

# Word clusters and reformulation markers in Chinese and English

## Implications for translation universal hypotheses\*

Richard Xiao

Edge Hill University (United Kingdom)

This article presents a corpus-based contrastive study of word clusters and reformulation markers in Chinese and English, and discusses the implications of the findings for translation universal hypotheses. The study is based on three balanced comparable corpora which represent British English, native Chinese and translational Chinese in addition to an English-Chinese parallel corpus which provides a basis for comparing native and translated English and investigating explicitation in translation. Our results show that word clusters are substantially more common in translated Chinese, suggesting a tendency in translations to use fixed and semi-fixed recurring patterns in an attempt to achieve improved fluency. The more frequent use of word clusters, especially those of high frequency and high coverage in translational Chinese, is also likely to be a result of the influence of the English source language because word clusters are significantly more prevalent in native English in relation to native Chinese. Chinese and English tend to use reformulation markers of different styles while on the other hand, reformulation markers are generally more common in both translated English and translated Chinese than in their native counterparts, suggesting that reformulation markers function as a strategy for explicitation in translations, which tend to use oral, stylistically simpler forms than non-translated texts.

**Keywords:** translation universal, word cluster, reformulation marker, Chinese/English

### 1. Introduction

Since its revival in the 1990s, for which the rapid development of multilingual corpus resources such as comparable and parallel corpora has been recognized as a principal reason (cf. Salkie, 2002), contrastive analysis has substantially expanded

its scope of research from contrastive linguistics in a narrow sense to the term in a broad sense. While the former is traditionally concerned with cross-linguistic contrast of different languages, the latter has developed over time to cover not only inter-lingual but also intra-lingual contrast by comparing different varieties of the same language. For example, in second language acquisition research, what Granger (1996) calls the CIA (Contrastive Interlanguage Analysis) approach compares learners' written or spoken production in their second or foreign language with comparable native data of the language they are learning, which helps to determine the extent of under- or overuse of particular linguistic features in their interlanguage (e.g. Xiao, 2007), or compares the interlanguages of learners from different mother tongue backgrounds, which helps to identify the common features of the second language acquisition process. Similarly, the intra-lingual contrastive analysis has been used in Translation Studies to compare translational and native varieties of the same language on the basis of monolingual comparable corpora to uncover the common features of translations, or the so-called translation universals (e.g. Mauranen and Kujamäki, 2004; Xiao, 2010).

Translation universal (TU) research has recently been an important area of corpus-based Translation Studies. A number of linguistic features of translations 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 (see Xiao and Yue (2009) for a comprehensive review). 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 connotative 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).

The present study seeks to demonstrate the value of contrastive analysis in translation universal research by combining intra-lingual and inter-lingual contrastive analysis via two case studies, namely, of word clusters, i.e. fixed or semi-

fixed lexical phrases closely associated with fluency, which is a “preferred strategy” that translators tend to adopt according to the translation universal hypothesis of normalization (Baker, 2004: 182), and of reformulation markers such as ‘that is to say’, which contribute substantially to making messages more explicit. On the one hand, the distributions of word clusters and reformulation markers in native and translational Chinese will be compared and their implications for translation universal hypotheses discussed, while on the other hand, such features in Chinese and English will be contrasted and the possible effects of the source language on translational Chinese explored.

The major corpus resources used in this study include three comparable corpora of English and Chinese. The Freiburg-LOB (FLOB) corpus is a balanced corpus representing written British English as used in the early 1990s (Hundt *et al.*, 1998). The Lancaster Corpus of Mandarin Chinese (LCMC) is designed as a Chinese match for FLOB for use in cross-linguistic contrast of English and Chinese (McEnery and Xiao, 2004), while the ZJU Corpus of Translational Chinese (ZCTC) is created as a translational counterpart of LCMC with the explicit aim of studying features of translated Chinese (Xiao, He and Yue, 2010).

The three corpora are each composed of one million words in five hundred 2,000-word text samples which are taken proportionally from 15 text categories

**Table 1.** The genres covered in FLOB, LCMC and ZCTC.

| Text category | Code | Genre                         | Samples | Percent |
|---------------|------|-------------------------------|---------|---------|
| News          | A    | Press reportage               | 44      | 8.8     |
|               | B    | Press editorials              | 27      | 5.4     |
|               | C    | Press reviews                 | 17      | 3.4     |
| Nonfiction    | D    | Religious writing             | 17      | 3.4     |
|               | E    | Instructional writing         | 38      | 7.6     |
|               | F    | Popular lore                  | 44      | 8.8     |
|               | G    | Biographies and essays        | 77      | 15.4    |
|               | H    | Reports/official documents    | 30      | 6       |
| Academic      | J    | Academic prose                | 80      | 16      |
| Fiction       | 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    | Humour                        | 9       | 1.8     |
| <i>Total</i>  |      |                               | 500     | 100     |

published in the UK and China respectively in the early 1990s. As can be seen in Table 1, these corpora are 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. These comparable corpora of English and the two varieties of Chinese will provide a reliable basis for the quantitative and qualitative analyses of word clusters and reformulation markers to be presented in Sections 3 and 4. A fourth corpus, an English-to-Chinese parallel corpus, will also be introduced and used in Section 4 to contrast reformulation markers in the two languages and to investigate the extent to which reformulation markers are explicated in the translation process.

