	-7.934
G	-4.05	152 d.f.	p<0.001	-2.184
H	-9.61	58 d.f.	p<0.001	-12.21

J	-9.13	158 d.f.	p<0.001	-4.777
K	-5.64	56 d.f.	p<0.001	-5.193
L	-6.28	46 d.f.	p<0.001	-5.984
M	-0.44	10 d.f.	p=0.667	-1.056
N	-13.66	56 d.f.	p<0.001	-9.122
P	-2.29	56 d.f.	p=0.026	-1.987
R	-8.85	16 d.f.	p<0.001	-9.215
Mean	-4.94	28 d.f.	p<0.001	-5.342

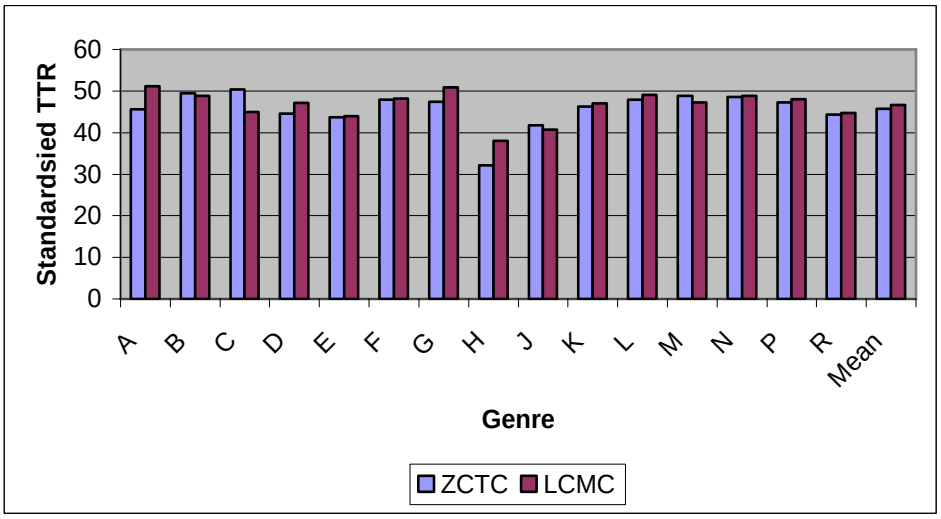

Figure 2. Standardized TTR in ZCTC and LCMC

However, if lexical density is measured by the STTR, then the LCMC corpus as a whole has a slightly higher STTR than ZCTC (46.58 vs. 45.73), but the mean difference (-0.847) is not statistically significant (t = - 0.573 for 28 d.f., p=0.571). This result is further confirmed by a closer look at individual genres (Figure 2). As can be seen, the differences for most genres are marginal.

While some genres display a greater STTR in native Chinese, there are also genres with a greater STTR in translated Chinese. This finding extends Xiao and Yue's (2008) observation of translated Chinese fiction to Mandarin Chinese in general.

