NCUE Journal of Humanities
Vol. 9, pp. 253-271
March 2014

Simplification in Translated Chinese Texts: A Corpus-Based Study on Mean Sentence Length

Ting-hui Wen[*]

Abstract

Mean sentence length is proposed as a measure of simplification: if a text has shorter mean sentence length, it is assumed to be simpler for readers to comprehend. Since translated texts are hypothesized to be simpler than non-translated text, they would presumably exhibit shorter mean sentence length. The present research aims to investigate, using corpus-based methods, the phenomenon of simplification in translated, compared to non-translated, Chinese texts. This paper focuses on measuring sentence length of the translated texts in the Corpus of Comparable Mystery Fiction, analyzing the results of mean sentence length and its additional measures: mean sentence length in terms of characters, mean sentence sub-unit length in terms of words and mean sentence sub-unit length in terms of characters.

The measure of mean sentence length and its additional measures render consistent results showing that the translated texts of the corpus under study exhibit shorter sentence length than the non-translated texts.

Keywords: corpus-based translation studies, simplification, mean sentence length

[*] Assistant Professor, Graduate Institute of Translation and Interpretation, National Changhua University of Education.

Received December 31, 2013.

國立彰化師範大學文學院學報
第九期，頁 253-271
二〇一四年三月

漢語翻譯文本中的簡化現象：以語料庫方法探究平均句長

溫婷惠[*]

摘要

平均句長是簡化現象的測量方法之一：如果一個文本的句長較短，通常被認為較簡單，讓讀者易於了解。假設翻譯文本比非翻譯文本簡單，翻譯文本的平均句長就比較短。本研究旨在以語料庫方法研究漢語翻譯文本中的簡化現象，而且特別針對平均句長做研究。本研究所使用的語料庫為自建的懸疑小說可比語料庫（The Corpus of Comparable Mystery Fiction），分析平均句長（以詞計算）以及其延伸的測量方法：以字計算的平均句長，以詞計算的平均小句長和以字計算平均小句長。

分析平均句長以及其延伸的測量方法的結果皆顯示本語料庫的翻譯文本的句長比非翻譯文本的句長短。

關鍵詞: 語料庫翻譯學、簡化現象、平均句長

[*]國立彰化師範大學翻譯研究所助理教授。

到稿日期：2013 年 12 月 31 日。

Introduction

In 1996, Mona Baker proposed four recurrent translation features: explicitation, simplification, normalisation and levelling out. Based on Baker's hypothesis, Sara Laviosa-Braithwaite (1996) then created the Translational English Corpus (TEC), adopted four measures from corpus linguistics—i.e., type/token ratio, percentage of high frequency words, lexical density and mean sentence length, to investigate simplification in translated texts. According to Laviosa-Braithwaite, translated texts tend to exhibit a higher type-token ratio, a higher percentage of high frequency words, a lower lexical density and a shorter mean sentence length.

The Comparable Corpus of the Chinese Mystery fiction (hereafter the CCCM, please refer to Appendix 1 for a complete list of texts included in the corpus under study), was created especially for the purpose of this study. Mystery fiction as a sub-genre was chosen as the object of this study due to its manageable size, its popularity in Taiwan and personal interest of the researcher. The CCCM consists of two subcorpora, one translated and one non-translated. The sizes of the two subcorpora are listed in Table 1:

Subcorpus	Number of texts	Number of words	Number of characters
Translated	8	787,128	1,191,734
Non-translated	12	823,600	1,245,455

Table 1 The size of the corpus

This paper focuses on one of the measures of simplification, i.e., mean sentence length, analyzing the results of mean sentence length and its additional measures: mean sentence length in terms of characters, mean sentence sub-unit length in terms of words and mean sentence sub-unit length in terms of characters. Mean sentence length is proposed as one of the measures of simplification: if a text has shorter mean sentence length, it is assumed to be simpler for readers to comprehend. Since translated texts are hypothesized to be simpler than non-translated texts, they would presumably exhibit shorter mean sentence length.

The texts in the CCCM have to be word-segmented before mean sentence length of a text is measured. Therefore, the basic concept of a Chinese word is discussed and the application of mean sentence length to the Chinese language is explained in section II. Mean sentence length in terms of Chinese characters is then proposed as an additional measure for the following two reasons: the definition of words in Chinese is still controversial; and the program of word-segmentation cannot achieve a hundred per cent accuracy.

Section III first presents the definition of a Chinese sentence. Punctuation is a relatively new and not well-developed concept in the Chinese language. A Chinese sentence is more a "discourse" unit: it can be as long as a whole paragraph and include different information relating to the same topic. Therefore, in addition to mean sentence length, an additional measure of mean sentence sub-unit length, with sentence sub-units defined as the segments between two commas, full stops, exclamation marks, question marks, semicolons, and colons, is proposed due to special characteristics of Chinese sentences. Mean sentence sub-unit length is also measured both in terms of words and characters.

The hypothesis, the results and tests of statistical significance of mean sentence length in terms of words (section IV), mean sentence length in terms of characters (section V), mean sentence sub-unit length in terms of words (section VI), and mean sentence sub-unit length in terms of characters (section VII) are discussed in detail from sections IV to VII respectively.

II. Measuring mean sentence length in Chinese

1. Definition of Chinese characters and words

Since "words" are basic units in calculating mean sentence length, a discussion of what constitutes a Chinese word is appropriate here. In written English, words are formed by the continuation of a series of letters, with a break or space indicating separate units. Unlike English, Chinese words are formed by "characters," rather than alphabetic letters, and there is no space between Chinese characters or words. Example (1a)[1] shows one Chinese sentence and its translation, and as noted, there are no spaces between words or characters.

(1a) 這幾年走在台北街頭，人們可以明顯感覺到日本觀光客似乎變多了，無論在故宮、永康街鼎泰豐、夜市、或各個捷運站，常可見東瀛客手持旅遊手冊，用好奇的眼神按圖索驥，期待貼近台灣民間的生活氣味。

Over the last few years, there has been a perceptible increase in the number of Japanese tourists in Taipei. You see them everywhere—the National Palace Museum, the famed restaurant DinTaiFung on Yungkang Street, night markets, MRT stations—taking in the sights with guidebooks in hand, trying to get a glimpse of Taiwanese life.

