Idioms, word clusters, and reformulation markers in translational Chinese: Can "translation universals" survive in Mandarin?

Richard Xiao

Edge Hill University

Abstract: This article is concerned with three linguistic features which have so far been rarely investigated in translation studies – namely idioms, word clusters and reformulation markers, in translational Chinese as represented in a one-million-word balanced corpus of translated Chinese texts in comparison with native Mandarin represented in a comparable corpus of non-translated Chinese texts. Our results show that idioms are more commonly used in native Chinese, meaning that the distribution patterns of idioms tend to be language-specific whereas word clusters are substantially more prevalent in translated Chinese, suggesting a tendency in translation to use fixed and semi-fixed recurring patterns in an attempt to achieve improved fluency. Reformulation markers function as a strategy for explicitation in Chinese translations, which tend to use informal, stylistically simpler forms than native Chinese texts.

1. Introduction

An important area of corpus-based translation studies has been translation universal (TU) research, which investigates the common features of translational language. The term 'translation universal' is, however, not without controversy. Gaspari and Bernardini (2010), for example, argue that translation universal might as well be called "mediation universal" because some features of translated language are found to be present in non-native language, both of which are mediated discourses. This argument echoes Granger's (1996: 48)

observation of the similarity between what she calls “translationese” and “learnerese”. Before further evidence is uncovered for the link between mediated discourses such as translational and non-native languages, however, we will follow the more conventional term for the purpose of the present study.

A number of linguistic features of translated texts have been observed, mainly on the basis of translated English, at lexical, syntactic and discourse level, which have motivated the formulation of TU hypotheses such as normalization, simplification, explicitation, sanitization, and levelling out/convergence. Simplification refers to the “tendency to simplify the language used in translation” (Baker 1996: 181-182), and as a result translated language is simpler than the target native language lexically, syntactically and/or stylistically. Normalization suggests that translational language displays a “tendency to exaggerate features of the target language and to conform to its typical patterns” so that translated texts are more “normal” than non-translated texts (Baker 1996: 183). Explicitation is manifested by the tendency in translations to “spell things out rather than leave them implicit” through more frequent use of connectives and increased cohesion (Baker 1996: 180). Sanitization means that translated texts, with lost or reduced connotational meaning, are “somewhat ‘sanitized’ versions of the original” (Kenny 1998: 515). Levelling out refers to “the tendency of translated text to gravitate towards the centre of a continuum” (Baker 1996: 184), which is also known as “convergence”, that is, the “relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features” (Laviosa 2002: 72). We will not review these TU hypotheses in great depth here. Interested readers are advised to refer to Xiao and Yue (2009), which provides a comprehensive review of the state of the art of corpus-based translation studies, including TU research.

This article is concerned with three linguistic features which have so far been rarely investigated in translation studies – namely idioms, word clusters and reformulation markers,

in translational Chinese as represented in a one-million-word balanced corpus of translated Chinese texts in comparison with native Mandarin represented in a comparable corpus of non-translated Chinese texts. These features have been chosen in this study because on the one hand, idioms and word clusters are fixed or semi-fixed lexical phrases closely associated with idiomaticity and fluency, which are a "preferred strategy" that translators tend to adopt according to the translation universal hypothesis of normalization (Baker 2004: 182), while on the other hand, reformulation markers such as *that is to say* contribute substantially to making messages more explicit.

In this article, we will review previous translation studies of the three features under investigation and then present the corpora and tools used in this study, on the basis of which quantitative and qualitative analyses will be undertaken to compare the use of idioms, word clusters, and reformulation markers in two matching corpora of translated and native Mandarin Chinese. The implications of our findings for TU hypotheses will also be discussed.

2. Idioms, word clusters and reformulation markers in translation studies

While idioms are pervasive in language use, there is unfortunately no universally agreed definition of the term. Baker (2007: 14) cites the following definition from the *Oxford English Dictionary* (1989), which she thinks is adequate: "A form of expression, grammatical construction, phrase, etc., peculiar to a language; a peculiarity of phraseology approved by the usage of the language, and often having a significance other than its grammatical or logical one." In practice, what Baker (2004, 2007) studies are "pre-packaged, recurring stretches of language," which might as well be used as a more operable definition of the term. Idiom in this sense also fits in well with Sinclair's (1991) "idiom principle", which operates in combination with the "open choice principle" to mirror the distinction between conventionality and flexibility in language use. Of the two, the principle of idiom plays the

central role in speech and writing, relying heavily on the speaker or writer's large inventory of prefabricated lexico-grammatical chunks at their disposal (cf. also McCarthy and Carter 2004).

Idioms in this study are broadly defined. They are similar to fixed and semi-fixed formulaic expressions based on collocations which are known as "word clusters", "lexical bundles", "multiword units", "prefabs", and "n-grams" and so on. However, the demarcation line between idioms and word clusters is actually fuzzy. Idioms in a narrow sense, that is, those characterized with a high degree of structural fixedness and semantic opacity, can be regarded as an "extreme example" of word clusters (Scott 2009: 286), as "all words have a tendency to cluster together with some others". On the other hand, there is an important difference between idioms and word clusters. While an idiom is a complete unit of meaning, whether literal or figurative, a word cluster may be complete or incomplete in meaning. Word clusters are purely structurally defined on the basis of co-occurrences with no regard to their semantic contents.

The distinction between the broad and narrow senses of idioms can also be found in Chinese. Idioms in Chinese are a complicated category commonly known as 熟语 *shuyu* ('familiar expression'). They refer to fused phrases or expressions recurring in language use such as 成语 *chengyu* ('idiomatic expression', typically composed of four Chinese characters), 习语 *xiyu* or *xiyongyu* ('conventional expression'), 惯用语 *guanyongyu* ('habitually used expression'), and 俗语 *suyu* ('common saying'). Although the Chinese term *chengyu* is often translated as 'idiom' in English, it only refers to a type of narrow-sense idioms in Chinese. *Chengyu* are conventionally used set phrases, which are historically allusive in origin, often highly fixed in structure (i.e. the four-character-mould), usually opaque in meaning and typically archaic in style (see Wu 1995 for a review of various definitions of *chengyu*). In relation to *chengyu*, fixed or semi-fixed phrases and expressions

which are highly frequent in language use but have a short history and are thus not historically allusive are often called *xiyu*, which can equally opaque in meaning (cf. An, Liu and Hou 2004). *Guanyongyu* ('habitually used expression') refer to recurring fused phrases or expressions which are usually transparent in meaning while *suyu* ('common saying') are similar except that they are more colloquial in style. Except for the narrow-sense idioms *chengyu*, Chinese *shuyu* ('familiar expression') of other types discussed here are broad-sense idioms that vary in structural fixedness, semantic opacity as well as in style.

It is clear from the above discussion that idioms in Chinese are more complex as a linguistic category than English idioms. As idioms are culturally rooted, they also embody different cultural traits such as historical backgrounds, natural environments, religious beliefs and world views (cf. Yang 2004). Nevertheless, idioms in English and Chinese are similar to each other in that they are both pre-packed, recurring formulaic expressions that help to achieve idiomaticity in their respective language.

According to Fernando (1996), idioms can be pure, semi- or literal idioms in terms of their idiomaticity while Halliday (2000) classifies idioms into ideational, interpersonal and relational types on the basis of their functions. Clearly, although idioms can only occur as one sentential constituent because of their holistic form and meaning, they can nevertheless play a number of roles in discourse and have numerous discourse functions. Such complexities, coupled with the pervasiveness and cultural specificity of idioms as well as the cultural diversity associated with language use, constitute a challenge in translation which translators must cope with if they are to translate idioms in the source language into appropriate idioms in the target language. In spite of their importance, however, idioms seem to have rarely been studied in translation research, with the exception of Baker (1992, 2007).

Baker (2007) studies the use of idioms in translated English in comparison with native English on the basis of the fiction and biography components of the Translational English

Corpus (TEC, see Baker 2004) and a comparable set of fiction samples from the British National Corpus (BNC, see Aston and Burnard 1998). Baker (2007: 14) assumes, on the basis of the normalization hypothesis, that "translators are likely to opt for safe, typical patterns of the target language and shy away from creative or playful uses", and consequently, "translators ought to be making heavy use of idioms, in the broad sense of pre-packed, recurring stretches of language." On the other hand, as idioms, especially those which are highly opaque in meaning (e.g. *chew the fat*), "tend to be highly informal in flavour", they are therefore expected to be avoided in translations, which "generally tend to be characterised by a higher level of formality than non-translations." These observations point in two opposite directions. On the one hand, translations are expected to make heavier use of idioms to confirm to the target language norm while on the other hand, idioms, and opaque idioms in particular, are expected to be avoided in translations because of their informal flavour. Unfortunately, Baker (2007) only gives some examples (*off the hook, out of order*) to show that opaque idioms are more likely to be avoided in translations than in non-translated texts, but does not provide statistics of the overall proportions of literal and opaque idioms in the translational versus native English data.