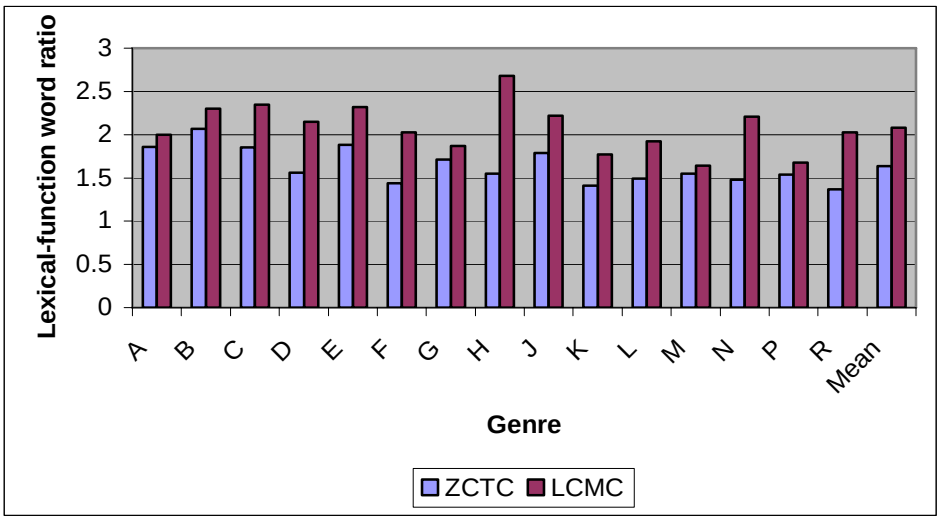

Figure 3. Lexical-function word ratios in ZCTC and LCMC

In terms of lexical versus function words,[5] a significantly greater ratio of lexical over function words is found in native Chinese than in translated Chinese (2.08 vs. 1.64, $t = -4.88$ for 28 d.f., $p<0.001$, mean difference = -0.441). As can be seen in Figure 3, which gives the lexical-function word ratios in ZCTC and LCMC, all genres have a greater ratio in native Chinese than in translated Chinese, and the mean differences for all genres other than science fiction (M) are statistically significant (see Table 5), especially in reports and official documents (H), adventure fiction (N) and humors (R). This result is in line with Xiao and Yue's (2008) observation of translated Chinese fiction and further confirms Laviosa's (1998b: 8) initial hypothesis that translational language has a relatively lower proportion of lexical words over function words.

Table 5. Mean differences in lexical-function word ratio across genres

Genre	t score	Degree of freedom	Significance level	Mean difference
A	-2.60	86 d.f.	p=0.011	-0.132
B	-3.34	52 d.f.	p=0.002	-0.228
C	-6.84	32 d.f.	p<0.001	-0.497
D	-7.94	32 d.f.	p<0.001	-0.588
E	-5.05	74 d.f.	p<0.001	-0.438
F	-9.10	86 d.f.	p<0.001	-0.590
G	-3.98	152 d.f.	p<0.001	-0.167
H	-9.88	58 d.f.	p<0.001	-1.125
J	-8.96	158 d.f.	p<0.001	-0.435
K	-5.42	56 d.f.	p<0.001	-0.364
L	-6.23	46 d.f.	p<0.001	-0.431
M	-0.59	10 d.f.	p=0.571	-0.097
N	-13.01	56 d.f.	p<0.001	-0.730
P	-2.34	56 d.f.	p=0.023	-0.139
R	-8.46	16 d.f.	p<0.001	-0.664
Mean	-4.88	28 d.f.	p<0.001	-8.441

On the other hand, as noted in section 2.2, there have been conflicting observations of mean sentence length as a sign of simplification. Figure 4 shows the mean sentence length scores of various genres in native and translated Chinese. It can be seen that while native Chinese has a slightly greater mean sentence length, the mean difference between ZCTC and LCMC (-1.533) is

not statistically significant (t = - 1.41 for 28 d.f., p = 0.17). In both native and translated Chinese, genres such as humor (R) use relatively shorter sentences whereas genres such as academic prose (J) use long sentences; in some genres there is a sharp contrast between native and translated Chinese (e.g. science fiction M) whereas in other genres the differences are less marked (e.g. academic prose J and press reportage A). It appears, then, that mean sentence length is more sensitive to genres than being a reliable indicator of native versus translational language.