If words are not segmented, *WordSmith* will count example (1a) as one word only. The first step to calculate mean sentence length of a Chinese text is thus to delimit Chinese words and segment them, inserting spaces between individual words. Therefore, Chinese words have to be defined first before they are segmented.

Chinese words can be categorised into four types, according to morphological rules: simple words, complex words, compound words, and repeated words (Tang 1988). Each type of word is discussed below.

A **simple word**, such as *pao* (跑, "run"), *shuo* (說, "say"), *ren* (人, "people"), *gao* (高, "tall"), *bo-li* (玻璃, "glass"), *pu-tao* (葡萄, "grapes") and *wei-sheng-su* (維生素, "vitamin"), is a word formed by only one morpheme, which can include one or more characters; in extreme cases (usually transliteration of foreign terms), one morpheme can consist of several characters, such as the proper nouns *nuo-si-te-la-da-mu-si* (諾斯特拉達姆斯) "Nostradamus" and *jiu-mo-luo-shi* (鳩摩羅什) "Kumaarajiiva". In both of these cases, the individual Chinese characters used to transliterate foreign names are not to be "read" as having separate meanings (which they might in other contexts); rather, they are treated as one unit. In the cases of *bo-li* (玻璃, "glass") and *pu-tao* (葡萄, "grapes"), the two characters together form a unit that means one thing; they are almost never used separately, and both must be present to form the meaning.

A **complex word**, such as *yi-zi* (椅 子, "chair"), *zhuo-zi* (桌 子, "desk/ table"), *lao-shi* (老 師, "teacher"), *lao-ba* (老 爸, "father"), *keneng-xing* (可能 性, "possibility") and *diannao-hua* (電腦 化, "computerisation"), is formed by one free morpheme and an affix. The affix may be at the end (*zi* 子, underlined in the first two examples) or the beginning (*lao* 老, underlined in the third and fourth examples). This may be likened to words in English such as "performer", formed from the morpheme "perform", plus the affix "-er". Many new words in Chinese have been coined in the twentieth and twenty-first century in this manner, following the morphology of European languages, as

[1] Example (1a) and its English translation is extracted from "The new tourism—Young Japanese discover Taiwan" by Chang, *Taiwan Panorama*, the issue of March 2006, pp. 34.

in the last two examples, where *xing* (性, “nature”) is used as an analogy for “-ity” in English, and *hua* (化, “change”) has been used as an analogy for “-isation”.

Compounding is a productive morphological process in Chinese; a compound word is formed by two free morphemes. Different types of compounds abound in Chinese, for example, Verb-Verb, such as *da-kai* (打 開; literally “hit-open”, i.e. “open”), Verb-Noun such as *sheng-qi* (生 氣; literally “bear-anger”, i.e. “angry”), Noun-Verb such as *xia-zhi* (夏 至; literally “arrival of summer”, i.e. “summer solstice”), and Noun-Noun *shu-fang* (書 房; literally “book-room”, i.e. “study”) (Tang 1988). Some terms, such as *mi-yue* (蜜月, “honey-moon”) and *tu-chun* (兔 唇, “hare-lip”) are translated according to the rules of compounding, just as their English equivalents are.

A **repeated word**, such as *cong-cong-mang-mang* (匆 匆 忙 忙, “hurry”) and *shi-shi-kan* (試 試 看, “try”), is formed by the repetition of one or two morphemes in the same word.

After this brief discussion of the nature of Chinese words, example (1b) shows the sample sentence about Japanese tourists in Taipei (example [1a] above) after spaces are inserted between words.

(1b) 這 幾年 走 在 台北 街頭 ， 人們 可以 明顯 感覺到 日本 觀光客 似乎 變多 了 ， 無論 在 故宮 、 永康街 鼎泰豐 、 夜市 、 或 各 個 捷運站 ， 常 可見 東瀛客 手持 旅遊 手冊 ， 用 好奇 的 眼神 按圖索驥 ， 期待 貼近 台灣 民間 的 生活 氣味 。

After words are segmented, there are 43 words in example (1b).

However, four issues emerge from example (1b): long compound words; *chengyu* and fixed expressions; proper nouns; and abbreviations.

First, **compounds** can be very long in modern Chinese and yet be treated as single words. In example (1b), *dong-ying-ke* (東瀛客, “Japanese tourist”) and *jie-yun-zhan* (捷運站, “Metropolitan Rapid Transit station”) include three Chinese characters, and are treated as single words. Extreme examples can be found in the Sinica Corpus: *guo-ji-hui-yi-ting* (國際會議廳, “international conference room”) and *di-er-ci-shi-jie-da-zhan* (第二次世界大戰, “The Second World War) include five or more characters and are treated as single words.

Second, ***chengyu*** is a set of traditional idiomatic expressions, consisting mainly of four characters. In (1b), *an-tu-suo-ji* (按圖索驥; literally “to follow a drawing to find a steed”) actually means to try to locate something by following up a clue. These *chengyu* are typically derived from sentences in classical Chinese, but then four main characters were extracted and now treated as a single “word”. For example, *gua-tian-li-xia* (瓜田李下; literally “in a melon patch, under a plum tree”) was derived from two lines of a poem *guo-tian bu na lü, li-xia bu zheng guan* (瓜田不納履, 李下不整冠; “don’t adjust your shoes in a melon field and don’t tidy your hat under the plum trees–i.e., to avoid being suspected of stealing the melons and plums”), which describe a code of conduct to avoid anything to arouse suspicion.

Moreover, other **fixed expressions** might include more characters than *chengyu*, and yet still be treated as one word. For example, *yi-bu-zuo-er-bu-xiu* (一不做二不休, “to carry the thing through, whatever the consequences are”) includes six characters and might be treated as one word, while the similar English expression “in for a penny, in for a pound” is treated as eight words.

Third, long and complex **proper nouns** might also be treated as single words. For example, the names of places, such as *Yong-kang-jie* (永康街, “Yungkang Street”) and *bei-ka-luo-lai-na-zhou* (北卡羅來納州, "North Carolina") include three and six characters respectively, and might be treated as single words. The full names of Chinese people, such as *Li-yuan-ze* (李遠哲, "Lee, Yuan-Tseh") and Jiang-jing-guo (蔣經國, "Chiang, Ching-kuo"), might be treated as single words.