In the remainder of this article, Section 2 will review previous translation studies of the two linguistic features under investigation, on the basis of which quantitative and qualitative analyses will be undertaken to compare the use of word clusters (Section 3) and reformulation markers (Section 4) in our Chinese and English corpora; the implications of our findings for TU hypotheses will also be discussed. Section 5 concludes the article by summarising the major findings.

## **2. Word clusters and reformulation markers in Translation Studies**

Word clusters are fixed and semi-fixed formulaic expressions based on collocations, which are also known as ‘lexical bundles’, ‘multiword units’, ‘prefabs’, and ‘n-grams’ and so on. Scott (2009: 286) observes that “all words have a tendency to cluster together with some others.” Word clusters are purely structurally defined on the basis of co-occurrences with no regard to their semantic contents. They can be computed automatically using corpus exploration tools such as WordSmith (Scott, 2009).

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 *et al.* 2004; Biber, 2006). In Translation Studies, Baker (2004) and Nevalainen (2005, cited in Mauranen, 2007) both 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 lexicalized collocations in the language.

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 tend to make heavy use of “pre-packed, recurring stretches of language”, or what she refers to as “idioms”. On the other hand, Baker (2007) also expects such formulaic expressions to be avoided in translations because of their often informal flavour arising from their highly opaque meaning (e.g. ‘chew the fat’). These observations point in two opposite directions.

A particular type of word clusters in Baker (2004, 2007) is the so-called reformulation markers such as ‘in other words’ and ‘that is to say’, though a reformulation marker can also be a single word instead of a word cluster (e.g. ‘namely’). 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), although according to Del Saz and Fraser (2005) the distinction between paraphrastic versus non-paraphrastic is 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). Blakemore (2007) also provides a nice discussion of the varieties of reformulations as well as various approaches to classifying reformulation markers.

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 (2004:2066) 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 explicating implicated premises and conclusions.”

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 Translational English Corpus than in the fiction subcorpus of the British National Corpus. 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.<sup>1</sup> He finds that reformulation markers are more common in translated Chinese, which supports the explicitation hypothesis.

These translation studies of 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 regional varieties of English. This means that what has been observed of word clusters and reformulation markers in the studies reviewed 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 ‘translation 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. Furthermore, cross-linguistic contrast of the use of word clusters and reformulation makers in two distinctly different languages such as Chinese and English is in itself of interest in contrastive linguistics.

In the present study, we will use two balanced comparable 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. A matching corpus of English and an English-Chinese parallel corpus will also be used to contrast Chinese and English in their use of word clusters and reformulation markers.

### 3. Word clusters

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 the present study, we use the default settings of the WordSmith Tools (version 5.0), that is, a minimum frequency of 5. Another useful parameter in computing word clusters is coverage, which measures how widespread a word cluster occurs in a given corpus. It is expressed as a percentage of the number of text samples containing a particular word cluster in the total number of text samples in that corpus.

As noted in Section 2, previous research has suggested that word clusters tend to be more commonly used in translated English. Is this also true in Chinese? What similarities and differences do English and Chinese show in their use of word clusters? In this section, we will study word clusters composed of 2-to-6 words because clusters comprising more than six words are rare in million-word corpora like FLOB, LCMC and ZCTC. Table 2 gives the frequencies of word clusters in the two Chinese corpora together with the results of log-likelihood (LL) 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

**Table 2.** 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        |

a keyword cluster analysis,<sup>2</sup> 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.

The more frequent use of word clusters in translational Chinese is probably a result of the translation process in which translators strive to achieve fluency; translators are also likely to be under the influence of the English source language. As can be seen in Figure 1, which compares the use of 2-4-word clusters (clusters of more than four words are infrequent and thus excluded in the graph) in the three comparable corpora, all of the three types of word clusters are most frequent in FLOB and least frequent in LCMC, with ZCTC between the two.

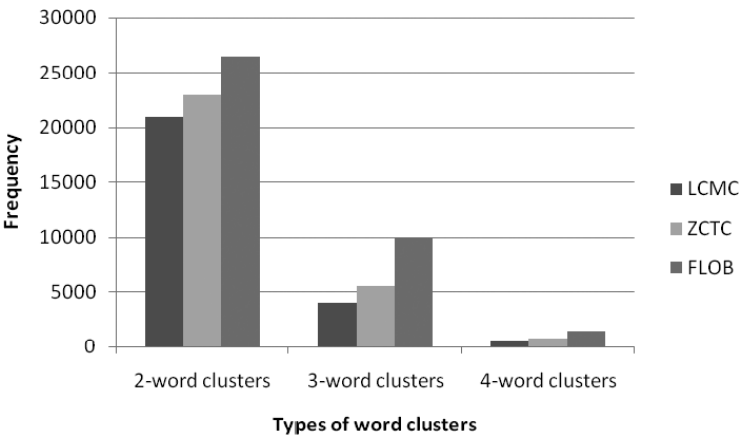


Figure 1. Word clusters in Chinese and English.