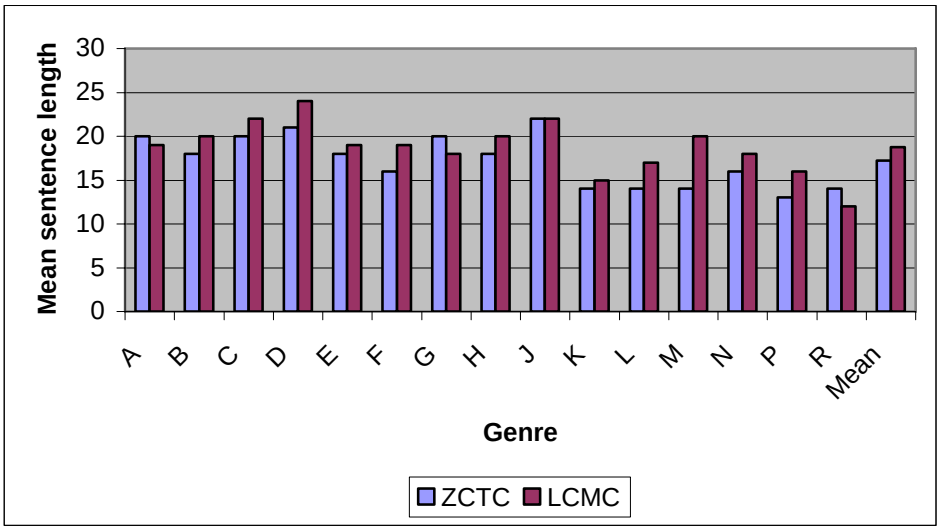

Figure 4. Mean sentence length in ZCTC and LCMC

4.2 Frequency profiles

Laviosa (1998b) defines ‘list head’ or ‘high frequency words’ as every item which individually accounts for at least 0.10% of the total tokens in a corpus. In Laviosa’s study, 108 items were high frequency words, most of which were function words. In the present case study, we also define high frequency words as those with a minimum proportion of 0.10%. But the numbers of items included can vary depending on the corpus being examined.

Table 6. Frequency profiles of ZCTC and LCMC

Type	ZCTC	LCMC
Number of items	114	108
Cumulative proportion	40.47%	35.70%
Repetition rate of high frequency words	3154.37	2870.37
Ratio of high / low frequency words	0.6988	0.5659

Table 6 shows the frequency profiles of translated and native Chinese corpora. As can be seen, while the numbers of high frequency words are very similar in the two corpora (114 and 108 respectively), high frequency words account for a considerably greater proportion of tokens in the translational corpus (40.47% in comparison to 35.70% for the native corpus). The ratio between high- and low-frequency words is also greater in translated Chinese (0.6988) than in native Chinese fiction (0.5659). Laviosa (1998b) hypothesizes on the basis of the results of lemmatization that there is less variety in the words that are most frequently used. As Chinese is a non-inflectional language, lemmatization is irrelevant; and as noted earlier, the standardized type-token ratios as a measure of lexical variability are very similar in translated and native Chinese. Nevertheless, it can be seen in Table 6 that high frequency words display a much greater repetition rate in translational than native Chinese (3154.37 versus 2870.37).

The above discussion suggests the core lexical features proposed by Laviosa (1998b) for translational English are essentially also applicable in translated Chinese, though the mean sentence length is less reliable as an indicator of simplification in translational Chinese.

4.3 Connectives as a device for explicitation

Chen (2006) finds that in his Chinese corpus of popular science books translated from English, connectives are significantly more common than in a comparable corpus of original Chinese scientific writing; some connectives are also found to be translationally distinctive, i.e. significantly more common in translated texts. Chen (2006) concludes that connectives are a device for explicitation in English-Chinese translation of popular science books. Xiao and Yue (2008) also note that connectives are significantly more frequent in translated than native Chinese fiction. In this section, we will compare the two balanced corpora of translated and native Chinese in terms of their frequency and use of connectives in an attempt to find out whether the observations by Chen (2006) and Xiao and Yue (2008) can also be generalized from specific genres to Mandarin Chinese in general.

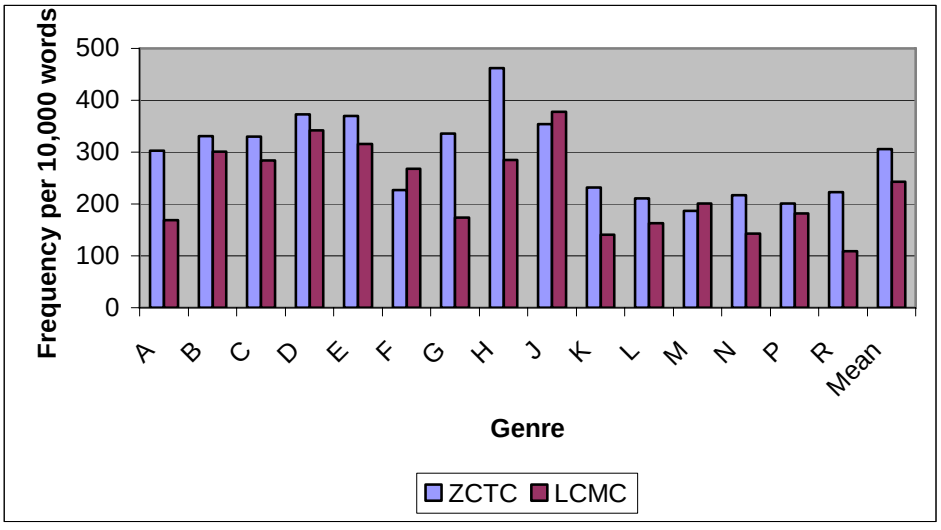

Figure 5. Normalized frequencies of conjunctions in ZCTC and LCMC

Figure 5 shows the normalized frequencies of conjunctions in the ZCTC and LCMC corpora. As