Finally, **abbreviations** are classified as a type of compound and treated as single words. For example, *gu-gong* (故宮) is an abbreviation of *guo-li gu-gong bo-wu-yuan* (國立 故宮 博物院, "National Palace Museum"), and *bei-yue* (北約) is an abbreviation of *bei-da-xi-yang gong-yue zu-zhi* (北大西洋 公約 組織, "North Atlantic Treaty Organization").

Whether these four types of words can really be treated as single words is a continued source of controversy among Chinese linguists. The Bureau of Standards, Metrology and Inspection, Ministry of Economic Affairs in Taiwan has authorised the Association for Computational Linguistics and Chinese Language Processing to draft *The Standard of Chinese Word Segmentation for Information Processing*, in order to establish a national standard for Chinese word processing. However, it only serves as a guideline, and its application varies. For example, in the largest Chinese corpus in Taiwan, the *Academia Sinica Balanced Corpus of Modern Chinese*, *lian-he-guo* (聯合國, "United Nations") is treated as one word while *bei-da- xi-yang gong-yue zu-zhi* (北大西洋 公約 組織, "North Atlantic Treaty Organization" is treated as three words; *bei-ka-luo-lai-na-zhou* (北卡羅來納州, "North Carolina") is sometimes treated as one word, and sometimes as two: *bei* (北, "North") and *ka-luo-lai-na-zhou* (卡羅來納州, "Carolina"). A Chinese name, such as Jiang-jing-guo (蔣經國, "Chiang, Ching-kuo") is treated as one word in the Sinica Corpus, while an English name, such as Bill Gates, is transliterated into *bier · gaizi* (比爾 · 蓋茲) and counted as two words. The inconsistency and controversy of the concept of words thus exerts an influence on the measures of simplification. For the purpose of this study, *Autotag*, the software is used to segment words and *The Standard of Chinese Word Segmentation for Information Processing* is adopted as the guideline when segmenting words for both subcorpora of the CCCM. An additional/alternative measure in terms of Chinese characters is proposed to solve the problem of inconsistency and inaccuracy in segmenting words (see the following section for a detailed discussion of measuring mean sentence length in terms of Chinese characters).

2. Measuring mean sentence length in terms of Chinese characters and words

For English speakers, the 'word' is a salient and intuitive concept; they can easily distinguish words by writing conventions: spaces must be inserted between words. However, for Chinese speakers, instead of *ci*, words, the basic unit for Chinese written language is *zi*, characters, which generally represent morphemes rather than "words." As noted by Jerome L. Packard (2000: 15), the status of characters in Chinese is as salient and robust as the status of words in English.

Packard gives examples of character puzzles in Chinese, instead of crossword puzzles (ibid). He also emphasises the fact that dictionaries and databases are arranged and searched according to characters: entries of a Chinese dictionary are always based on characters. The basic definition of a character is listed under each character, and the definitions and usages of different words consisting of the same character are listed as sub-entries underneath.

Moreover, from 1981 to 1997, there was a TV programme, called *Mei Ri Yi Zi* "*A Character a Day*", which was produced and broadcast by The Chinese Television System in Taiwan. During the seventeen years, more than 1,600 characters were introduced to

Taiwanese students and foreigners learning the Chinese language. Characters are often perceived by learners to be the basic unit of the Chinese language.

Even in the world of computing, *Microsoft Word* has a tool for counting words in a document, but it actually counts characters in a Chinese document, instead of words. Writers and translators translating from different languages into Chinese are usually paid according to the characters they write/translate instead of words.

Calculating mean sentence length in terms of characters, instead of words, might also be an index in investigating simplification in the Chinese language. Since a Chinese word is formed by characters, the number of characters a word has will also have an influence on the total length of a sentence. For example, the issue of long compounds, *chengyu* and fixed expressions, and abbreviation can also be solved by calculating mean sentence length in terms of characters. Take the following two sentences (2)[2] and (3)[3] for example.

(2) 後來 ， 他們 在 一 條 偏僻 的 街道 上 看見 一 家 小 旅館 ， 以為 可以 找到 過夜 的 地方 ， 偏偏 這 家 旅館 也 客滿 了。

They finally saw a small hotel on a deserted street and thought that they could find a place to stay overnight. However, this hotel was also full.

(3) 就是 這 種 拋棄 科技 產品 令人生畏 的 繁複 功能 ， 方便 阿公 、 阿嬤 隨時 看見 生活 的 快樂 記憶 ， 在 歐洲 一舉 攻下 該 項 產品 的 一半 市場。

Dispensing with daunting, complicated functions makes it easier for grandma and grandpa to look back upon fond memories of their lives, and this product has captured half the market in Europe.

Example (2) is a sentence extracted from a Chinese textbook for Grade Four students in Taiwan, which contains 27 words and 41 characters, while example (3) is a sentence extracted from *Taiwan Panorama*, a bilingual magazine in Taiwan written for adult readers, which includes 28 words, but a significantly higher number of characters (52). Example (2) has only one more word than example (3), but it contains 11 more characters, which might indicate that sentences written for adult readers tend to use longer words which consist of more characters.

It might render insightful results if we count mean sentence length in terms of characters, which also takes word length somewhat into consideration. As shown in examples (2) and (3), if sentence length is calculated in terms of words, the difference is not obvious: example (2) is only one word less than example (3). If sentence length is calculated in terms of characters, example (2) is eleven characters less than example (3), which at the same time reflects that example (3) consists of longer words and might be more difficult for readers to comprehend.

Therefore, in this study, the texts included in the two subcorpora of the CCCM are both word-segmented and character-segmented, and sentence length is computed twice, once using the number of words, and once using the number of characters, to see if any significant patterns emerge.

III. Mean sentence sub-unit length as an additional measure

1 The definition of a Chinese sentence

[2] Example (2) is a Chinese sentence extracted from *Chinese VIII*, Taipei: National Institute for Compilation and Translation, and translated by the researcher of the current study.

[3] Example (3) and its translation is extracted from “Philips’ simplicity revolution” by Teng, *Taiwan Panorama*, the issue of May 2007, pp. 22.

Chinese linguists have tried to define what constitutes a Chinese sentence from a functional perspective: Li and Thompson (1981) pointed out that Chinese is a discourse-oriented language whereas English is a sentence-oriented language. It should be noted that a comma in the Chinese language can occur after a phrase, a clause or even a sentence, and indicate a pause for readers, while a full stop in Chinese sometimes indicates a larger linguistic unit than a sentence in English. Therefore, a Chinese sentence should be regarded as "a discourse unit consisting of several information units bearing some relation to the same topic" (Gao 1997: 11).