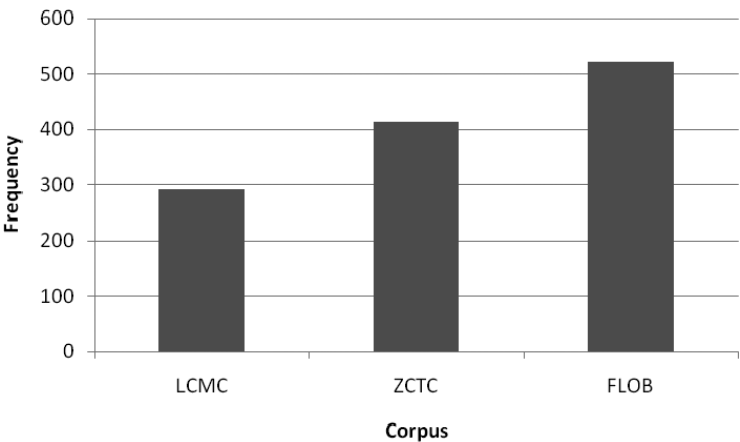


Figure 2. High-frequency word clusters in Chinese and English.

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 d.f.,  $p<0.001$ ). Given that translated Chinese tends to use high-frequency words (Xiao, 2010), it is hardly surprising to find a more frequent use of high-frequency word clusters in ZCTC. While word clusters in different languages cannot be directly compared against each other qualitatively, it is also of interest to note in Figure 2 that the use

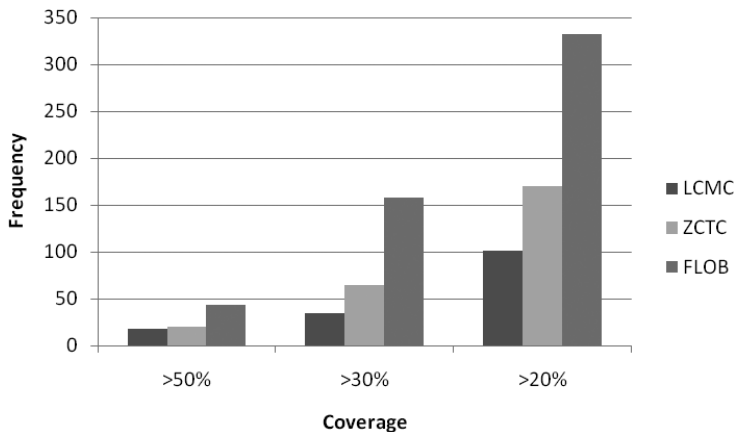


Figure 3. Coverage of 2-word clusters in Chinese and English.

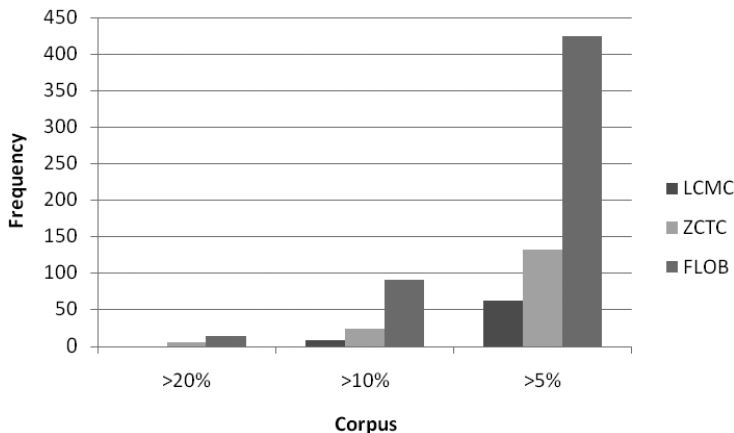


Figure 4. Coverage of 3-word clusters in Chinese and English.

of high-frequency word clusters in the ZCTC corpus of translational Chinese is quantitatively more akin to the English corpus FLOB, which yields 522 instances of high-frequency clusters (498 2-word clusters and 24 3-word clusters).

On the other hand, word clusters have a much wider coverage in translated Chinese in comparison with native Chinese, which is possibly a result of the influence of English (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, 20 and 44 instances in LCMC, ZCTC and FLOB respectively) and 3-word clusters with a minimum coverage rate of 20% (zero, four and 13 instances respectively), their frequencies are quite similar in the three corpora. However, there is a marked contrast in the frequencies of 2-word clusters with a minimum coverage rate of 30% (35, 65 and 158 instances respectively) and 3-word clusters with a minimum coverage rate of 10% (eight, 23 and 90 instances respectively) in the three corpora. This contrast displays an accelerating tendency as the coverage rate drops: there are 101, 170 and 332 occurrences of 2-word clusters with a minimum coverage rate of 20%, and 61, 132 and 425 instances of 3-word clusters with a minimum coverage rate of 5%, in the native and translated Chinese corpora and the comparable English corpus respectively.

The higher frequency and wider coverage of word clusters in translational Chinese suggest that translators demonstrate a higher propensity for striving for fluency than writers of native Chinese texts. Translators are also likely to be under the influence of English, the principal source language of the ZCTC corpus.