can be seen, the mean frequency of conjunctions is significantly greater in the translational corpus (306.42 instances per 10,000 tokens) than in the native (243.23) corpus (LL=723.12 for 1 d.f., $p<0.001$). However, a genre-based comparison reveals more subtleties. Genres of imaginative writing (five types of fiction K-P and humor R) generally demonstrate a significantly more frequent use of conjunctions in translational Chinese,[6] a finding which supports Xiao and Yue's (2008) observation of literary translation. Of expository writing, on the other hand, while connectives are considerably more frequent in most genres in translated Chinese (particularly reports and official documents H and press reportage A), there are also genres in which conjunctions are more common in native Chinese (namely popular lore F and academic prose J).

Xiao and Yue (2008) find that a substantially greater variety of frequent connectives are used in translated fiction in comparison with native Chinese fiction. While this finding is supported by ZCTC and LCMC, the two balanced corpora yield even more interesting results. Figure 6 compares the frequencies of conjunctions of different usage bands, as measured in terms of their proportion of the total numbers of tokens in their respective corpus of translational / native Chinese. As can be seen, more types of conjunctions of high frequency bands - i.e. with a proportion greater than 0.10% (7 and 4 types for ZCTC and LCMC respectively), 0.05% (13 and 7 types) and 0.01% (43 and 39 types) - are used in translational corpus. There are an equal number of conjunctions (56 types) with a proportion greater than 0.005% in translational and native corpora. After this balance point, the native corpus displays a greater number of less frequent conjunctions of the usage band 0.001% and below. This finding further confirms our earlier observation of the use of high and low frequency words in translated Chinese (cf. section 4.2). It also provides evidence that helps to extend the explicitation hypothesis from English to Chinese and to

generalize Chen (2006) and Xiao and Yue's (2008) observations from popular science translation and literary translation to the Mandarin language as a whole.

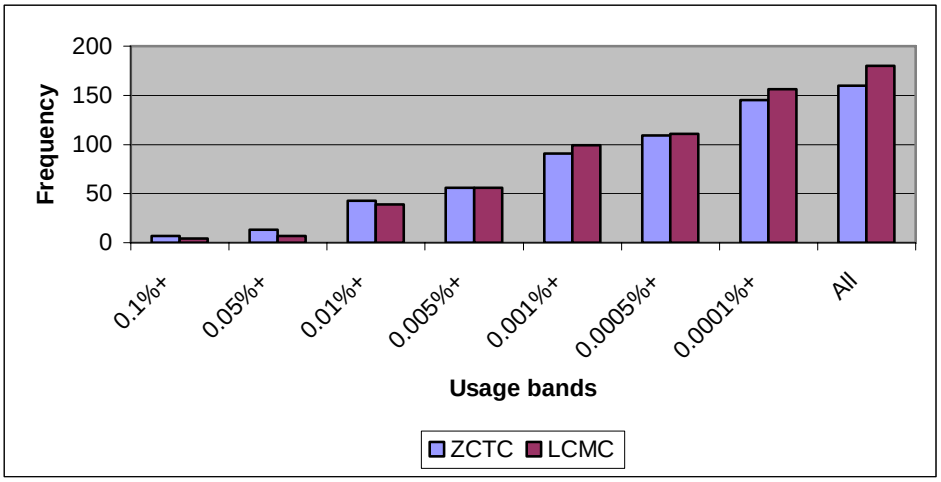

Figure 6. Distribution of conjunctions across usage bands

While the tendency to use conjunctions more frequently can be taken as a sign of explicitation, a closer comparison of the lists of conjunctions with a proportion of 0.001% in their respective corpus also sheds some new light on simplification. There are 91 and 99 types of conjunctions of this usage band. Of these, 86 items overlap in the two lists. Five conjunctions that appear on the ZCTC list but not on the LCMC list are all informal, colloquial, and simple, e.g. 以至于 'so…that…', 换句话说 'in other words', 虽说 'though', 总的来说 'in short', 一来 'first', which usually have more formal alternatives, e.g. 虽然 for 虽说 'though', and 总之 for 总的来说 'in a word'. In contrast, the 13 conjunctions that appear on the LCMC list but not on the ZCTC list are typically formal and archaic including, for example, 故 'hence', 可见 'it is thus clear', 进而 'and then', 加之 'in addition', 固然 'admittedly', 继而 'afterwards', 非但 'not only', 然 'nevertheless', and 尔后 'thereafter'. This appears to suggest that translators tend to use simpler forms than those used in

native language, thus providing evidence for the simplification hypothesis but against the normalization hypothesis.