The use of punctuation marks in the Chinese language was introduced and proposed in the early twentieth century following the vernacularization of Chinese. Therefore, punctuation is still a relatively new and not well-developed concept in the Chinese language. Sometimes a comma "，" appears in a position where a full stop "。" is expected, simply because frequent use of full stops in a paragraph would be considered awkward. It is actually not uncommon to see a whole Chinese paragraph with only one full stop, as shown in example (4). Example (4)[4] is a paragraph extracted from *Taiwan Panorama*. The texts in this magazine are first written in Chinese, and then translated into English and other languages.

(4)「設計師對空間永遠是貪得無饜的，」身穿粉紅色襯衫、牛仔褲、帆布鞋，一派輕鬆打扮的華碩工業設計部副理李政宜說，幾年前華碩新大樓規劃好後，由於立體隔板的隔間不利於團隊溝通討論，設計部門犧牲了個人隱私，自願搬來舊大樓，自己動手作室內設計，打掉一面牆，引進大片光源，讓視線延伸到戶外；常埋首於電腦螢幕前的設計師，累了就可以走到戶外陽台呼吸新鮮空氣或者吞雲吐霧，遠眺關渡平原的落日餘暉，偶而還會開拔到水鳥公園開會，一邊討論，一邊看著水鳥在旁飛來飛去。

(4a)"Designers are never satisfied with either the size or layout of their workspaces," says Li Cheng-yi, deputy director of Asus Design.

(4b)He says that when Asustek completed plans for its new building a few years ago, they called for offices separated by solid dividers.

(4c) But such a layout doesn't lend itself to discussion and communication among the members of a team.

(4d) Upon consideration, Asus Design decided to forego the privacy that the new offices would have afforded and instead moved into the old building.

(4e) There, they redesigned their workspace by knocking out a wall, installing large light sources, and opening up sightlines to the outdoors.

(4f) The new layout allowed the designers, who spend most of their time in front of computer monitors, to walk out onto an exterior balcony for a breath of fresh air or a smoke, or for a glimpse of sunset on the Kuantu Plain.

(4g) The department now sometimes even holds meetings in Kuantu's waterfowl refuge, alternating between discussing issues and watching the birds.

As we can see from example (4), there is only one full stop "。" in the whole paragraph, which means that by Laviosa-Braithwaite's (1996) definition, there is only one sentence in this example. However, it is translated into seven sentences in English, as illustrated in (4a), (4b), (4c), (4d), (4e), (4f) and (4g). In example (4), the Chinese source text could also be divided into seven or more grammatical sentences as its English translation and end with a full stop under the syntactic notion of a sentence, i.e., a set of expressions consisting minimally of a noun phrase, followed by an auxiliary, followed by

[4] Example (4) and its translation are extracted from "Taiwanese design takes flight" by Teng, *Taiwan Panorama*, the issue of May 2007, pp. 6.

a verb phrase in deep structure (Fromkin *et al.* 2003: 594). The writer and the editor, however, preferred to use commas and semicolons to group all these units into one Chinese sentence only because they bear information relating to the same topic, i.e., the office layout.

According to Show-lin Lin (2002: 95), articles written for adult readers have to be simplified to be included in the Chinese textbooks for students aged from 6 to 15 by using shorter sentences. Therefore, counting mean sentence length might still serve as an appropriate measure of simplification if we are comparing translated and non-translated texts both written in the Chinese language. We continue to define a Chinese sentence in the same way as an English sentence, for both of the two subcorpora, taking words between two full stops, exclamation marks and question marks as one sentence, to see whether translated texts have significantly shorter sentences than texts composed in Chinese.

In order to apply this measure to the Chinese language and make it possible for *WordSmith* to calculate mean sentence length automatically, the first step is that the texts should be word-segmented. Then, since *WordSmith* does not recognise Chinese punctuation marks, all the marks are substituted with their English counterparts.

2. Additional measure: mean sentence sub-unit length

As we have discussed in section 3.1, a Chinese sentence, i.e., words between two stops, exclamation marks and question marks, is actually a much larger linguistic unit than an English sentence. In Chinese, commas are usually used to connect sentences, where full stops are usually used in English (Tsao 1979, cited in Lin 2002: 17).

Unlike commas in English, which can be used in a parallel construction to separate words and short phrases, commas in the Chinese language are usually used to separate larger units, such as longer phrases, clauses and sentences. Semi-pauses (、), a punctuation mark which is specific to the Chinese language, are usually used to separate words and short phrases in a parallel construction.

According to the analysis of an article in a Chinese textbook for junior high school students, Lin (2002) discovered that 66% of the commas are used in this article to connect *xiao ju* (literally “small sentences”, or “short sentences”, similar to clauses in English). Example (5) is one Chinese sentence consisting of three *xiao ju* connecting with two commas (punctuation marks in its English translation are used according to the Chinese sentence).

(5) 外面正下著雨，雨雖然不大，卻一直沒有停歇。

(Literal translation) It is raining outside, although the rain is not heavy, it never stops.

A comma can sometimes be used after a phrase as well. In example (6), the phrase is underlined, and punctuation marks in its English translation are used according to the Chinese sentence.

(6) 那時候，他是小學三年級吧，阿地他們已是國中生了。

(Literal translation) At that time, he was in his third grade, A-di and others were junior high school students.

Segments between commas in the Chinese language tend to indicate smaller sub-units than a full Chinese sentence, and it might render insightful results to measure mean length of these sentence sub-units.

In classical Chinese, although no punctuation marks were officially employed, *judou* was often marked after the Han Dynasty (206 BC–220 AD): *ju* (句), marked as “○”, is similar to a full stop, usually indicated at the end of discourse; *dou* (讀), marked as “ 、 ”, is similar to commas, usually indicating a short break. *Judou* was used to clarify ambiguity and to increase comprehension, and words between *judou* are similar to the sentence sub-units proposed here.

In order to measure mean sentence sub-unit length automatically using *WordSmith*, punctuation marks which indicate these sub-units, i.e., commas (，), semicolons (；) and colons (：), are replaced with English full stops (.). The results of mean sentence length calculated by *WordSmith* after the replacement of punctuation marks are actually the results of mean sentence sub-unit length. As with sentence length, we also measure it in terms of both words and characters.