The use of word clusters is linked to style and is therefore likely to vary across genres.<sup>3</sup> Figure 5 illustrates the distribution of high-frequency (with a minimum proportion of 0.01% in a given corpus) word clusters of high coverage (with a minimum coverage of 20% for FLOB and 10% for LCMC and ZCTC) across four broad text categories.<sup>4</sup> As can be seen, in all of the four broad text categories, high-frequency high-coverage word clusters are consistently more common in the translational Chinese corpus ZCTC than in the native Chinese corpus LCMC ( $LL = 8.02$ , 3 d.f.,  $p = 0.045$ ), which suggests that contrary to the normalization hypothesis, translational Chinese is essentially distinct from native Chinese because translations are likely to be influenced by the source language, as can be seen from a comparison of ZCTC and FLOB in Figure 5. With the exception of academic writing, which is the least variable of the four text categories in our corpora (cf. Xiao, 2009), high-frequency high-coverage word clusters are even more common in the English corpus FLOB in all other three categories.

Figure 5 also shows that in both English and Chinese high-frequency high-coverage word clusters are most commonly used in fiction and are least frequent in nonfiction. On the other hand, the two languages differ in that the news category makes marginally more frequent use of such word clusters than academic writing

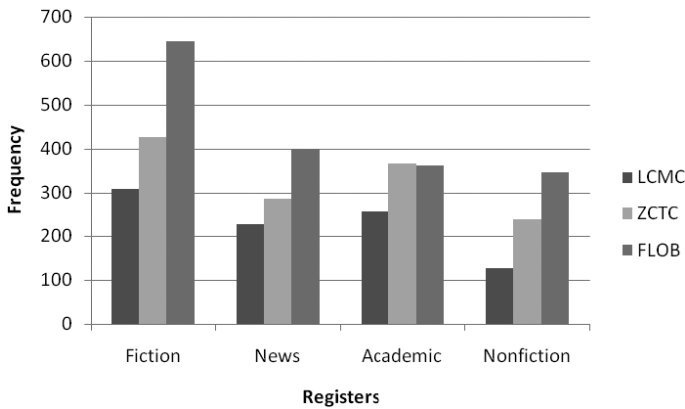


**Figure 5.** Distribution of high-frequency high-coverage word clusters in Chinese and English.

in English whereas in both native and translational Chinese,<sup>5</sup> academic writing makes more frequent use of such word clusters than news.

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 Chinese corpora. For example, for 2-word clusters, the highest coverage rate in LCMC is 69.8% (的 ‘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’). The highest coverage of some clusters of function words in English can be as high as 100% (e.g. ‘of the’ and ‘in the’ in FLOB).

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 3 and 4 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).<sup>6</sup>

It is clear that some 2-word and 3-word clusters, such as those listed as ‘common word clusters’ in Tables 3 and 4, 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).

Furthermore, it is of interest to note in Tables 3 and 4 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

Table 3. 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                 |

Table 4. 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        |

Table 4. (continued)

| Type           | Word clusters | Literal glosses |
|----------------|---------------|-----------------|
| Unique in LCMC | 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           |
|                | 的是            | DE be           |
|                | 到了            | arrive ASP      |
|                | 的发展           | DE development  |

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 5 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.

The analyses of word clusters in English and Chinese at the macro- and micro levels suggest that word clusters are more prevalent in English than in Chinese. They are also substantially more common in translated than native Chinese, suggesting a tendency in translations to use fixed and semi-fixed recurring patterns in an attempt to achieve improved fluency. The more common use of word clusters, especially those of high frequency and high coverage in translational Chinese, is also likely to be a result of the influence of the English source language. The quantitative and qualitative differences between native and translational Chinese

Table 5. 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 |

in terms of their use of word clusters suggest that translational Chinese is at best a special variant of Chinese distinct from native Chinese. They also 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.

#### 4. Reformulation markers

This section compares the use of reformulation markers in native and translational Chinese as well as in native and translational English. Reformulation markers in Chinese and English will first be discussed separately and then the results will be contrasted. Finally the explicitation of reformulation markers in English-to-Chinese translation will be investigated.

This study focuses on commonly used paraphrastic reformulation markers for elaboration and explicitation in Chinese as listed in Table 6. The list was compiled on the basis of the author’s knowledge about Mandarin Chinese as a competent native speaker. While the list may not be exhaustive, it covers all commonly used reformulation markers in Chinese. 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 6 gives the frequencies of reformulation markers in the two Chinese corpora following the human analysis.

Table 6. Literate and oral reformulation markers in LCMC and ZCTC.

| Style    | Reformulation markers | Gloss                         | LCMC | ZCTC |
|----------|-----------------------|-------------------------------|------|------|
| Literate | 即                     | namely, i.e.                  | 267  | 274  |
|          | 换言之                   | to put it differently         | 5    | 8    |
| Oral     | 也就是说                  | that is to say                | 27   | 28   |
|          | 或者说                   | or rather                     | 14   | 25   |
|          | 换句话说/换句话说讲            | 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  |

Table 6 shows 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, 1 d.f.,  $p = 0.033$ ). It is even more interesting to note the contrast between reformulation markers of different styles, which are labelled as ‘literate’ and ‘oral’ reformulation markers, following Biber’s (1988) terms. The two literate 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.<sup>7</sup> Although they are terse in form, these reformulation markers are not 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 oral reformulation markers in this study.