Idioms which are opaque in meaning also tend be structurally tight whereas literal idioms are more likely to be structurally loose. They correspond to the narrow and broad senses of idioms. Unless the corpus used is annotated semantically for the two different senses of idioms, Baker's (2007) practice of using a few selected examples is probably the only feasible way of studying opaque idioms in a large corpus. In contrast, idioms in broad sense based on their collocational behaviour are easier to study because corpus exploration tools (e.g. WordSmith, see Scott 2009) are available for computing word clusters (or called 'lexical bundles', 'multiword unit' or 'n-grams' in the literature).

Generally speaking, the frequency of word clusters tends to drop sharply as their length grows. For example, the frequency of 4-word clusters is significantly lower than that of 3-word clusters, which are in turn substantially less frequent than 2-word clusters. The statistical significance of word clusters is usually measured by their recurring rate, e.g. 5 or 10 occurrences in a million words. In addition, the dispersion or coverage rate can be used in combination with the recurring rate to avoid extracting word clusters which are frequent in only a few texts in a corpus. In the present study, we use the default settings of the WordSmith Tools (5.0), that is, a minimum frequency of 5 and a maximum coverage of 10%.

While word clusters may not necessarily be complete in structure or meaning, they are nevertheless of great importance in language studies. Word clusters have recently been investigated in areas such as genre analysis and language teaching (e.g. Granger 1998; De Cock 1998, 2000; Cortes 2002; Biber, Conrad and Cortes 2004; Biber 2006). In contrast, word clusters have rarely been researched in translation studies, with the exceptions of Baker (2004) and Nevalainen (2005, cited in Mauranen 2007). Both of them find that recurring word clusters are more common in translations in comparison with non-translated texts. This finding echoes Baroni and Bernardini's (2003: 379) observations based on their investigation of collocations in translated and native texts, which even differentiate between two types of repetition patterns:

> […] translated language is repetitive, possibly more repetitive than original language. Yet the two differ in what they tend to repeat: translations show a tendency to repeat structural patterns and strongly topic-dependent sequences, whereas originals show a higher incidence of topic-independent sequences, i.e. the more usual lexicalised collocations in the language.

A particular type of idioms or word clusters in Baker (2004, 2007) is the so-called reformulation markers such as *in other words* and t*hat is to say*, though a reformulation marker can also be a single word instead of a word cluster (e.g. *namely*). Reformulation markers are a kind of discourse markers which function to enhance connectivity in discourse (Schourup 1999: 230). Murillo (2004: 2066) calls them "markers of the explicit" as these discourse markers "assist, to varying degrees, in the inferential process by making explicit reference assignment, disambiguation, further enrichment and elliptic material in connection with the recovery of the propositional form." Murillo (2004) observes, from the viewpoint of Relevance Theory (Sperber and Wilson 1995), that reformulation markers not only function to recover the propositional form of an utterance, but they also operate in relation to its explicatures and implicatures "by explicitating implicated premises and conclusions" (2004: 2066).

The glossing and explicating functions of reformulation markers render them particularly relevant to the explicitation hypothesis in translation universal research. For example, Baker (2004) finds that reformulation markers such as *that is*, *that is to say*, and *in other words* are substantially more frequent in the fiction and biography components of the TEC corpus than the fiction subcorpus in the BNC. Mutesayire (2005) views the higher frequency of reformulation markers in translated English as evidence of explicitation. In the same vein, Chen (2006: 152) compares the distribution of similar Chinese reformulation markers in a corpus of translated popular science books and the science section of the Sinica corpus which represents native Mandarin Chinese as used in Taiwan.[1] He finds that reformulation markers are more common in translated Chinese, which supports the explicitation hypothesis.

These translation studies of idioms, word clusters, and reformulation markers have uncovered some interesting features of translated English, and in the case of Chen (2006), of

translated Chinese. Or to be more precise, they reveal some features in translations that might be characteristic of specific genres such as fiction and biography (as in Baker 2007) or popular science writing (as in Chen 2006). Biber (1995: 278) notes that language can vary substantially across genres while Xiao (2009) demonstrates that the genre of scientific writing is the least diversified of all genres across various varieties of English. This means that what has been observed of idioms, word clusters and reformulation markers in the studies cited above might be specific to particular genres rather than applicable to translational English or translational Chinese as a whole.

More importantly, it is debatable whether the features uncovered on the basis of translational English can be generalized to other translated languages. Existing evidence has largely come from translational English and related European languages. If such features are to be generalized as "translational universals", the languages involved must not be restricted to English and closely related languages. Clearly, evidence from "genetically" distinct languages such as English and Chinese is undoubtedly more convincing, if not indispensable.

In the present study, we will use two comparable balanced corpora of translational and native Chinese to verify whether the above English-based, genre-specific features of translations can be generalized to Mandarin Chinese in general.

3. The corpora and tools

Two comparable monolingual corpora are used in this study, namely the *Lancaster Corpus of Mandarin Chinese* (LCMC) and the *ZJU Corpus of Translational Chinese* (ZCTC), which represent native and translational Chinese respectively. LCMC is designed as a Chinese match for the FLOB corpus of British English (Hundt et al 1998) and the Frown corpus of American English (Hundt et al 1999) for use in cross-linguistic contrast of English and Chinese (McEnery and Xiao 2004), while ZCTC is created as a translational counterpart of

LCMC with the explicit aim of studying features of translated Chinese (Xiao, He and Yue 2010).

Table 1. The genres covered in LCMC and ZCTC

Code	Genre	LCMC & ZCTC		LCMC		ZCTC	
		Samples	Percent	Tokens	Percent	Tokens	Percent
A	Press reportage	44	8.8	89,367	8.73	88,196	8.67
B	Press editorials	27	5.4	54,595	5.33	54,171	5.32
C	Press reviews	17	3.4	34,518	3.37	34,100	3.35
D	Religious writing	17	3.4	35,365	3.46	35,139	3.45
E	Instructional writing	38	7.6	77,641	7.59	76,681	7.54
F	Popular lore	44	8.8	89,967	8.79	89,675	8.81
G	Biographies and essays	77	15.4	156,564	15.30	155,601	15.29
H	Reports/official documents	30	6	61,140	5.97	60,352	5.93
J	Academic prose	80	16	163,006	15.93	164,602	16.18
K	General fiction	29	5.8	60,357	5.90	60,540	5.95
L	Mystery and detective fiction	24	4.8	49,434	4.83	48,924	4.81
M	Science fiction	6	1.2	12,539	1.23	12,267	1.21
N	Adventure fiction	29	5.8	60,398	5.90	59,042	5.80
P	Romantic fiction	29	5.8	59,851	5.85	59,033	5.80
R	Humour	9	1.8	18,645	1.82	19,072	1.87
Total		500	100	1,023,387	100.00	100.00	100.00

The two monolingual Chinese corpora are each composed of one million words in five hundred 2,000-word text samples which are taken proportionally from fifteen text categories published in China in the 1990s as shown in Table 1, which also gives the actual numbers of tokens in different genres as well as their corresponding proportions across genres in the ZCTC and LCMC corpora. As can be seen, the two corpora are roughly comparable in terms of both overall size and proportions for different genres. 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.

Both corpora are annotated with word class information using the same tool to ensure comparability (see Xiao et al 2010 for details). They are marked up in XML and encoded in Unicode, applying the Unicode Transformation Format 8-Bit (UTF-8) to facilitate cross-platform operations and data interchange.

The two comparable corpora of Mandarin Chinese will provide a reliable basis for the quantitative and qualitative analyses of idioms, word clusters and reformulation markers to be presented in the following section. A third corpus, which is an English-to-Chinese parallel corpus to be introduced in section 6, is also used to investigate the extent to which reformulation markers are explicated in the translation process.