4.4. Passives constructions

This section compares the distribution patterns of passive constructions in translational and native Chinese. While passives in Chinese can be marked lexically or syntactically (Xiao, McEnery and Qian 2006), we will only consider the 'default' passive form marked by *bei* (被), which is also the most important and frequent type of passive construction in Mandarin. Figure 7 shows the normalized frequencies of passives the fifteen genres as well as their mean frequencies in the ZCTC and LCMC corpora. As indicated by the mean frequencies, passives are more frequent in translational Chinese, and the log-likelihood (LL) test indicates that difference is statistically significant (LL=65.59 for 1 d.f., $p<0.001$, see Table 7). The figure also shows that there is considerable variability across genres. Table 7 gives the result of log-likelihood test for difference in each genre, with those significant results highlighted.

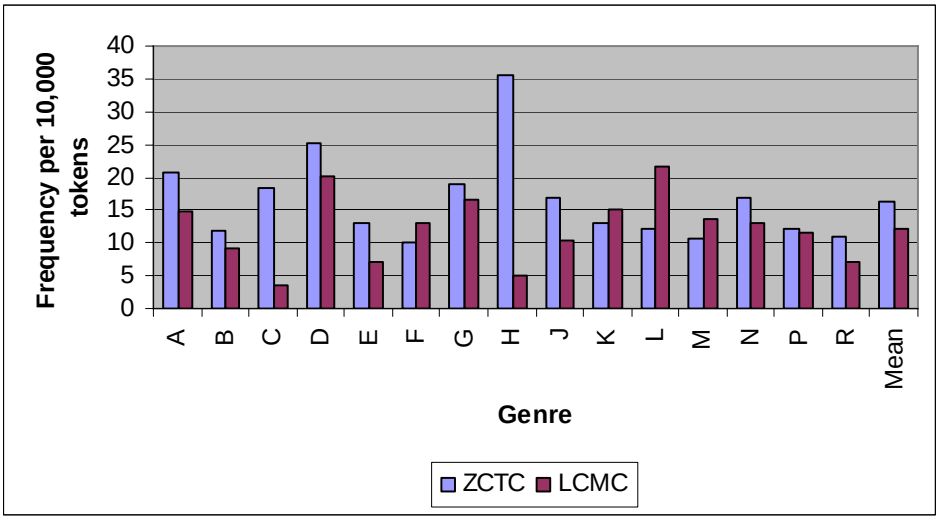

Figure 7. Distribution of passives in ZCZC and LCMC

Table 7. Log-likelihood tests for passives in ZCTC and LCMC

Genre	LL score	Significance level
A	**8.65**	**0.003**
B	1.83	0.176
C	**38.61**	**<0.001**
D	1.93	0.165
E	**13.29**	**<0.001**
F	3.17	0.075
G	2.16	0.142
H	**155.68**	**<0.001**
J	**27.75**	**<0.001**
K	0.88	0.347
L	**13.56**	**<0.001**
M	0.45	0.502
N	3.24	0.072
P	0.06	0.802
R	1.72	0.189
Mean	**69.59**	**<0.001**

A combined reading of Figure 7 and Table 7 reveals that in genres of expository writing such as press reportage (A), press reviews (C), skills / trade / hobbies (E), reports / official documents (H), and academic prose (J), passives are significantly more frequent in translational Chinese. The contrast is less marked in genres of imaginative writing (K-R). In imaginative writing, significant