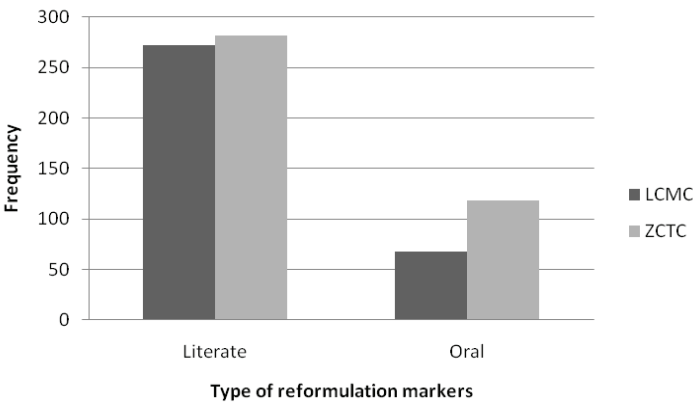


Figure 6. Literate and oral reformulation markers in LCMC and ZCTC.

As can be seen in Figure 6, although both literate and oral reformulation markers are more common in the translational corpus ZCTC, the frequencies of the literate forms of reformulation markers in the two corpora are very close and the difference is not significant ( $LL = 0.127$ , 1 d.f.,  $p = 0.772$ ), whereas the oral forms are substantially more common in translational than native Chinese ( $LL = 13.31$ , 1 d.f.,  $p < 0.001$ ).<sup>8</sup> 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).

For easy manipulation of data and comparison with English, we will regroup the fifteen genres in our corpora into two broad text categories: literature (K-R) and nonliterature (A-J). Table 7 gives the frequencies of literate and oral reformulation markers in native and translational Chinese in these two broad categories. It can be seen that in literature, literate reformulation markers are more common in native Chinese, whereas oral reformulation markers are more frequent in translated Chinese. In nonliterature, while both native and translated Chinese varieties make more frequent use of literate reformulation markers, oral reformulation markers display a greater proportion in translated (25%) than native (18%) Chinese. This means that translational Chinese differs quantitatively from native Chinese in its use of reformulation markers, especially the oral forms.

The different behaviours of literate and oral reformulation markers are more clearly illustrated in their distribution across genres in native and translational Chinese. As can be seen in Figure 7, although the overall frequencies of literate reformulation markers in the two corpora do not differ significantly, there are considerable variations across genres. Literate reformulation markers are infrequent in imaginative genres of literature (K-R) in both native and translational Chinese corpora.<sup>9</sup> In informative genres of nonliterature (A-J), they are significantly more common in news (A-C), biographies and essays (G) and reports/official documents (H) in translated texts, but much more frequent in religious writing (D), popular lore (F) and academic prose (J) in native Chinese texts.<sup>10</sup> These three genres of the second group are all formal types of writing which demonstrate a high propensity for a formal style.

**Table 7.** Distributions of literate and oral reformulation markers in Chinese.

| Text category       | Corpus | Literate reformulation markers |     | Oral reformulation markers |     |
|---------------------|--------|--------------------------------|-----|----------------------------|-----|
| Literature (K-R)    | LCMC   | 26                             | 76% | 8                          | 24% |
|                     | ZCTC   | 20                             | 42% | 28                         | 58% |
| Nonliterature (A-J) | LCMC   | 272                            | 82% | 61                         | 18% |
|                     | ZCTC   | 277                            | 75% | 91                         | 25% |

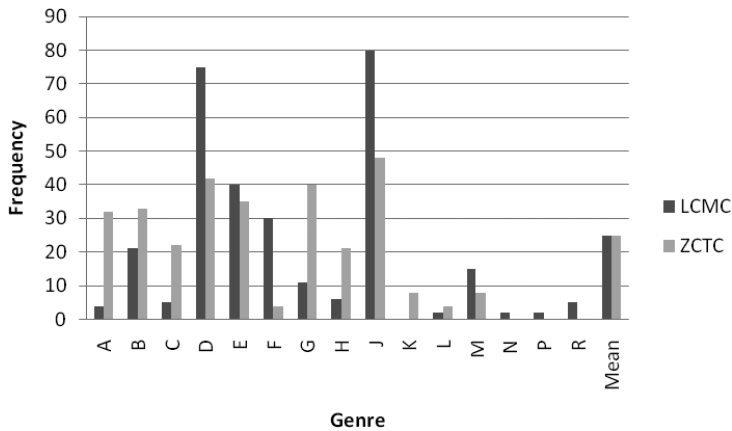


Figure 7. Distribution of literate reformulation markers across genres in Chinese.

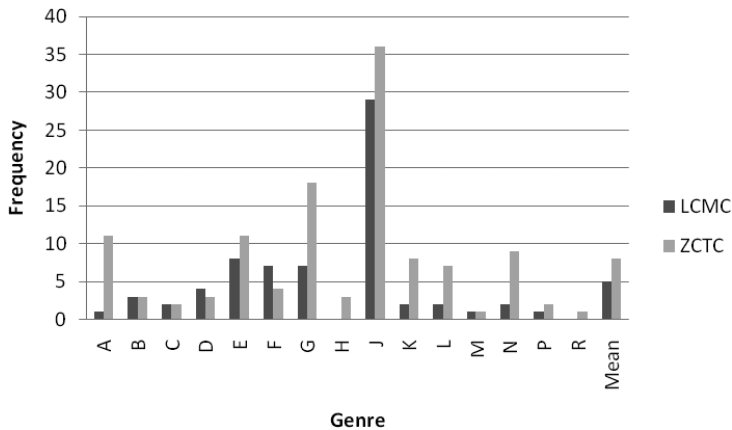


Figure 8. Distribution of oral reformulation markers across genres in Chinese.