Two publicly available, XML-aware and Unicode-compliant corpus exploration tools are used to explore the monolingual Chinese corpora. They are Xaira (i.e. XML Aware Indexing and Retrieval Architecture, Burnard and Todd 2003; see Xiao 2006 for a review) and Wordsmith 5.0 (Scott 2009), which are used for distribution analysis word cluster analysis respectively. A parallel concordancer, ParaConc (Barlow 1995), is used to search the English-Chinese parallel corpus.

4. Idioms

Having reviewed previous research of idioms, word clusters and reformulation markers in translation studies and presented our corpora and tools, we will explore these linguistic features in translated Chinese in comparison with native Chinese on the basis of two comparable balanced corpora. We will first examine idioms.

In the LCMC and ZCTC corpora, idioms are tagged according to their word classes: *nl* for nominal idioms, *vl* for verbal idioms, *al* for adjectival idioms, *dl* for adverbial idioms, and *bl* for nominal modifying idioms (cf. Liu et al 2008). Hence it is quite straightforward to extract idioms from our corpus data with the help of these tags. Figure 1 shows the normalized frequencies (per 100,000 words) of idioms across the fifteen genres covered in the LCMC corpus of native Chinese and the ZCTC corpus of translated Chinese. As can be seen, with a few exceptions (e.g. E, L, and N), idioms in the native corpus LCMC are considerably higher than in the translational corpus ZCTC.

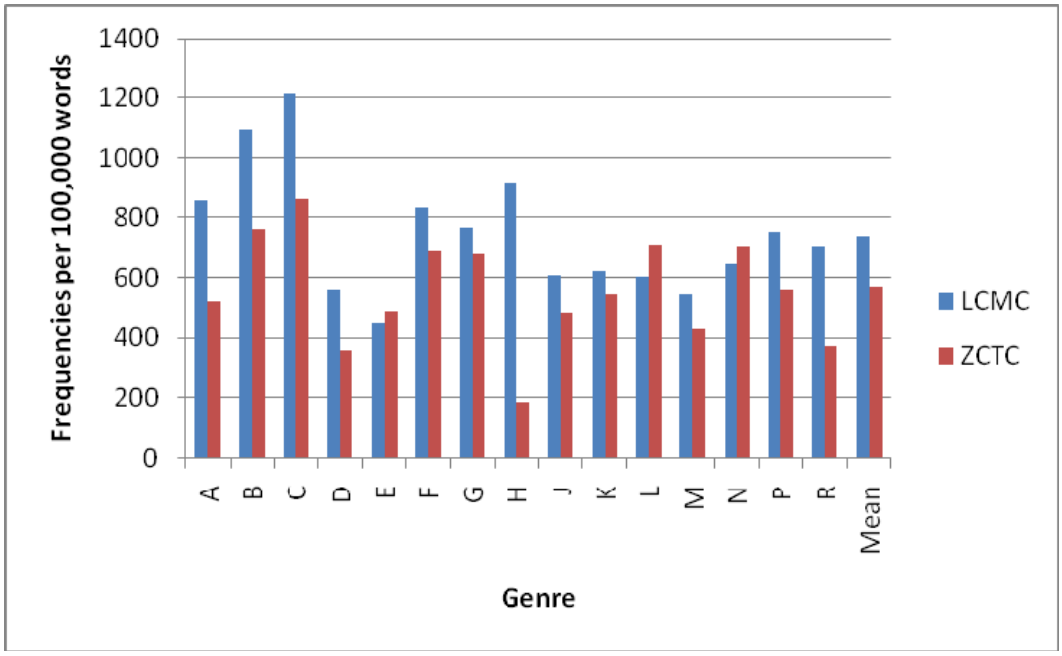

Figure 1. Normalized frequencies of idioms in LCMC and ZCTC

Table 2 gives the raw frequencies of idioms in the two corpora, which are used in log-likelihood (LL) tests to measure the statistical significance of the difference in the frequencies between the two corpora. For a difference to be statistically significant, the LL score must be greater than 3.84 (1 degree of freedom) for a significance level less than 0.05, which means that we can be more than 95% confident that the observed difference is not due

to chance. It can be seen in the table that translated Chinese only displays a marginally higher use of idioms in the genre of mystery and detective stories (L), which is statistically significant (LL=5.41, p=0.02), while the differences in the genres of instructional writing (E) and adventure fiction (N) have no statistical significance. In all other genres, idioms are more frequently used in the native corpus LCMC, though the differences in general fiction (K) and science fiction (M) are not significant. The table also shows that when the two corpora are taken as a whole, the overall frequency of idioms in LCMC (7,979 occurrences) is significantly higher than that in ZCTC (6,265 occurrences), with an LL score of 196.81 and a significance level less than 0.001.

Table 2. Raw frequencies of idioms in LCMC and ZCTC with significance tests

Genre	LCMC	ZCTC	LL score	Significance
A	809	499	72.37	0.000
B	632	457	27.23	0.000
C	443	320	19.20	0.000
D	209	135	15.61	0.000
E	368	405	2.00	0.158
F	794	666	10.50	0.001
G	1,272	1,141	6.37	0.012
H	590	121	334.23	0.000
J	1,051	858	18.43	0.000
K	398	355	2.21	0.137
L	315	374	5.41	0.020
M	72	57	1.66	0.197
N	412	447	1.64	0.200
P	475	354	17.02	0.000
R	139	76	18.37	0.000
Total	7,979	6,265	196.81	0.000

In section 2 we noted Baker's (2007) observations of the two conflicting tendencies in using idioms in translations. On the one hand, idioms are supposed to be used heavily in translations to conform to the norm of the target language while on the other hand idioms, especially those characterized with a high degree of opacity, are expected to be avoided in translations because of their informal tone. Clearly, our Chinese data supports the second tendency. This is because Chinese idioms are different from their English counterparts in terms of their formation and etymological sources (see section 2). Many Chinese idioms, especially the so-called *chengyu*, have allusive stories from ancient times behind them, which render them highly opaque, with their actual meaning different from their surface meaning. As a result, Chinese idioms other than those called *suyu* ('common saying') tend to carry a formal tone and sometimes an archaic flavour.[2] In contrast, English idioms, especially those with an opaque meaning (e.g. *kick the bucket*), tend to have an informal flavour of slangs. Such cross-linguistic differences indicate that idioms can have different distribution patterns in different languages, and language-specific features of idioms determine that the high use of idioms is unlikely to be a universal feature of translational language. The substantially more common use of idioms in native than translated Chinese also suggests that the TU hypothesis of normalization is unsupported in our corpora.

5. Word clusters

Now we will examine word clusters at the loose end of the idiomatic continuum. Previous research has suggested that word clusters tend to be more commonly used in translations. Is this also true in Chinese? In this article, we will study word clusters composed of 2-to-6 words because clusters comprising more than six words are quite rare in million-word corpora like LCMC and ZCTC. Table 3 gives their frequencies in the two Chinese corpora together with the results of log-likelihood tests. As can be seen, word clusters of all types are

much more frequent in the translational corpus ZCTC than in the native corpus LCMC, and the differences are all statistically significant as indicated by their LL scores and significance levels. The higher use of word clusters in the translational corpus is also evidenced by a keyword cluster analysis,[3] which shows that for 3-to-5-word clusters, 123 clusters are significantly more common in ZCTC as opposed to just one such cluster which is significantly more frequent in LCMC; and for 2-to-6-word clusters, 958 clusters are significantly more common in ZCTC in contrast to 59 such clusters which are significantly more frequent in LCMC.

Table 3. Word clusters in LCMC and ZCTC

Word clusters	LCMC	ZCTC	LL score	Significance
2-word clusters	21002	23006	103.44	0.000
3-word clusters	4015	5523	248.36	0.000
4-word clusters	580	732	16.58	0.000
5-word clusters	160	197	4.06	0.044
6-word clusters	70	105	7.25	0.007

In addition to their significantly higher frequencies in translational Chinese, word clusters demonstrate two other interesting characteristics. On the one hand, high-frequency word clusters (defined here as those accounting for at least 0.01% of the respective corpus) are more common in Chinese translations. As can be seen in Figure 2, the number of high-frequency word clusters in ZCTC (a total of 413, including 403 2-word clusters and ten 3-word clusters) is greater than that in LCMC (a total of 291, including 287 2-word clusters and four 3-word clusters), which is a statistically significant difference (LL=21.96, 1 degree of freedom, $p<0.001$). Given that translated Chinese tends to use high-frequency words (Xiao 2010), it is hardly surprising to find a more common use of high-frequency word clusters in ZCTC.