difference is found only in the genre of mystery / detective fiction (L), where passives are significantly more common in native Chinese. The different distribution patterns of passives in translational and native Chinese provide evidence that translated Chinese is distinct from native Chinese. Such patterns are closely related to the different functions of passives in Chinese and English, the overwhelmingly dominant source language in our translational corpus (cf. section 3.1). In addition to a basic passive meaning, the primary function of passives in English is to mark an impersonal, objective and formal style whereas passives in Chinese are typically a pragmatic voice carrying a negative semantic prosody (Xiao, McEnery and Qian 2006: 143-144). Since mystery / detective fiction is largely concerned with victims who suffer from various kinds of mishaps and the attentions of criminals, it is hardly surprising to find that the inflictive voice is more common in this genre in native Chinese. On the other hand, expository genres like reports / official documents (H), and press reviews (C), and academic prose (J), where the most marked contrast is found between translational and native Chinese, are all genres of formal writing that make greater use of passives in English. When texts of such genres are translated into Chinese, passives tend to be overused; that is, native speakers of Chinese would not normally use the passive when they express similar ideas. For example, the translated example 该证书就必须被颁发 (ZCTC_H) is clearly a direct translation of the English passive *Then the certificate must be issued*. In such cases, a native speaker of Mandarin is very likely to use the so-called unmarked 'notational passive', i.e. the passive without a passive marker, which is very common in Chinese, as in 该证书就必须颁发. It is presently not clear to what extent translated Chinese is affected by the translation process, which is part of our future investigation based on parallel corpus research in our project. However, available evidence of this kind does suggest that normalization may not

be a universal feature of translational language (cf. section 2.3).

5. Conclusions

This article first provided a review of the state of the art of research in the so-called translation universals, namely the characteristic features of translational language. The limitations of the previous research in this area as revealed in our review led to the discussion of our proposal for a new project specifically designed to overcome such limitations. We also presented a new balanced corpus of translational Chinese created on this project which, together with a comparable corpus of native Chinese, provided a quantitative basis for our case studies of some lexical and syntactic features of translational Chinese.

Our case studies have shown that Laviosa's (1998b) observations of the core patterns of lexical features of translational English are supported by our monolingual comparable corpora of translational and native Chinese. Translational Chinese has a significantly lower lexical density (i.e. the proportion of lexical words) than native Chinese, but there is no significant difference in the lexical density as defined by the standardized type-token ratio. In relation to native Chinese, translated Chinese has a relatively low proportion of lexical words over function words, a higher proportion of high-frequency words over low-frequency words, and a greater repetition rate of high frequency words. Beyond the lexical level, our data shows that the mean sentence length is sensitive to genres and may not be a reliable indicator of simplification, but a comparison of frequent connectives in native and translational Chinese corpora appears to suggest simpler forms tend to be used in translations. In spite of some genre-based subtleties, translational Chinese also uses connectives more frequently than native Chinese, which provides evidence in favor of the explicitation hypothesis. Our analysis of passives in the two corpora provides further evidence

supporting the previous finding that translational language is affected by the translation process, though the extent of such influence is yet to be investigated. The source-induced difference between translational and native Chinese in their use of passives also suggests that normalization may be language specific and does not apply in translational Chinese.

Finally, we believe that the newly created *ZJU Corpus of Translational Chinese* (ZCTC) will play a leading role in the study of translational Chinese by producing more empirical evidence, and it is our hope that the study of translational Chinese will help to address limitations of imbalance in the current state of translation universal research.

Acknowledgements

We are grateful to the China National Foundation of Social Sciences for supporting our project 'A corpus-based quantitative study of translational Chinese in English-Chinese translation' (Grant Reference 07BYY011). We also thank our postgraduate assistants Wei Huang, Wenying Hu, Shulin Yu, and Shangchao Min for their help in data collection.

Notes

1. Readers are reminded of this modification when they interpret the results based on a comparison of the LCMC and ZCTC corpora. Those who are interested in potential change during this decade in Mandarin Chinese are advised to use the *UCLA Written Chinese Corpus* (http://www.lancs.ac.uk/fass/projects/corpus/UCLA/), which models LCMC but samples texts one decade apart.