In contrast, as can be seen in Figure 8, oral 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 literate 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 forms in comparison with native Chinese, thus providing fresh evidence for simplification in translations but serving as a counter-example of the normalization hypothesis.

The remainder of this section will compare the above findings about reformulation markers in translational Chinese with observations of reformulation markers in translational English, and explore the explicitation of reformulation markers in English-to-Chinese translation. As there is no corpus of translational

**Table 8.** Reformulation markers in English.

| Type of reformulation markers  | Examples                                                                                                                                                                                                                                                                                                                                                                                          |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Oral reformulation markers     | that is, in other words, that is to say, I mean, this means, that means, it means, which means, this meant, that meant, it meant, which meant, or more generally, to be more precise, to be more accurate, or more precisely, or more accurately, or rather, put differently, or better still, or better yet, or to say the same thing in a different way, to put it simply, to put it concretely |
| Literate reformulation markers | i.e., viz, namely                                                                                                                                                                                                                                                                                                                                                                                 |

**Table 9.** Distribution of oral and literate reformulation markers in English.

| Text category | Data type     | Literate reformulation markers |     | Oral reformulation markers |      |
|---------------|---------------|--------------------------------|-----|----------------------------|------|
| Literature    | Native        | 0                              | 0%  | 17                         | 100% |
|               | Translational | 2                              | 10% | 19                         | 90%  |
| Nonliterature | Native        | 5                              | 17% | 25                         | 83%  |
|               | Translational | 26                             | 26% | 75                         | 74%  |

English that is comparable to the FLOB corpus of native English that mirrors the relationship between LCMC and ZCTC for Chinese, we will use the BFSU General Chinese-English Parallel Corpus, a balanced bi-directional parallel corpus composed of a roughly equal amount of literary and non-literary English and Chinese originals with their respective Chinese and English translations (Wang, 2004). The comparable components in this parallel corpus will not only reveal the usage patterns of reformulation markers in native and translational English as well as in literature and nonliterature, but they will also allow us to examine reformulation markers as an explication device.

Table 8 gives a list of commonly used paraphrastic reformulation markers in English, which was established on the basis of a literature review, including all items that have been studied in the research reviewed in this article. As for Chinese, each concordance line containing these reformulation markers was evaluated manually to remove noise. Table 9 gives the frequencies of literate and oral reformulation markers in native and translated English texts. It can be seen that in both literature and nonliterature, oral reformulation markers are more frequent than literate forms, and both oral and literate forms are more commonly used in translational than native English, though the differences in the literature component are marginal.

A comparative analysis of the distribution of the two types of reformulation markers in native and translational English with that in native and translational

Chinese (Tables 7 and 9) suggests that oral reformulation markers are more common than their literate counterparts in both native and translational English, and in both literature and nonliterature, possibly because two out of the three literate forms ('viz' and 'i.e.') in this study are archaic and are largely used only in very formal writing. In contrast, literate reformulation markers are more common than oral forms in Chinese, with the exception of translated literature, which makes more frequent use of oral reformulation markers possibly to simplify the translated language. In relation to native English, both literate and oral reformulation markers are more commonly used in translational English, with an even more marked contrast in translated nonliterature. The same is true of Chinese except for the use of literate reformulation markers in translated literature, but the contrast between native and translational Chinese varieties is less marked than their English counterparts. In terms of the two broad text categories, reformulation markers are more common in nonliterature than literature in both English and Chinese, but in this case, the contrast is more marked in Chinese than in English.

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 the English-to-Chinese translations of literature and nonliterature components in the BFSU parallel corpus, totalling 814,269 English words and 677,126 Chinese words.

Table 10 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 explication rates. As can be seen, typically 10%–30% of reformulation markers are explicated (i.e. added by

**Table 10.** Explication of reformulation markers in English-to-Chinese translation.

| Reformulation markers | Glosses                         | Transferred from source text | Explicated in translation | Explication 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/<br>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                |

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 to be 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 the following examples.<sup>11</sup>

- (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  
1878 year Lewes die after Eliot again not write what stuff two year  
后, 即 1880 年, 艾略特 也 去 世 了。  
after 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 :  
US DE west frontier foster RVC one kind viewpoint i.e. think  
每 个 人 都 是 创 业 家, 或 者 说 至 少 可 以 说 每 个 人 都  
everyone all be entrepreneur or at-least may say everyone all  
有 权 试 一 试。  
have-right try one try

In examples like these, the reformulation marker 即 ‘namely, i.e.’ in Chinese translations (underlined 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.