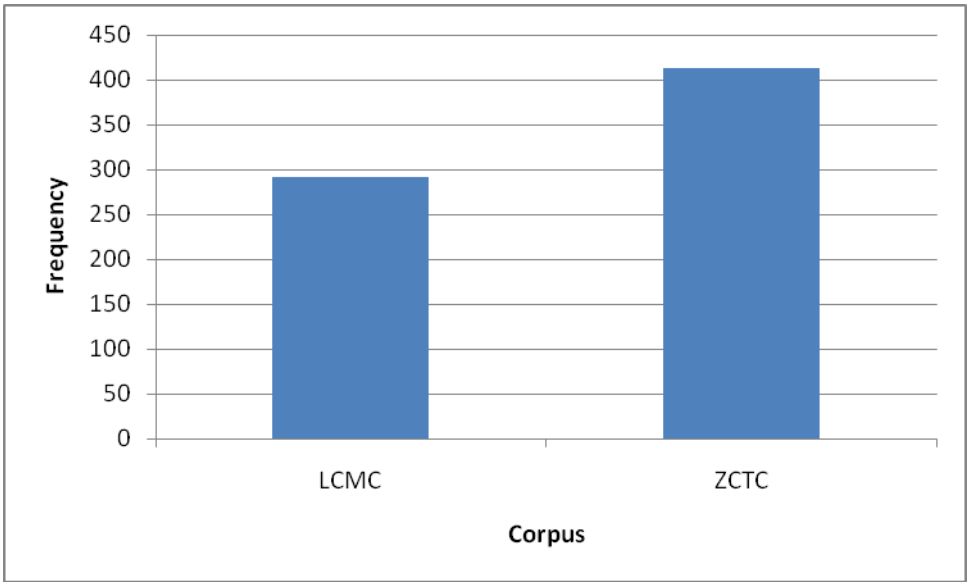

Figure 2. High-frequency word clusters in LCMC and ZCTC

On the other hand, word clusters have a much wider coverage in translated Chinese in comparison with native Chinese (see Figures 3 and 4). As can be seen in the figures, because of the low overall frequencies of 2-word clusters with a minimum coverage rate of 50% (18 and 20 instances in LCMC and ZCTC respectively) and 3-word clusters with a minimum coverage rate of 20% (zero and four instances in LCMC and ZCTC respectively), their frequencies are quite similar in native and translated Chinese. However, there is a marked contrast in the frequencies of 2-word clusters with a minimum coverage rate of 30% (35 and 65 instances in LCMC and ZCTC respectively) and 3-word clusters with a minimum coverage rate of 10% (eight and 23 instances in LCMC and ZCTC respectively) in the two corpora. This contrast displays an accelerating tendency as the coverage rate drops: there are 101 and 170 occurrences of 2-word clusters with a minimum coverage rate of 20%, and 61 and 132 instances of 3-word clusters with a minimum coverage rate of 5%, in the native and translated corpora respectively. The higher frequency and wider coverage of word clusters in translational Chinese suggests that translators demonstrate a higher propensity for striving for fluency than writers of native Chinese texts.

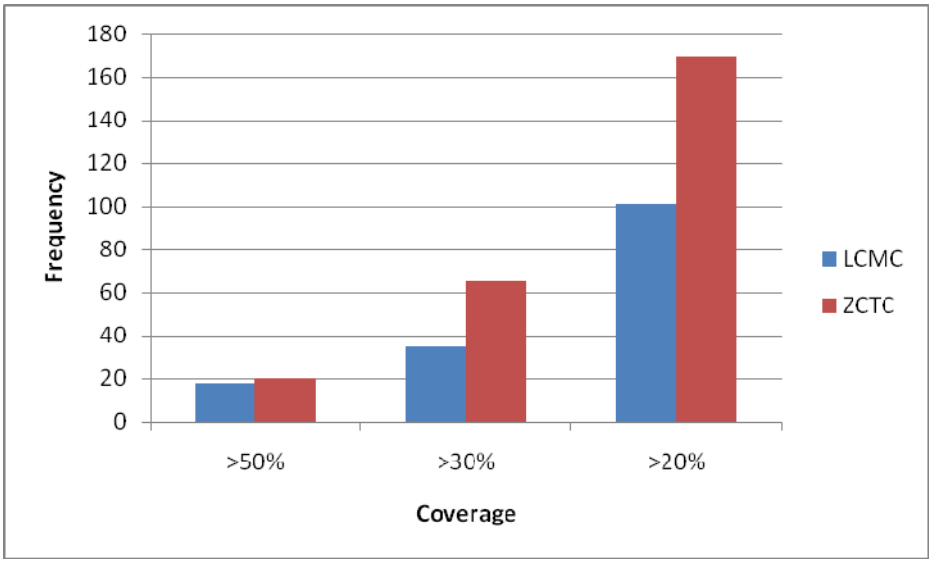

Figure 3. Coverage of 2-word clusters in LCMC and ZCTC

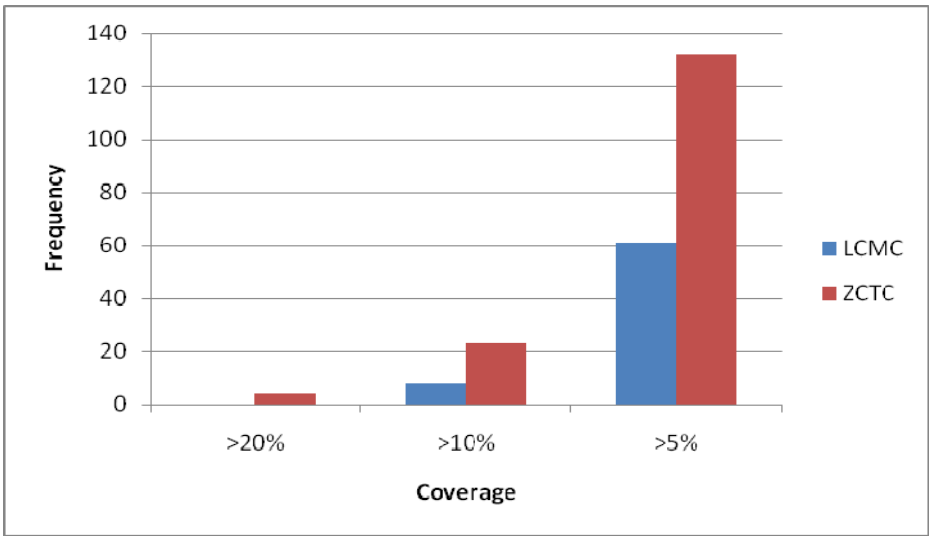

Figure 4. Coverage of 3-word clusters in LCMC and ZCTC

In addition to the difference in coverage of word clusters in general, there is also a sharp contrast in the highest coverage rate in the two corpora. For example, for 2-word clusters, the highest coverage rate in LCMC is 69.8% (的 一 'DE one') whereas the highest coverage rate in ZCTC is 79.8% (不 是 'not be'); similarly for 3-word clusters, the highest

coverage rate in LCMC is 19.6% (是 一 种 'be one kind') whereas the highest rate in ZCTC is 27.6% (而 不 是 'but not be').