2. The number of tokens given here for the LCMC corpus may be different from earlier releases, because this edition of LCMC has been retagged using ICTCLAS2008, which was used to tag the ZCTC corpus (see section 3.2).
3. See the official website of the ZCTC corpus (www.lancs.ac.uk/fass/projects/corpus/ZCTC/) for the full part-of-speech tagset as applied on the corpus.
4. See the official website of ICTCLAS (www.ictclas.org) for the history and test results of the software tool. In order to ensure maximum comparability between translated and native Chinese corpora, a new version of the LCMC corpus has also been produced for use on our project, which is retagged using this same tool.
5. In this study, we follow Xiao, Rayson and McEnery (2008) in treating adjectives (including non-predicate noun modifiers and descriptive adjectives), adverbs, nouns, and verbs as lexical words. Function words include the following POS categories: auxiliaries, classifiers, conjunctions, interjections, numerals and quantifiers, onomatopoeias, particles, place words, prefixes, pronouns, prepositions, space words, suffixes, and time words. Unclassified words and symbols and punctuations are excluded in our computations.
6. Note that the difference in science fiction (M) is not significant (LL=0.641 for 1 d.f., p=0.423).

References

Aijmer, K. (2007), 'Translating discourse markers: A case of complex translation', in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 95-116. Clevedon: Multilingual Matters.

Baker, M. (1993), 'Corpus linguistics and Translation Studies: Implications and applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honor of John*

Sinclair, 233-250. Amsterdam: John Benjamins.

Baker, M. (1995), 'Corpora in translation studies: An overview and some suggestions for future research.' *Target* 7(2): 223-243.

Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Blum-Kulka, S. (1986), 'Shifts of cohesion and coherence in Translation', in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, 17-35. Tübingen: Gunter Narr.

Blum-Kulka, S. and Levenston, E. (1983), 'Universals of lexical simplification', in C. Faerch and G. Kasper (eds.) *Strategies in Interlanguage Communication*, 119-139. London: Longman.

Chen, W. (2006), *Explication Through the Use of Connectives in Translated Chinese: A Corpus-based Study*. PhD thesis, University of Manchester.

Cheong, H. (2006), 'Target text contraction in English-into-Korean Translations: A contradiction of presumed translation universals?' *Meta* 51(2): 343-367.

Chesterman, A. (2004), 'Beyond the particular', in A. Mauranen and P. Kuyamaki (eds.) *Translation Universals: Do they exist?* 33-49. Amsterdam: John Benjamins.

Frawley, W. (1984), 'Prolegomenon to a theory of translation', in W. Frawley (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives*, 159-175. London: Associated University Press.

Gellerstam, M. (1996), 'Translations as a source for cross-linguistic studies', in K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, March 1994*, 53-62. Lund: Lund University Press.

Hartmann, R. (1985), 'Contrastive textology.' *Language and Communication* 5: 107-110.

Hundt, M., Sand, A. and Siemund, R. (1998), *Manual of Information to Accompany the Freiburg-LOB Corpus of British English*. Freiburg: University of Freiburg.

Hundt, M., Sand, A. and Skandera, P. (1999), *Manual of Information to Accompany the Freiburg-Brown Corpus of American English*. Freiburg: University of Freiburg.

Jantunen, J. (2001), 'Synonymity and lexical simplification in translations: A corpus-based approach.' *Across Languages and Cultures* 2(1): 97-112.

Jantunen, J. (2004), 'Untypical patterns in translations. Issues on corpus methodology and synonymity', in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator,* 101-126. Clevedon: Multilingual Matters.

Kenny, D. (1998), 'Creatures of habit? What translators usually do with words.' *Meta* 43(4): 515-523.

Kenny, D. (1999), 'The German-English parallel corpus of literary texts (GEPCOLT): A resource for translation scholars.' *Teanga* 18: 25-42.

Kenny, D. (2000), 'Translators at play: exploitations of collocational norms in German-English translation', in B. Dodd (ed.) *Working with German Corpora*, 143-160. Birmingham: University of Birmingham Press.

Kenny, D. (2001), *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester: St. Jerome Publishing.

Laviosa, S. (1997), 'How comparable can "comparable corpora" be?' *Target* 9(2): 289-319.

Laviosa, S. (1998a), 'The corpus-based approach: A new paradigm in translation studies.' *Meta* 43(4): 474-479.

Laviosa, S. (1998b), 'Core patterns of lexical use in a comparable corpus of English narrative prose.' *Meta* 43(4): 557-570.

Laviosa, S. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Laviosa-Braithwaite, S. (1996), *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. PhD Thesis, University of Manchester.

Laviosa-Braithwaithe, S. (1997), 'Investigating simplification in an English comparable corpus of newspaper articles', in K. Klaudy and J. Kohn (eds.) *Transferre necesse est. Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting*, 531-540. Budapest: Scholastica.