- (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  
it include ASP 1989 year second edition DE all content i.e.  
原 先 的 牛 津 辞 典、 补 遗 部 分 及 新 增 材 料。  
former DE 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  
Butler Dr do ASP make-history DE thing i.e. she is first CL former  
人员 全为男人的 学院 中担任 校长的 女性。  
personnel 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, 即  
radical campaign reform be PSV this one kind desire start DE i.e.  
竞选 运动 要 摆脱 不受 约束 的政治 喧闹  
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  
this kind plan chiefly have two type i.e. individual pension account  
和 基奥 计划。  
and Keogh plan

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.<sup>12</sup>

## 5. Conclusions

In this article, we have taken a corpus-based approach to exploring word clusters and reformulation markers in English and Chinese on the basis of three balanced comparable corpora of the two languages in addition to an English-to-Chinese parallel corpus. The study has two focuses. The first focus relates to an inter-lingual contrastive analysis of Chinese and English, seeking to uncover the cross-linguistic differences as well as similarities between the two languages in their use

of word clusters and reformulation markers, while the second falls on an intra-lingual contrastive analysis of native and translational varieties of Chinese and English 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 word clusters are substantially more common in translated Chinese in terms of frequency, coverage as well as keyword clusters. This observation is in line with findings reported in previous translation studies such as Baker (2004) and Nevalainen (2005), suggesting a tendency in translations to use fixed and semi-fixed recurring patterns in an attempt to achieve improved fluency. The more frequent use of word clusters, especially those of high frequency and high coverage in translational Chinese, is also likely to be a result of the influence of the English source language because word clusters are significantly more prevalent in native English in relation to native Chinese.

Our corpora also reveal some qualitative difference between native and translational Chinese in terms of word cluster use. 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.

The quantitative and qualitative differences uncovered about word clusters and reformulation markers in this study between native and translated Chinese suggest 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.

Chinese and English tend to use reformulation markers of different styles while on the other hand, reformulation markers are generally more common in both translated English and translated Chinese than in their native counterparts. Our finding based on balanced comparable corpora of English and Chinese supports previous observations in some specific genres that reformulation markers function as a strategy for explicitation in translations, which also tend to use oral, stylistically simpler forms than native texts. More interestingly, it is found that literate and oral 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 oral reformulation markers, the distribution of literate 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 the study of translational Chinese exemplified in this article has brought fresh evidence and insights to translation universal research from the perspective of a language distinctly different from English, which has largely been the basis for TU research. Methodologically, it is our hope that the study has demonstrated great potential of both inter-lingual and intra-lingual contrastive analysis in linguistic research in general and in Translation Studies in particular.

## Notes

\* 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). I also thank the two anonymous reviewers for their insightful and constructive comments and suggestions, which have helped greatly to improve this article.

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 2 for further discussion of the term in narrow and broad senses.

2. 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.

3. I would like to thank an anonymous reviewer for pointing this out to me.

4. A higher minimum is specified for FLOB because clusters of function words in English usually have very high coverage (e.g. 100%).

5. The difference between academic writing and news in LCMC and ZCTC is not significant ( $LL = 1.25$ , 1 d.f.,  $p = 0.26$ ).

6. The 3-word cluster 蒲式耳 ‘bushel’ in Table 3 is a technical term unknown to the ICTCLAS tagger used to annotate our Chinese corpora, which incorrectly tokenized it as three separate words. In the table, the character string ‘S A N’ is an acronym for ‘Storage Area Network’.

7. The literate 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 literate 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).

8. The stylistic differences between literate and oral reformulation markers in Chinese, unfortunately, cannot be retained in their English glosses.
9. Because of the low overall frequencies, the differences between the native and translational corpora are not statistically significant in imaginative genres.
10. The differences between native and translated Chinese in B (i.e. news editorials,  $LL=2.62$ , 1 d.f.,  $p=0.106$ ) and E (i.e. instructional writing,  $LL=0.36$ , 1 d.f.,  $p=0.547$ ) are not significant.
11. 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 is used before a verb to form a nominal construction.
12. Syntactically, Chinese grammar is much more tolerant of the so-called run-on sentences than English.

## References

- Aijmer, K. 2007. "The Meaning and Functions of the Swedish Discourse Marker *alltså* — Evidence from Translation Corpora". *Catalan Journal of Linguistics* 6: 31–59.
- Baker, M. 1996. "Corpus-based Translation Studies: the Challenges that Lie Ahead". In *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, H. Somers (ed), 175–186. Amsterdam: John Benjamins.
- Baker, M. 2004. "A Corpus-based View of Similarity and Difference in Translation". *International Journal of Corpus Linguistics* 9(2): 167–193.
- Baker, M. 2007. "Patterns of Idiomaticity in Translated vs. Non-translated text". *Belgian Journal of Linguistics* 21: 11–21.
- Baroni, M. and Bernardini, S. 2003. "A Preliminary Analysis of Collocational Differences in Monolingual Comparable Corpora". In *Proceedings of Corpus Linguistics 2003*, D. Archer, P. Rayson, A. Wilson and T. McEnery (eds), 82–91. Lancaster: UCREL, Lancaster University.
- Biber, D. 1988. *Variations across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. 1995. *Dimensions of Register Variation: a Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
- Biber, D. 2006. *University Language: a Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Biber, D., Conrad, S. and Cortes, V. 2004. "If You Look at...: Lexical Bundles in University Teaching and Textbooks". *Applied Linguistics* 25(3): 371–405.
- Blakemore, D. 2007. "'Or'-parentheticals, 'that is'-parentheticals and the Pragmatics of Reformulation". *Journal of Linguistics* 43(2): 311–339.
- Chen, W. 2006. *Explicitation through the Use of Connectives in Translated Chinese: a Corpus-based Study*. PhD thesis, University of Manchester.
- Cortes, V. 2002. "Lexical Bundles in Freshman Composition". In *Using Corpora to Explore Linguistic Variation*, R. Reppen, S. Fitzmaurice and D. Biber (eds), 131–145. Amsterdam: John Benjamins.
- Cuenca, M.-J. 2003. "Two Ways to Reformulate: a Contrastive Analysis of Reformulation Markers". *Journal of Pragmatics* 35(7): 1069–1093.