Apart from the macro-level quantitative analysis above of frequency and coverage, a qualitative analysis of high-frequency word clusters at micro level yields equally interesting findings. Tables 4 and 5 show the distribution of high-frequency 3-word clusters (defined here as those with a minimum normalized frequency of 50 instances per million words) and 2-word clusters (defined as those with a minimum normalized frequency of 300 instances per million words).[4]

Table 4. High-frequency 3-word clusters in LCMC and ZCTC

Type	Word clusters	Literal glosses
Common word clusters	是 一 种	be one kind
	并 不 是	but not be
	是 不 是	be not be
	的 一 种	DE one kind
	很 大 的	very large DE
	这 是 一	this is one
	更 多 的	more many DE
	是 我 的	be I DE
	最 大 的	most large DE
	的 情况 下	DE situation under
	而 不 是	but not be
	有 一 种	have one kind
	所 说 的	SUO say DE
Unique in ZCTC	最 重要 的	most important DE
	了 他 的	ASP he DE
	最 好 的	most good DE
	在 他 的	in he DE

	蒲 式 耳	transliteration of ‘bushel’
	这 件 事	this CL matter
	这 两 个	this two CL
	的 一 部分	DE one part
	了 一 种	ASP one kind
	和 他 的	and he DE
	重要 的 是	important DE be
	的 两 个	DE two CL
	更 大 的	more large DE
	是 一 位	be one CL
	有 一 天	have one day
	表面 活性 剂	surface active agent
	S A N	S A N
	也 不 会	also not will
	一 段 时间	one length time
	在 我 的	in I DE
	所 做 的	SUO do DE
	了 她 的	ASP she DE
	有 两 个	have two CL
	在 她 的	in she DE
Unique in LCMC	的 基础 上	DE basis on

It is clear that some 2-word and 3-word clusters, such as those listed as “common word clusters” in the tables, are frequently used in both translational and native Chinese texts. On the other hand, however, there are a much greater number of high-frequency word clusters which are unique in ZCTC (24 high-frequency 3-word clusters and 19 high-frequency 2-word clusters) than those which are unique in LCMC (one high-frequency 3-word cluster and three high-frequency 2-word clusters).

Table 5. High-frequency 2-word clusters in LCMC and ZCTC

Type	Word clusters	Literal glosses
Common word clusters	不 是	not be
	他 的	he DE
	了 一	ASP one
	一 种	one kind
	的 人	DE person
	的 一	DE one
	这 是	this be
	不 能	not can
	自己 的	self DE
	就 是	precisely be
	这 一	this one
	中 的	middle DE
	我 的	I DE
	是 一	be one
	上 的	above DE
	两 个	two CL
	人 的	person DE
	都 是	all be
	也 是	also be
	也 不	also not
	一 次	one CL
	有 一	have one
	一 位	one CL
	并 不	but not
	的 时候	DE time
	大 的	large DE
	新 的	new DE
	她 的	she DE
	是 在	be on

Unique in ZCTC	是 一	be one
	你 的	you DE
	不 会	not will
	他们 的	they DE
	了 他	ASP he
	是 一个	be one-CL
	公司 的	company DE
	的 问题	DE issue
	重要 的	important DE
	了 我	ASP I
	的 话	DE word (‘if’)
	是 个	be CL
	他 说	he say
	多 的	many DE
	这样 的	this DE
	它 的	it DE
	的 一个	DE one
	说 我	say I
	我们 的	we DE
Unique in LCMC	的 是	DE be
	到 了	arrive ASP
	的 发展	DE development

Furthermore, it is of interest to note in Tables 4 and 5 that high-frequency 2-word and 3-word clusters unique in ZCTC are mostly demonstrative structures (e.g. 了 他 的 ‘ASP he DE’, 你 的 ‘you DE’, and 这样 的 ‘this DE’) and modifying structures (e.g. 最 重要 的 ‘most important DE’, and 公司 的 ‘company DE’). Indeed many of these demonstrative structures are also modifying structures. In contrast, high-frequency word clusters which are

unique in LCMC appear to be mainly head structures (e.g. 的 基础 上 'DE basis on, on the basis of', and 的 发展 'DE development, the development of').

While the comparison of 2-word and 3-word clusters above has revealed some interesting similarities and differences between native and translated Chinese in terms of their use of recurring formulaic expressions, it does not tell us much about how word clusters can help translators to achieve fluency. Although short word clusters such as 的 一 'DE one' and 不 是 'not be' do not appear to contribute to fluency, longer clusters as exemplified in Table 6 show that such structurally defined recurring formulaic expressions (which may not necessarily be a complete unit of meaning) are certainly as useful in helping the translator to achieve native-like fluency in translation as they are in a native speaker's language production.

Table 6. Some examples of 3-to-6-word clusters

Cluster type	Chinese example	English gloss
3-word clusters	世界 上 最	most… in the world
	主要 是 因为	mainly because
	一点 也 不	not at all
	是 不可避免 的	is unavoidable
4-word clusters	很 大 程度 上	to a large extent
	是 可以 理解 的	is understandable
5-word clusters	从 某种 意义 上 说	in a sense
	这 并 不 是 说	this does not mean
6-word clusters	一 遍 又 一 遍 地	again and again
	最 强劲 的 增长 是 在	the strongest growth is in

The analyses of word clusters at macro- and micro levels suggest that there are a number of quantitative and qualitative differences between native and translational Chinese

in terms of their use of word clusters. Such differences highlight the "relatively higher level of homogeneity of translated texts" in terms of word cluster use (Laviosa 2002: 72), thus providing fresh evidence in support of the TU hypothesis of convergence or levelling out.

6. Reformulation markers

Now we will compare the use of reformulation markers in native and translational Chinese. There are different opinions in the literature about what counts as a reformulation marker. The term used in a narrow sense refers to discourse markers which are strictly paraphrastic, i.e. indicating equivalence (e.g. Murillo 2004). On the other hand, as Cuenca (2003: 1072) observes, "reformulation is more than a strict paraphrase." She argues that "[i]t should be considered a complex semantic category that ranges from strict paraphrase to other values such as specification, explanation, summary or denomination, and even to non-paraphrastic meanings such as implication, conclusion and contrast" (Cuenca 2003: 1073), though only paraphrastic reformulation markers are investigated in her study. While "non-paraphrastic reformulations typically recapitulate or resume something from the preceding discourse", paraphrastic reformulation markers have "the metalinguistic function of clarifying, specifying, expanding or elaborating without changing the semantic content" (Aijmer 2007: 44-45). On other hand, Del Saz and Fraser (2005) argue that paraphrastic versus non-paraphrastic is a distinction difficult to maintain, at least for English. Instead, they classify reformulations in English into four categories: 'expansion' (i.e. providing more information), 'compression' (i.e. summarizing or recapitulating with a single expression), 'modification' (i.e. modifying a prior segment), and 'reassessment' (i.e. revising the speaker's opinion of an implication conveyed by a prior segment). Blackmore (2007) provides a nice discussion of the varieties of reformulations as well as various approaches to classifying reformulation markers.

This study focuses on commonly used paraphrastic reformulation markers for elaboration and explicitation in Chinese as listed in Table 7. Syntactically, some of them can be used either as a connective or as a predicate. As we are only interested in those instances which are used as reformulation connectives, all instances of these items were retrieved from the two Chinese corpora and evaluated manually in KWIC (key-word in context) concordances to avoid errors in automatic annotation of word classes. Table 7 gives the frequencies of the reformulation markers in the two corpora following the human analysis.

Table 7. Frequencies of reformulation markers in LCMC and ZCTC

Style	Reformulation mark	Gloss	LCMC	ZCTC
Formal	即	namely, i.e.	267	274
	换言之	to put it differently	5	8
Informal	也就是说	that is to say	27	28
	或者说	or rather	14	25
	换句话说/换句话讲	put in other words	8	18
	这就是说	that is to say	15	10
	这就意味着	this means…	1	20
	我的意思是	what I mean is…	1	10
	更确切/准确/具体地说	to be more precise/specific	2	7
Total			340	400

It can be seen in the table that the translational corpus ZCTC makes more frequent use of reformulation markers than the matching native Chinese corpus LCMC, and the difference in the overall frequencies in the two corpora is statistically significant (LL = 4.52, a degree of freedom, p=0.033). It is even more interesting to note that the contrast between formal and informal reformulation markers. The two formal reformulation markers 即 'namely, i.e.' and 换言之 'to put it differently' have an archaic flavour, but they are no longer viewed as

archaisms because of their prevalence in modern Chinese discourse, especially in formal writing.[5] Although they are terse in form, these reformulation markers are not more stylistically simpler than the more colloquial forms such as 也就是说 “that is to say” and 换句话说 “put in other words”, which are referred to as informal reformulation markers in this study.