Examples (2) and (3) in section II have almost the same sentence length (27 and 28 words respectively). If we replace the punctuation marks with full stops, the mean sentence sub-unit length in example (2a) is 6.75 words and 10.25 characters, while the mean sentence sub-unit length for example (3a) is 9.33 words and 17.33 characters.

(2a) 後來 . 他們 在 一 條 偏僻 的 街道 上 看見 一 家 小 旅館 . 以為 可以 找到 過夜 的 地方 . 偏偏 這 家 旅館 也 客滿 了 .

(3a) 就是 這 種 拋棄 科技 產品 令人生畏 的 繁複 功能 . 方便 阿公 阿嬤 隨時 看見 生活 的 快樂 記憶 . 在 歐洲 一舉攻下 該 項 產品 的 一半 市場 .

Example (2a), the sentence written for Grade Four students, has fewer words and much fewer characters in terms of sentence sub-unit length than example (3a), the sentence written for adult readers. Therefore, the mean sentence sub-unit length might also serve as an index in measuring simplification in the Chinese language.

Moreover, since the translated texts in the CCCM are translations from English source texts, it might be expected that the punctuation marks of the source texts, especially full stops, exerted a great influence on the translations. The translated texts would have shorter mean sentence length as a result of the interference from their source texts. Measuring mean sentence sub-unit length of both the subcorpora of the CCCM would render more robust results, free from the influence of the English source texts on the translations.

IV. Mean sentence length in terms of words

1. Hypothesis

Since simplified texts tend to have shorter mean sentence length in terms of words, and we might assume that translated texts tend to be simpler than non-translated texts, it can be hypothesized that translated texts have shorter mean sentence length in terms of words than non-translated texts.

2. Results

First, all the word-segmented texts were processed by *WordSmith.* The overall results of the two subcorpora are shown in Figure 1.

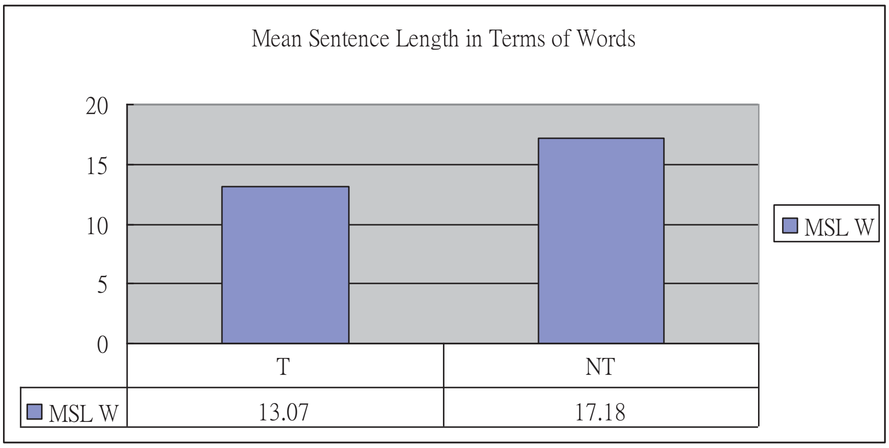

Figure 1 The results of mean sentence length in terms of words

The mean sentence length of the translated subcorpus is 13.07 words, 4.11 words fewer than the mean sentence length of the non-translated subcorpus (17.18 words). The translated subcorpus has shorter mean sentence length in terms of words than the non-translated one, as we have predicted in the hypothesis. However, a statistical test is required to confirm that the difference between the two subcorpora is statistically significant and does not happen by chance.

3. Statistical tests

In statistics, significance means "reaching a degree of statistical certainty at which it is unlikely that a result is due purely to chance" (Oakes, 1998: 255). There are several methods of testing significance in statistics, such as the *Chi*-square test, the *t*-test, and the *z*-test (Butler, 1985; Woods et al., 1986; Oakes, 1998). Since we are measuring mean sentence length of the two subcorpora in the CCCM, and there is only one variable concerned (mean sentence length), and the sample size of the corpora is large, according to the Central Limit Theorem[5], the *z*-test is employed to test statistical significance[vi] (Butler, 1985: 78-83; Oakes, 1998: 250; Baroni and Evert, 2008).

The significance level of a statistical test is the probability level below which the observed differences are treated as significant, and in linguistics, where the measurement is often less exact, a significance level of 0.05 (written as $\alpha = 0.05$) is common (Butler 1985: 71). Therefore, the significance level of the current study is set as $\alpha = 0.05$.

	Translated (word)	Non-translated (word)
Sample size	60138	47891
Mean sentence length	13.07	17.18
Standard deviation	10.17	14.32
Z score	-53.08	

Table 1 The statistical test of mean sentence length in terms of words

After applying the formula, *z* = -53.08, it can be concluded that mean sentence length in terms of words of the translated population is significantly shorter than that of the non-translated population.

[5] A test of significance can only be applied when the populations from which the samples are taken are normally distributed, but according to the Central Limit Theorem, which states that "when samples are repeated drawn from a population, the means of the samples will be normally distributed around the population mean" (Oakes 1998: 250), the requirement can be relaxed in the case of large samples. In other words, if the sample size is large, then the Central Limit Theorem will assure the validity of the test.

The conclusion supports the hypothesis that modern translated mystery fiction in Taiwan (with the source texts in English) tends to have shorter mean sentence length in terms of words than the modern non-translated mystery fiction, and therefore, might be simpler for readers to comprehend, which serves as an index of simplification.

V. Mean sentence length in terms of characters

1. Hypothesis

Since simplified texts tend to have shorter mean sentence length in terms of characters, and we might assume that translated texts tend to be simpler than non-translated texts, it can be hypothesized that translated texts have shorter mean sentence length in terms of characters than non-translated texts.

2. Results

First, all the character-segmented texts were processed by *WordSmith*. The overall results of the two subcorpora are shown in Figure 2.