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*, 151-162. Amsterdam: John Benjamins.

Mauranen, A. (2000), 'Strange strings in translated language: A study on corpora', in M. Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, 119-141. Manchester: St. Jerome Publishing.

Mauranen, A. (2007), 'Universal tendencies in translation', in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 32-48. Clevedon: Multilingual Matters.

Mauranen, A. and Kujamäki, P. (2004), *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.

McEnery, T. and Wilson, A. (2001), *Corpus Linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T. and Xiao, R. (2002), 'Domains, text types, aspect marking and English-Chinese translation.' *Languages in Contrast* 2(2): 211-229.

McEnery, T. and Xiao, R. (2004), 'The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study', in M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, 1175-1178. Lisbon, 24-30 May 2004.

McEnery, T. and Xiao, R. (2007), 'Parallel and comparable corpora: What is happening?', in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 18-31. Clevedon: Multilingual Matters.

McEnery, T., Xiao, R. and Mo, L. (2003), 'Aspect marking in English and Chinese.' *Literary and Linguistic Computing* 18(4): 361-378.

McEnery, T., Xiao, R. and Tono, Y. (2006), *Corpus-based Language Studies*: *An Advanced Resource Book*. London and New York: Routledge.

Nevalainen, S. (2005), 'Köyhtyykö kieli käännettäessä? Mitätaajuuslistat kertovat suomennosten sanastosta', in A. Mauranen and J. Jantunen (eds.) *Käännössuomeksi*, 141- 162. Tampere: Tampere University Press.

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

Olohan, M. and 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): 557-570.

Pym, A. (2005), 'Explaining explicitation', in K. Károly and Á. Fóris (eds.) *New Trends in Translation Studies*, 29-43. Budapest: Akadémiai Kiadó.

Scott, M. (2004), *The WordSmith Tools* (v. 4.0). Oxford: Oxford University Press.

Stubbs, M. (1986), ‘Lexical density: A computational technique and some findings’, in M. Coultard (ed.) *Talking about Text. Studies Presented to David Brazil on His Retirement.* Birmingham: English Language Research, University of Birmingham.

Stubbs, M. (1996), *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. London: Blackwell

Teubert, W. (1996), ‘Comparable or parallel corpora?’ *International Journal of Lexicography* 9(3): 238-64.

Tirkkonen-Condit, S. (2002), ‘Translationese – A myth or an empirical fact? A study into the linguistic identifiability of translated language.’ *Target* 14(2): 207-220.

Tirkkonen-Condit, S. (2005), ‘Do unique items make themselves scarce in translated Finnish?’, in K. Károly and Á. Fóris (eds.) *New Trends in Translation Studies. In Honor of Kinga Klaudy*, 177-189. Budapest: Akadémiai Kiadó.

Toury, G. (1995), *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Toury, G. (2004), ‘Probabilistic explanations in translation studies. Welcome as they are, would they qualify as universals?’, in A. Mauranen and P. Kuyamaki (eds.) *Translation Universals: Do they exist?* 15-32. Amsterdam: John Benjamins.

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

Váradi, T. (2007), ‘NP modification structures in parallel corpora’, in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 168-186. Clevedon: Multilingual Matters.

Xiao, R. (2008), ‘Using an enhanced MDA model in study of world Englishes.’ Paper presented at the Fourth Inter-Varietal Applied Corpus Studies (IVACS) Conference. University of Limerick,

13-14 June 2008.

Xiao, R., McEnery, T., Baker, P. and Hardie, A. (2004), 'Developing Asian language corpora: Standards and practice', In *Proceedings of the 4th Workshop on Asian Language Resources*, 1-8. Sanya, Hainan Island, March 25, 2004.

Xiao, R., McEnery, T. and Qian, Y. (2006), 'Passive constructions in English and Chinese: A corpus-based contrastive study.' *Languages in Contrast* 6(1): 109-149.

Xiao, R., Rayson, P. and McEnery, A. (2008), *A Frequency of Mandarin Chinese: Core Vocabulary for Learners*. London and New York: Routledge.

Xiao, R. and Yue, M. (2008) 'Using corpora in Translation Studies: The state of the art', in P. Baker (ed.) *Contemporary Approaches to Corpus Linguistics*. London: Continuum.