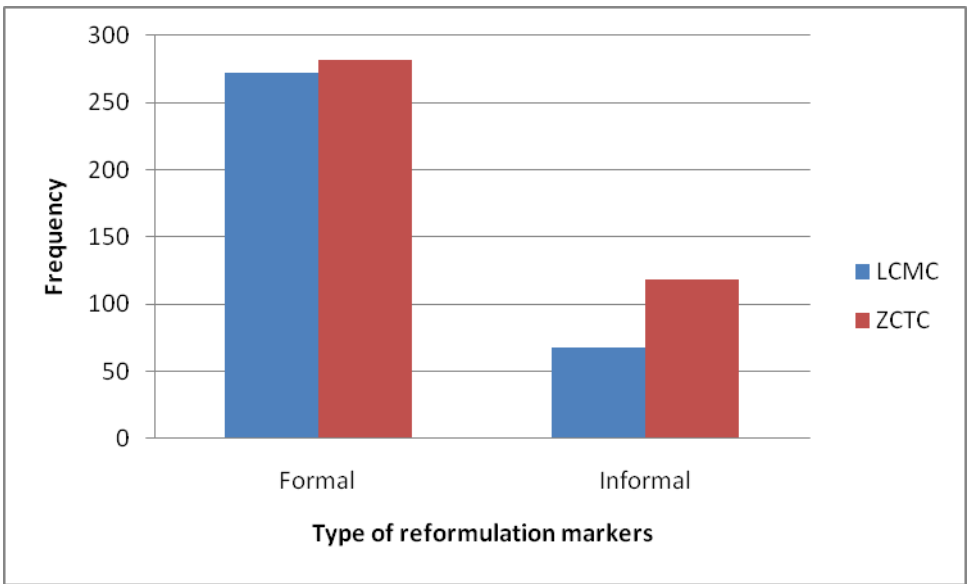

Figure 5. Formal and informal reformulation markers in LCMC and ZCTC

As can be seen in Figure 5, although both formal and informal reformulation markers are more common in the translational corpus ZCTC, the frequencies of the formal forms of reformulation markers in the two corpora are very close and not significant (LL=0.127, 1 degree of freedom, p=0.772), whereas the colloquial informal forms are substantially are more common in translational than native Chinese (LL=13.31, 1 degree of freedom, p<0.001).[6] This finding is in line with the distribution patterns of formal and informal conjunctions in native and translational Chinese as observed in Xiao and Yue (2009).

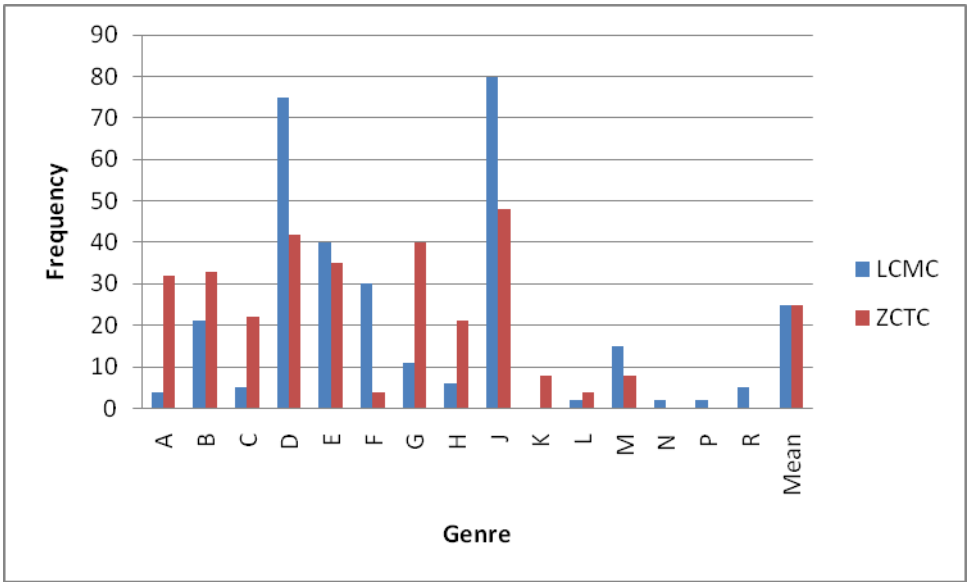

Figure 6. Distribution of formal reformulation markers across genres

The different behaviours of formal and informal reformulation markers are more clearly shown in their distribution across genres in native and translational Chinese. As can be seen in Figure 6, although the overall frequencies of formal reformulation markers in the two corpora do not differ significantly, there are considerable variations across genres. Formal reformulation markers are infrequent in genres of imaginative writing (i.e. various types of fiction K-P and humour R) in both native and translational corpora.[7] In informative writing (i.e. A-J), they are significantly more common in news (A-C), essays and biography (G) and reports/official documents (H) in translated texts but much more frequent in religious writing (D), popular reading (F) and academic prose (J) in native texts.[8] These three genres of the second group are all formal types of writing which demonstrate a high propensity for a formal style.

In contrast, as can be seen in Figure 7, informal reformulation markers are more frequently used in most genres in translational Chinese. While they are still more frequent in religious writing (D) and popular reading (F) in native Chinese, the contrast between native and translated texts is less marked than that in formal reformulation markers. These

observations suggest that, on the one hand, the use of reformulation markers varies across genres while, on the other hand, translational Chinese has a tendency to use stylistically simpler markers in comparison with native Chinese, thus providing fresh evidence for simplification in translations but a counter example of the normalization hypothesis.

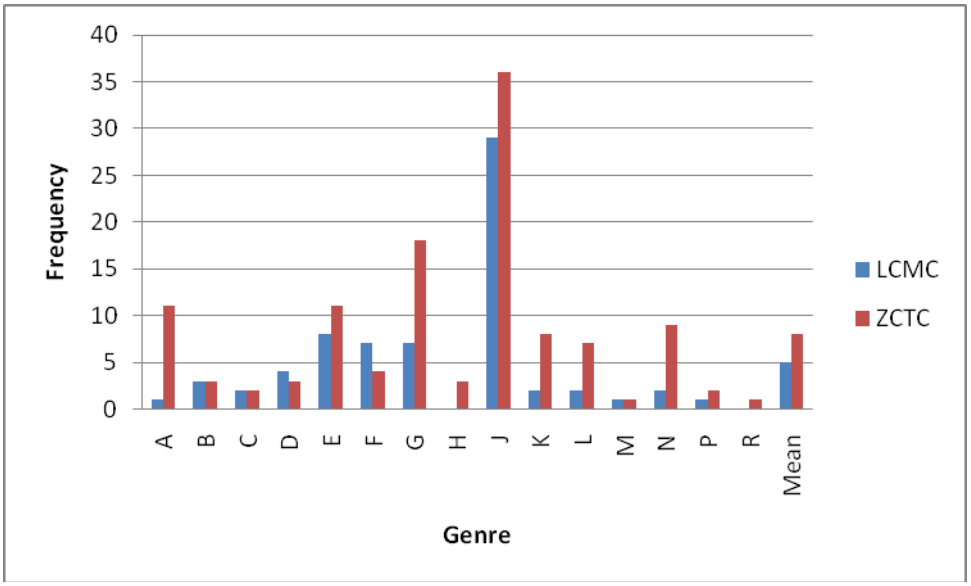

Figure 7. Distribution of informal reformulation markers across genres

We noted earlier that paraphrastic reformulation markers are an explicative device in translation (cf. section 2). Then to what extent are reformulation markers explicated in the translation process – that is, they are used in Chinese translations but not in English source texts? To answer this question, we searched a balanced English-to-Chinese parallel corpus which is composed of a roughly equal amount of literary and non-literary English texts translated into Chinese, totalling 814,269 English words and 677,126 Chinese words.[9]

Table 8 gives the frequencies of reformulation markers in Chinese translations which are transferred from the English source texts and those which are supplied in the translation process as well as their explicitation rates. As can be seen, typically 10%-30% of reformulation markers are explicated (i.e. added by translators), with explicitation rates

varying markedly from zero to over 85% in extreme cases. It is also clear that the more semantic content a reformulation marker contains, the less likely it is explicated. For example, the highly semantically loaded reformulation markers like 更确切/具体地说 'to be more precise/specific', 我的意思是 'what I mean is…', and 这就意味着 'this means…' are all transferred from the English source texts, whereas the purely paraphrastic and explicative reformulation marker 即 'namely, i.e.' has a very high rate of explicitation, as demonstrated in examples below.[10]

Table 8. Explicitation of reformulation markers in English-to-Chinese translation

Reformulation marker	Gloss	Transferred from source text	Explicated in translation	Explicitation rate
即	namely, i.e.	14	82	85.4%
也就是说/这就是说	that is to say	11	5	31.3%
换言之	to put it differently	3	1	25%
或者说	or rather	15	2	11.8%
换句话说	put in other words	18	2	10%
更确切/具体地说	to be more precise/specific	7	0	0
我的意思是	what I mean is…	7	0	0
这就意味着	this means…	1	0	0

1) After Lewes' death in 1878, Eliot wrote nothing further, dying just two years later, in 1880.