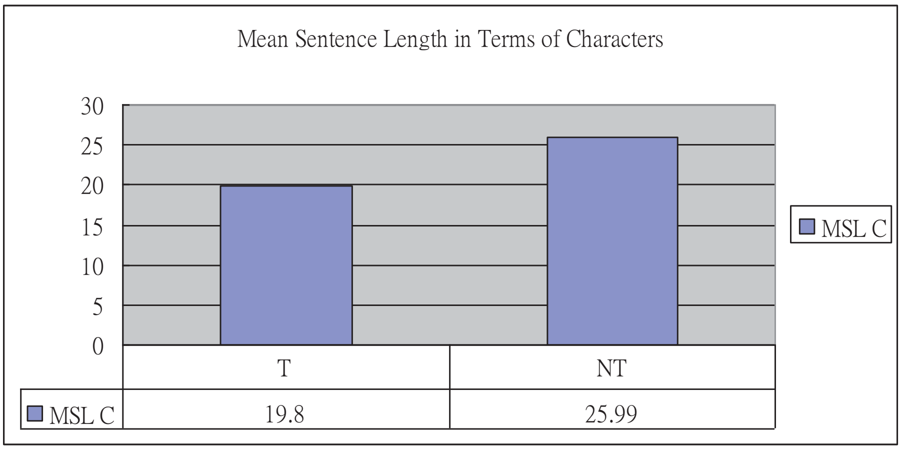

Figure 2 The results of mean sentence length in terms of characters

The mean sentence length of the translated subcorpus is 19.8 characters, 6.91 characters fewer than the mean sentence length of the non-translated subcorpus (25.99 characters). The translated subcorpus has shorter mean sentence length in terms of characters than the non-translated subcorpus, as we have predicted in the hypothesis. However, a statistical test is required to confirm that the difference between the two subcorpora is statistically significant and does not happen by chance.

3. Statistical tests

Using the same statistical *z*-test that is discussed above, we find:

	Translated (characters)	Non-translated (characters)
Sample size (n_i)	60139	47902
Mean sentence length ($\overline{X}_i$)	19.80	25.99
Standard deviation (s_i)	15.64	21.82
Z score	-52.28	

Table 2 The statistical test of mean sentence length in terms of characters

After applying the formula, $z = -52.28$. Since the significance level is still set as $\alpha = 0.05$, the values of z greater than 1.96 or less than -1.96 will be significant. Since $z = -52.28 < -1.96$, H_0 is rejected. In other words, the results of mean sentence length in terms of characters of the translated and non-translated subcorpora are significantly different in this test. Moreover, as $\overline{X}_1 = 19.8 < \overline{X}_2 = 25.99$, it can be concluded that mean sentence length in terms of characters of the translated population is significantly shorter than that of the non-translated population.

The conclusion supports the hypothesis that modern translated mystery fiction in Taiwan (with the source texts in English) tends to have shorter mean sentence length in terms of characters than the modern non-translated mystery fiction, and therefore, might be simpler for readers to comprehend, which serves as an index of simplification.

VI. Mean sentence sub-unit length in terms of words

1. Hypothesis

Since simplified texts tend to have shorter mean sentence sub-unit length in terms of words, and we might assume that translated texts tend to be simpler than non-translated texts, it can be hypothesized that translated texts have shorter mean sentence sub-unit length in terms of words than non-translated texts.

2. Results

First, all the word-segmented texts with commas, colons and semicolons replaced by full stops were processed by *WordSmith*. The overall results of the two subcorpora are shown in Figure 3.

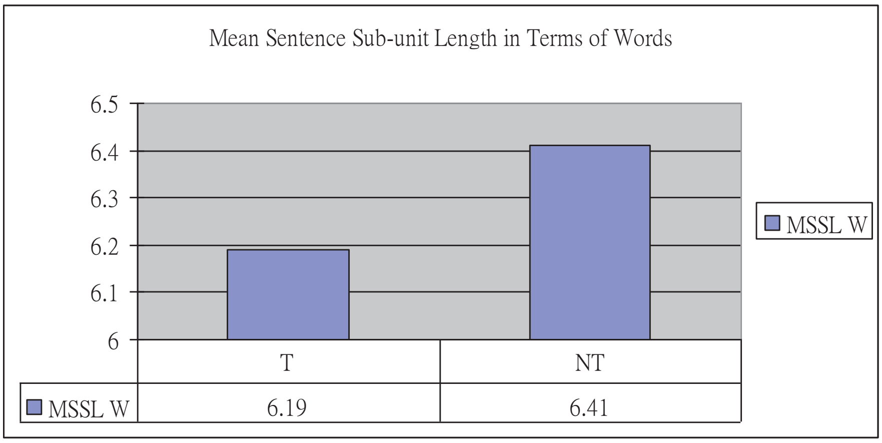

Figure 3 The results of mean sentence sub-unit length in terms of words

The mean sentence sub-unit length in terms of words of the translated subcorpus is 6.19 words, 0.22 words fewer than that of the non-translated subcorpus (6.41 words). The translated subcorpus has shorter mean sentence sub-unit length in terms of words than the non-translated subcorpus, as we have predicted in the hypothesis. However, a statistical test is required to confirm that the difference between the two subcorpora is statistically significant and does not happen by chance.

3. Statistical tests

Using the same statistical *z*-test that is discussed above, we find:

	Translated (word)	Non-translated (word)
Sample size (n_i)	126994	128465
Mean sentence sub-unit length ($\overline{X}_i$)	6.19	6.41
Standard deviation (s_i)	3.70	3.60
Z score	-15.02	

Table 3 The statistical test of mean sentence sub-unit length in terms of words

After applying the formula, $z = -15.02$. Since the significance level is still set as $\alpha = 0.05$, the values of z greater than 1.96 or less than -1.96 will be significant. Since $z = -15.02 < -1.96$, H_0 is rejected. In other words, the results of mean sentence sub-unit length in terms of words of the translated and non-translated subcorpora are significantly different in this test. Moreover, as $\overline{X}_1 = 6.19 < \overline{X}_2 = 6.41$, it can be concluded that mean sentence sub-unit length in terms of words of the translated population is significantly shorter than the mean sentence sub-unit length of the non-translated population.

The conclusion supports the hypothesis that modern translated mystery fiction in Taiwan (with the source texts in English) tends to have shorter mean sentence sub-unit length in terms of words than the modern non-translated mystery fiction, and therefore, might be simpler for readers to comprehend, which serves as an index of simplification.