- De Cock, S. 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, S. 2000. "Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing". In *Corpus Linguistics and Linguistic Theory. Papers from ICAME 20 1999*, C. Mair and M. Hundt (eds), 51–68. Amsterdam: Rodopi.
- Del Sz, M-M. and Fraser, B. 2005. "Reformulation in English". Paper presented at the 9th International Pragmatics Conference, Riva Del Garda, Italy, 10–15 July 2005.
- Granger, S. 1996. "From CA to CIA and Back: an Integrated Approach to Computerized Bilingual and Learner Corpora". In *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, March 1994*, K. Aijmer, B. Altenberg and M. Johansson (eds), 37–51. Lund: Lund University Press.
- Granger, S. 1998. "Prefabricated Writing Patterns in Advanced EFL Writing: Collocations and Formulae". In *Phraseology: Theory, Analysis and Applications*, A. Cowie (ed), 145–160. Oxford: Clarendon Press.
- Hundt, M., Sand, A. and Siemund, R. 1998. *Manual of Information to Accompany the Freiburg-LOB Corpus of British English*. Freiburg: University of Freiburg.
- Kenny, D. 1998. "Creatures of Habit? What Translators Usually Do with Words". *Meta* 43(4): 515–523.
- Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.
- Mauranen, A. 2007. "Universal Tendencies in Translation". In *Incorporating Corpora. The Linguist and the Translator*, M. Rogers and G. Anderman (eds), 32–48. Clevedon: Multilingual Matters.
- Mauranen, A. and Kujamäki, P. (eds). 2004. *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.
- McCarthy, M. and Carter, R. 2004. "This That and the Other: Multi-word Clusters in Spoken English as Visible Patterns of Interaction". In *Explorations in Corpus Linguistics*, M. McCarthy (ed.), 7–26. Cambridge: Cambridge University Press.
- McEnery, T. and Xiao, R. 2004. "The Lancaster Corpus of Mandarin Chinese: a Corpus for Monolingual and Contrastive Language Study". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*. Lisbon, 24–30 May 2004. 1175–1178.
- Murillo, S. 2004. "A Relevance Reassessment of Reformulation Markers". *Journal of Pragmatics* 36(11): 2059–2068.
- Mutesayire, M. 2005. *Cohesive Devices and Explicitation in Translated English — A Corpus-based Study*. PhD thesis, University of Manchester.
- Nevalainen, S. 2005. "Köyhtyykö Kieli Käännettäessä? Mitä Taajuuslistat Kertovat Suomenosten Sanastosta". In *Käännössuomeksi*, A. Mauranen and J. H. Jantunen (eds), 141–162. Tampere: Tampere University Press.
- Salkie, R. 2002. "How Can Linguists Profit from Parallel Corpora". In *Parallel Corpora, Parallel Worlds*, L. Borin (ed), 93–109. Amsterdam: Rodopi.
- Schourup, L. 1999. "Discourse Markers". *Lingua* 107: 227–265.
- Scott, M. 2009. *The WordSmith Tools* (version 5.0). Oxford: Oxford University Press.
- Sperber, D., and Deirdre, W. 1995. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

- Wang, K. 2004. "Han Ying Duiying Yuliaoku de Sheji (Design of the Chinese-English Parallel Corpus)". In *Shuangyu Duiying Yuliaoku Yanzhi yu Yingyong (The Development and Application of Bilingual Parallel Corpora)*, K. Wang (ed), 36–53. Beijing: Foreign Language Education and Research Press.
- Xiao, R. 2007. "What Can SLA Learn from Contrastive Corpus Linguistics? The case of Passive Constructions in Chinese Learner English". *Indonesian Journal of English Language Teaching* 3(2): 1–19.
- Xiao, R. 2009. "Multidimensional Analysis and the Study of World Englishes". *World Englishes* 28 (4): 421–450
- Xiao, R. 2010. "How Different is Translated Chinese from Native Chinese?". *International Journal of Corpus Linguistics* 15(1): 3–33.
- Xiao, R. and Tao, H. 2007. *The Lancaster Los Angeles Spoken Chinese Corpus*. Lancaster: UCREL.
- Xiao, R. and Yue, M. 2009. "Using Corpora in Translation Studies: the State of the Art". In *Contemporary Corpus Linguistics*, P. Baker (ed), 237–262. London: Continuum.
- Xiao, R., He, L. and Yue, M. 2010. "In Pursuit of the Third Code: Using the ZJU Corpus of Translational Chinese in Translation Studies". In *Using Corpora in Contrastive and Translation Studies*, R. Xiao (ed), 182–214. Newcastle: Cambridge Scholars Publishing.

#### *Author's address*

Richard Xiao  
Department of English  
Edge Hill University  
Ormskirk  
Lancashire, L39 4QP  
United Kingdom

[Richard.Xiao@edgehill.ac.uk](mailto:Richard.Xiao@edgehill.ac.uk)