1878 年 路易斯 死 后， 艾略特 再 没 写 什么 东西 。两 年 后，

1878 year Lewes die after Eliot again not write what stuff two year after

即 1880 年， 艾略特 也 去世 了 。

i.e. 1880 year Eliot also die ASP

2) The American frontier fostered the notion that “everybody is an entrepreneur,” or that everybody has the right to try his hand.

美国 的 西部 边疆 产生 出 一 种 观点， **即** 认为：每个人

US DE west frontier foster RVC one kind viewpoint **i.e.** think everyone

都 是 创业家， 或者说 至少 可以 说 每个人 都 有权 试 一 试。

all be entrepreneur or at least may say everyone all have right try one try

In examples like these, the reformulation marker 即 ‘namely, i.e.’ in Chinese translations (highlighted in the examples) cannot find its equivalence in the English originals; in other words, they are supplied by translators as an explicitation strategy. Explicitation of reformulation markers in English-to-Chinese translation typically occurs where appositions are used in the English source texts as in (1), where “just two years later” actually refers to the prepositional phrase ‘in 1880’. Appositions can not only be phrases but also clauses, as exemplified in (2), where the noun phrase ‘the notion’ refers to what is expressed by the two *that*-clauses.

In addition to appositions in the English source texts, other structures are translated into Chinese using explicated reformulation markers as well. For example, in (3) the *which* relative clause is translated using a reformulation; in (4) the *by* prepositional phrase is reformulated; in (5) the Chinese translation reformulates what is expressed by the infinitival complement in the source text (i.e. *to take campaigns out of unregulated hurly-burly of politics*) while Chinese translation uses a totally different sentence structure in (6). In addition to their explicative function in translation, reformulation markers, especially the semantically less full marker 即 ‘namely, i.e.’, also function to break long sentences in the English source texts into shorter sentence segments, which are characteristic of Chinese.[11]

3) This is the full text of the 1989 second edition, which consolidates the original OED, all the supplementary volumes and additional new material.

它 包括 了 1989 年 第二 版 的 全部 内容，**即** 原先 的

it include ASP 1989 year second edition DE all content **i.e.** former DE

牛津 辞典、 补遗 部分 及 新 增 材料。

Oxford dictionary supplementary part and new add material

4) Dr Butler made history by being the first woman head of a formerly all-male college <...>

布特赖 博士 做 了 永垂史册 的 事情，**即** 她 是 第一 位 原来 人员

Butler Dr do ASP make history DE thing **i.e.** she is first CL former personnel

全 为 男人 的 学院 中 担任 校长 的 女性。

all be man DE college in hold post head DE female

5) Drastic campaign reform is motivated by the desire to take campaigns out of unregulated hurly-burly of politics <...>

激进的 竞选 改革 是 受 这样 一 种 愿望 启动 的，

radical campaign reform be PSV this one kind desire start DE

即 竞选 运动 要 摆脱 不 受 约束 的 政治 喧闹 <...>

i.e. campaign movement must get rid of not PSV restrict DE political noise

6) The two chief types of these programs are Individual Retirement Accounts (IRAs) and Keogh plans.

这 种 计划 主要 有 两 类，**即** 个人 退休金 账户 和

this kind plan chiefly have two type **i.e.** individual pension account and

基奥 计划。

Keogh plan

7. Conclusions

In this article, we have explored idioms, word clusters and reformulation markers in translated Chinese in comparison with native Chinese on the basis of two comparable balanced corpora of Mandarin Chinese and an English-to-Chinese parallel corpus, in an attempt to verify whether some English-based, genre-specific features of translations can be generalized as translation universals in the light of evidence from Chinese, a language which is "genetically" distinct from English.

Our results show that idioms are significantly more common in native Chinese as a whole and also in nearly all genres, a finding which runs counter to Baker's (2007: 14) first expectation that "translators ought to be making heavy use of idioms" but supports her second expectation that idioms characterized with a high degree of opacity are more likely to be avoided in translations than those less opaque ones. This finding is closely associated with the formation and etymological source of Chinese idioms, especially the so-called *chengyu*, which are quite different from their English counterparts (see section 2). As the distribution patterns of idioms tend to be language-specific, the heavy use of idioms may not be a universal feature of translational language. The statistically significant quantitative contrast in the use of idioms in native and translational Chinese also tells a different story from the translation universal hypothesis of normalization.

On the other hand, word clusters are substantially more common in translational Chinese in terms of frequency, coverage as well as key clusters, an observation which is in

line with findings reported in previous translation studies such as Baker (2004) and Nevalainen (2005). Such cross-linguistic evidence shows that translators tend to use fixed and semi-fixed recurring patterns which are purely structurally defined on the basis of their collocational behaviour in an attempt to achieve improved fluency. Our corpora also reveal some qualitative difference between native and translational Chinese. While word clusters in translated texts tend to be modifying structures, those in native texts are more likely to be head structures. It will nevertheless require substantial further research and cross-linguistic evidence to claim this feature as a universal feature of translation. On the other hand, the quantitative and qualitative differences uncovered in this study between native and translated Chinese show that translated texts are more similar to each other than to non-translated texts, which means that the universal hypothesis of convergence or levelling out is upheld in the light of Chinese evidence.

Our finding based on comparable balanced corpora of native and translational Chinese supports previous observations in some specific genres that reformulation markers can function as an explicitation strategy. More interestingly, it is found that formal and informal reformulation markers are sensitive to genre variation and may behave differently in native and translated texts. While translational Chinese generally makes more frequent use of informal colloquial reformulation markers, the distribution of formal markers seems to interact with the formality of genres, suggesting that translations are stylistically simpler than native Chinese texts. This means that while explicitation is supported in translational Chinese, the patterns of reformulation marker use in translated Chinese also provide evidence in support of simplification and convergence but against the normalization hypothesis.

As a final remark, we believe that translation universal research should follow the approach to universals in language studies in general. This means that it must not be based on one language and confined to its closely related languages alone, which largely characterizes

the current situation of TU research. If the features observed of translational English are to be generalized as translation or mediation universals, evidence from distinctly different languages such as Chinese is clearly useful. It is our hope that the study of translational Chinese exemplified in this article will bring fresh insights to translation universal research.

Acknowledgements

I am obliged to the China National Foundation for Social Sciences for their support of our project "A corpus-based quantitative study of translational Chinese in English-Chinese translation" (Award reference 07BYY011).

Notes

1. See the official website of the corpus (http://dbo.sinica.edu.tw/SinicaCorpus/) for more details about the Sinica corpus. Note that some of the reformulation markers in Chen's (2006) study are not strictly paraphrastic denoting equivalence, e.g. 总而言之 'in summary', 总的来说 'in summary', 归根到底 'to be more precise', and 不用说 'needless to say'. See section 4 for further discussion of the term in narrow and broad senses.

2. Since ICTCLAS (version 3.0), the lexical analysis system used to annotate our corpora differentiates between idioms belonging to different parts of speech but not between idioms of different semantic types discussed in this study, it is impossible to know the proportions of idioms in narrow and broad senses.

3. Like a wordlist word cluster, a keyword cluster is composed of two or more words which co-occur with each other repeatedly. A keyword cluster differs in that it only uses keywords (cf. Scott 2009: 145). In this study, the native and translational Chinese corpora LCMC and ZCTC are used against each other as the reference corpus in keyword extraction.

4. The 3-word cluster 蒲 式 耳 'bushel' in Table 4 is a technical term unknown to the ICTCLAS tagger, which incorrectly tokenized it as three separate words. In the table, the character string 'S A N' is an acronym for 'Storage Area Network'.

5. The formal marker 即 'namely, i.e.' occurs 267 times and 41 times per million tokens respectively in written and spoken Chinese as represented by LCMC and the Lancaster Los Angeles Spoken Chinese Corpus (LLSCC, see Xiao and Tao 2007), with all instances in the spoken corpus being found in edited oral narrative, the most 'literate' genre covered in LLSCC. Similarly, the formal reformulation marker 换言之 'to put it differently' is also more frequent in LCMC than LLSCC (with a normalized frequency of five and one respectively in the two corpora).

6. The stylistic differences between formal and informational reformulation markers in Chinese, unfortunately, cannot be retained in their English glosses.