VII. Mean sentence sub-unit length in terms of characters

1. Hypothesis

Since simplified texts tend to have shorter mean sentence sub-unit length in terms of characters, and we might assume that translated texts tend to be simpler than non-translated texts, it can be hypothesized that translated texts have shorter mean sentence sub-unit length in terms of characters than non-translated texts.

2. Results

First, all the character-segmented texts with commas, colons and semicolons replaced by full stops were processed by *WordSmith*. The overall results of the two subcorpora are shown in Figure 4.

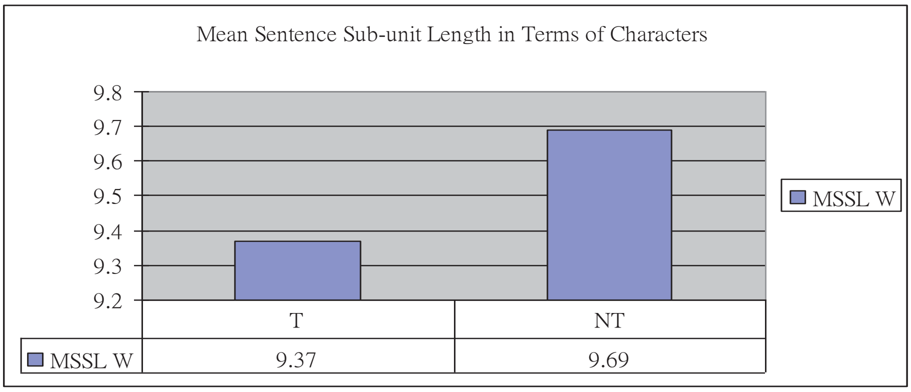

Figure 2 The results of mean sentence sub-unit length in terms of characters

The mean sentence sub-unit length in terms of characters of the translated subcorpus is 9.37 characters, 0.32 characters fewer than that of the non-translated subcorpus (9.69 characters). The translated subcorpus has shorter mean sentence sub-unit length in terms of characters than the non-translated subcorpus, as we have predicted in the hypothesis. However, a statistical test is required to confirm that the difference between the two subcorpora is statistically significant and does not happen by chance.

3. Statistical tests

Using the same statistical *z*-test that is discussed above, we find:

	Translated (characters)	Non-translated (characters)
Sample size (n_i)	127027	128518
Mean sentence sub-unit length ($\overline{X}_i$)	9.37	9.69
Standard deviation (s_i)	5.63	5.36
Z score	-14.40	

Table 4 The statistical results of mean sentence sub-unit length in terms of characters

After applying the formula, $z = -14.40$. Since the significance level is still set as $\alpha = 0.05$, the values of *z* greater than 1.96 or less than -1.96 will be significant. Since $z =$

-14.40 < -1.96, H_0 is rejected. In other words, the results of mean sentence sub-unit length in terms of characters of the translated and non-translated subcorpora are significantly different in this test. Moreover, as $\overline{X}_1$ = 9.37 < $\overline{X}_2$ = 9.69, it can be concluded that mean sentence sub-unit length in terms of characters of the translated population is significantly shorter than the mean sentence sub-unit length of the non-translated population.

The conclusion supports the hypothesis that modern translated mystery fiction in Taiwan (with the source texts in English) tends to have shorter mean sentence sub-unit length in terms of characters than the modern non-translated mystery fiction, and therefore, might be simpler for readers to comprehend, which serves as an index of simplification.

VIII. Summary

In conclusion, the modern translated mystery fiction in Taiwan (with source texts in English only) has significantly shorter mean sentence length both in terms of words and characters and significantly shorter mean sentence sub-unit length both in terms of words and characters than its non-translated counterpart. In other words, the results of the measure of mean sentence length and its additional measures all indicate that modern translated mystery fiction in Taiwan (with source texts in English only) does have shorter sentences and shorter sentence sub-units and exhibits syntactic simplification than its non-translated counterpart.

Due to the controversy over the definition of a Chinese word and the consideration of word length, the measure of mean sentence length in terms of characters has been proposed. Moreover, regarding the specific characteristics of Chinese sentences and Chinese punctuation marks, mean length of sentence sub-units, words between full stops, exclamation marks, question marks, commas, semicolons and colons, has been proposed as an additional measure. This additional measure of mean sentence sub-unit length is also measured both in terms of words and characters. The differences and the *z* scores of the measure of mean sentence length in terms of words, and its additional measures, i.e., mean sentence length in terms of characters, mean sentence sub-unit length in terms of words and mean sentence sub-unit length in terms of characters, are listed in table 5.

	Words		Characters	
	differences (T – NT)	*z* scores	differences (T – NT)	*z* scores
Mean sentence length	-4.11	-53.08	-6.91	-52.28
Mean sentence sub-unit length	-0.22	-15.02	-0.32	-14.40

Table 5 The *z* scores of mean sentence length and its additional measures

In statistics, however, once the test is chosen (the *z*-test) and the level of significance (α = 0.05) is established, every result (*z* > 1.96 or *z* < -1.96) will be either statistically significant or not. Although the degrees of statistical significance are not usually distinguished, the table of *z* scores above indicates that the translated subcorpus has significantly shorter sentence length whether mean sentence length is measured in terms of words or in terms of characters. While the controversy of the definition of words remains unsolved, with the value of *z* much smaller than -1.96, mean sentence length in terms of characters cannot only serve as an additional measure, but also as an adequate alternative measure.

Moreover, the table shows that the values of *z* of mean sentence length both in terms of words and characters are relatively much larger than 1.96 or much smaller than -1.96

compared with the values of *z* of mean sentence sub-unit length. The measure of mean sentence sub-unit length is proposed because a Chinese sentence is a much larger unit than an English sentence and commas in Chinese are often used to connect sentences, clauses, and longer phrases. This measure can further eliminate the possible influence of English punctuation marks, especially full stops, on the translated texts. It was expected that the translated texts might have shorter sentence length due to the usage of full stops in their English source texts, and mean sentence sub-unit length would further suggest that regardless of the interference from their source texts, the translated texts are still syntactically simplified by having shorter sentence sub-units length. The results show that the measures of mean sentence length and mean sentence sub-unit length both render consistent results.

According to the results and the statistical tests in this study, we can conclude that the modern translated mystery fiction published in Taiwan (with source texts in English only) tends to have shorter sentence length and shorter sentence sub-unit length than its non-translated counterpart, and is therefore simpler syntactically.