7. Because of the low overall frequencies, the differences between the native and translational corpora are not statistically significant in imaginative genres.

8. The differences between native and translated Chinese in text categories B (i.e. news editorials, LL=2.62, 1 degree of freedom, p=0.106) and E (i.e. instructional writing, LL=0.36, 1 degree of freedom, p=0.547) are not significant.

9. This is part of the *General Chinese-English Parallel Corpus* created by Beijing Foreign Studies University (BFSU). See Wang (2004) for more information.

10. These examples are cited from our English-to-Chinese parallel corpus. In the literal glosses of Chinese examples, ASP stands for 'aspect marker', CL for 'classifier', DE for

structural particle *de*, PSV for 'passive', RVC for 'resultative verb complement', and SUO for the particle *suo*, which used before a verb to form a nominal construction.

11. Syntactically, Chinese grammar is much more tolerant of the so-called 'run-on sentences' than English.

References

Aijmer, Karin (2007) "The meaning and functions of the Swedish discourse marker *alltså* – Evidence from translation corpora". *Catalan Journal of Linguistics* 6: 31-59.

An, Na, Liu, Haitao and Hou, Min (2004) "Yuliaoku zhong shuyu de biaoji wenti (Tagging of the idiom in the corpus)". *Zhongwen Xinxi Xuebao (Journal of Chinese Information Processing)* 2004(1): 20-25.

Aston, Guy and Burnard, Lou (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Baker, Mona (1992) *In Other Words*. London: Routledge.

Baker, Mona (1996) "Corpus-based translation studies: The challenges that lie ahead", in Harold Somers (ed.) *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, 175-186. Amsterdam: John Benjamins.

Baker, Mona (2004) "A corpus-based view of similarity and difference in translation". *International Journal of Corpus Linguistics* 9(2): 167-193.

Baker, Mona (2007) "Patterns of idiomaticity in translated vs. non-translated text". *Belgian Journal of Linguistics* 21: 11-21.

Barlow, Michael (1995) *A Guide to ParaConc*. Houston: Athelstan.

Baroni, Marco and Bernardini, Silvia (2003) "A preliminary analysis of collocational differences in monolingual comparable corpora", in Dawn Archer, Paul Rayson, Andrew

Wilson and Tony McEnery (eds) *Proceedings of the Corpus Linguistics 2003*, 82-91. Lancaster: UCREL, Lancaster University.

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

Biber, Douglas (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, Douglas, Conrad, Susan and Cortes, Viviana (2004) "If you look at…: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25(3): 371-405.

Blackmore, Diana (2007) "'Or'-parentheticals, 'that is'-parentheticals and the pragmatics of reformulation". *Journal of Linguistics* 43(2): 311–339.

Burnard, Lou and Todd, Tony (2003) "Xara: An XML aware tool for corpus searching", in Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds) *Proceedings of Corpus Linguistics 2003*, 142-144. Lancaster: UCREL, Lancaster University.

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

Cortes, Viviana (2002) "Lexical bundles in freshman composition", in Randi Reppen, Susan Fitzmaurice and Douglas Biber (eds) *Using Corpora to Explore Linguistic Variation*, 131-145. Amsterdam: John Benjamins.

Cuenca, Maria-Josep (2003) "Two ways to reformulate: A contrastive analysis of reformulation markers". *Journal of Pragmatics* 35(7): 1069-1093.

De Cock, Sylvie (1998) "A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English". *International Journal of Corpus Linguistics* 3(1): 59-80.

De Cock, Sylvie (2000) “Repetitive phrasal chunkiness and advanced EFL speech and writing”, in Christian Mair and Marianne Hundt (eds.) *Corpus Linguistics and Linguistic Theory. Papers from ICAME 20 1999*, 51-68. Amsterdam: Rodopi.

Del Saz, Mª Milagros and Fraser, Bruce (2005) “Reformulation in English”. Paper presented at the 9th International Pragmatics Conference, Riva Del Garda, Italy, 10-15 July 2005.

Fernando, Chitra (1996) *Idioms and Idiomaticity*. Oxford: Oxford University Press.

Gaspari, Federico and Bernardini, Silvia (2010) “Comparing non-native and translated language: Monolingual comparable corpora with a twist”, in Richard Xiao (ed.) *Using Corpora in Contrastive and Translation Studies*, 215-234. Newcastle: Cambridge Scholars Publishing.

Granger, Sylviane (1996) “From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora”, in Karin Aijmer, Bengt Altenberg and Mats Johansson (eds) *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, March 1994*, 37-51. Lund: Lund University Press.

Granger, Sylviane (1998) “Prefabricated writing patterns in advanced EFL writing: Collocations and formulae”, in Anthony Cowie (ed.) *Phraseology: Theory, Analysis and Applications*, 145-160. Oxford: Clarendon Press.

Halliday, M. A. K. (2000) *An Introduction to Functional Grammar* (2nd ed.). London: Arnold.

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

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

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

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

Liu, Qun, Zhang, Huaping and Zhang, Hao (2008) *The ICTCLAS Tagset* (version 3.0). http://www.ictclas.org/. Accessed on 5th March 2010.

Mauranen, Anna (2007) “Universal tendencies in translation”, in Margaret Rogers and Gunilla Anderman (eds) *Incorporating Corpora. The Linguist and the Translator*, 32-48. Clevedon: Multilingual Matters.

McCarthy, Michael and Carter, Ronald (2004) “This that and the other: Multi-word clusters in spoken English as visible patterns of interaction”, in Michael McCarthy (ed.) *Explorations in Corpus Linguistics*, 7-26. Cambridge: Cambridge University Press.

McEnery, Tony and Xiao, Richard (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.

Murillo, Silvia (2004) “A relevance reassessment of reformulation markers”. *Journal of Pragmatics* 36(11): 2059–2068.

Mutesayire, Martha (2005) *Cohesive Devices and Explicitation in Translated English – A Corpus-based Study*. PhD thesis, University of Manchester.

Nevalainen, Sampo (2005) “Köyhtyykö kieli käännettäessä? Mitä taajuuslistat kertovat suomennosten sanastosta”, in Anna Mauranen and Jarmo Harri Jantunen (eds) *Käännössuomeksi*, 141-162. Tampere: Tampere University Press.

Oxford University Press (1989) *Oxford English Dictionary* (2nd ed.). Oxford: Oxford University Press.

Schourup, Lawrence (1999) “Discourse markers”. *Lingua* 107: 227-265.

Scott, Mike (2009) *The WordSmith Tools* (version 5.0). Oxford: Oxford University Press.

Sinclair, John (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sperber, Dan, and Deirdre, Wilson (1995) *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

Wang, Kefei (2004) “Han Ying duiying yuliaoku de sheji (Design of the Chinese-English parallel corpus)”, in Kefei Wang (ed.) *Shuangyu Duiying Yuliaoku Yanzhi yu Yingyong (The Development and Application of Bilingual Parallel Corpora)*, 36-53. Beijing: Foreign Language Education and Research Press.

Wu, Chu-hsia (1995) “On the cultural traits of Chinese idioms”. *Intercultural Communication Studies* V(1): 61-81.

Xiao, Richard (2006) “Review of Xaira: An XML Aware Indexing and Retrieval Architecture”. *Corpora* 1(1): 99-103.

Xiao, Richard (2009) “Multidimensional analysis and the study of world Englishes”. *World Englishes* 28 (4): 421-450

Xiao, Richard (2010) “How different is translated Chinese from native Chinese?”. *International Journal of Corpus Linguistics* 15(1): 3-33.

Xiao, Richard and Tao, Hongyin (2007) *The Lancaster Los Angeles Spoken Chinese Corpus*. Lancaster: UCREL.

Xiao, Richard and Yue, Ming (2009) “Using corpora in Translation Studies: The state of the art”, in Paul Baker (ed.) *Contemporary Corpus Linguistics*, 237-262. London: Continuum.

Xiao, Richard, He, Lianzhen and Yue, Ming (2010) “In pursuit of the third code: Using the ZJU Corpus of Translational Chinese in Translation Studies”, in Richard Xiao (ed.) *Using Corpora in Contrastive and Translation Studies*, 182-214. Newcastle: Cambridge Scholars Publishing.

Yang, Changrong (2004) “Ying Han chengyu de jieding yu duiying (Defining criteria for English and Chinese idioms and correspondences between them)”. *US-China Foreign Language* 2(7): 43-47.