References

Academia Sinica Balanced Corpus of Modern Chinese: online. Available at: http://proj1. sinica.edu.tw/~tibe/2-words/modern-words/ index.html (accessed January 2009).

Baker, M. (1996). "Corpus-based translation studies: The challenges that lie ahead". In H. Somers (Ed.), *Terminology, LSP and translation studies in language engineering: In honour of Juan C. Sager*. Amsterdam and Philadelphia: John Benjamins, 175-186.

Baroni, M. and S. Evert. (2008). "Statistical methods for corpus exploitation". In A. Lüdeling and M. Kytö (Eds.), *Corpus linguistics: An international handbook*. Berlin and New York: Mouton de Gruyter, 777–803.

Butler, C. (1985). *Statistics in Linguistics*. Oxford and New York: Basil Blackwell.

Chang, S. (2006): online. "The new tourism—Young Japanese discover Taiwan". In *Taiwan Panorama*, March 2006: 34. Available at: http://www.taiwan-panorama. com/index.php. (accessed January 2009).

Chinese VIII. (2004). Taipei: National Institute for Compilation and Translation.

Fromkin, V. A., R. Rodman and N. M. Hyams. (2003). *An Introduction to Language*. 7th edition. Boston and Massachusetts: Heinle.

Gao, Z. (1997). *Automatic Extraction of Translation Equivalents from a Parallel Chinese-English Corpus*. Unpublished PhD thesis, University of Manchester.

Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): a resource and a methodology for the empirical study of translation*. Unpublished PhD thesis, UMIST.

Li, C. N. and S. A. Thompson. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Lin, S. (2002). *A Linguistic Study of the Use of Commas and Periods in Chinese*. Unpublished MA dissertation, National Hsinchu University of Education.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Packard, J. L. (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge: Cambridge University Press.

Scott, M. (2007): online. *WordSmith Tools. Version 5. Online manual*. Available at: http://www.lexically.net/wordsmith/index.html (accessed January 2009).

Standard of Chinese Word Segmentation for Information Processing: online. Available at: http://rocling.iis.sinica.edu.tw/CKIP/paper/ wordsegment_standard. pdf (accessed January 2009).

Tang, T. (1988). *Hanyu Cifa Jufa Lunji* (*Studies on Chinese Morphology and Syntax*). Taipei: Student Book Company.

Teng, S. (2007): online. "Philips' simplicity revolution". *Taiwan Panorama* May 2007: 22. Available at: http://www.taiwan-panorama. com/index.php (accessed January 2009).

----- (2007): online. "Taiwanese Design Takes Flight". *Taiwan Panorama* May 2007: 6. Available at: http://www.taiwan-panorama. com/index.php (accessed January 2009).

Woods, A., P. Fletcher and A. Hughes. (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Appendix 1: Texts included in the CCCM

The Translated Subcorpus

	Title	File name	Author	Translator	Date (TT)	Date (ST)	Publisher	Size: pages	Size: characters	Size: words
1	繁花將盡 *All Flowers Are Dying*	Flowers	勞倫斯・卜洛克 (Lawrence Block)	尤傳莉 (You, Chuan-li)	31/01/2005	2005	臉譜 (Faces Publishing Ltd)	415	147976	100376
2	第 12 張牌 *The Twelfth Card*	Twelfth	傑佛瑞・迪佛 (Jeffrey Deaver)	劉永毅 (Liu, Yong-yi)	21/11/2005	2005	皇冠 (Crown Culture Corporation)	536	245387	160545
3	第三隻魔 *3rd Degree*	Third	詹姆斯・派特森 (James Patterson)	鄭家瑾 (Cheng, Chia-ching)	08/03/2005	2005	宏道文化 (Elegant Books Ltd)	272	88395	57342
4	泣！死神的哀悼 *Monday Mourning*	Mourning	凱絲・萊克斯 (Kathy Reichs)	吳俊宏 (Wu, Chiun-hung)	28/11/2005	2004	皇冠 (Crown Culture Corporation)	400	153605	101443
5	噬血聖母的噤聲曲 *Hush*	Hush	安・佛萊瑟 (Anne Frasier)	胡洲賢 (Hu, Chou-hsian)	01/11/2005	2002	果藤書房 (Vine Publisher)	288	142311	94052
6	搖籃曲 *Lullaby*	Lullaby	恰克・帕拉尼克 (Chuck Palahniuk)	盧慈穎 (Lu, Tsi-ying)	09/10/2005	2002	麥田 (Rye Field Publishing Co.)	384	104583	69852
7	死亡傳喚 *The Summons*	Summons	約翰・葛里遜 (John Grisham)	宋偉航 (Song, Wei-hang)	01/10/2005	2002	遠流 (Yuan-Liou Publishing Co., Ltd)	454	137062	93739
8	人質 *Hostage*	Hostage	羅伯・克里斯 (Robert Crais)	莊綉雲 (Chuang, Xu-yun)	28/09/2005	2001	墨文堂文化 (Cheng Chung Books Co.)	480	171283	108692

The Non-Translated Subcorpus

	Title	File name	Author	Date	Publisher	Size: pages	Size: characters	Size: words
1	天人菊殺人事件 *Tian-ren-chu Murder Case*	TRJ	藍霄 (Lan, Hsiao)	15/08/2005	小知堂 (Wisdom Publisher)	240	90964	57795
2	雨夜莊謀殺案 *Murder Case at Yu-ye Village*	YYZ	林斯諺 (Lin, Si-yan)	10/12/2005	小知堂 (Wisdom Publisher)	384	123144	81235
3	超能殺人基因 *ESP Murderous Gene*	Gene	既晴 (Chi, Ching)	15/11/2005	皇冠 (Crown Culture Corporation)	272	111809	71735
4	殺人上癮 *Addicted to Killing*	Addicted	夏佩爾、烏奴奴 (Hsia, Pei-er & Wu, Nu-nu)	30/06/2005	小知堂 (Wisdom Publisher)	288	93697	62916
5	失落的印記 *Curse of the Eagle God*	Curse	伍臻祥 (Wu, Chen-hsiang)	15/09/2004	高寶 (Gobooks Publisher)	304	133984	88964
6	消失的天堂鳥 *Disappearing Bird of Paradise*	Paradise	吳國棟 (Wu, Kuo-tong)	04/07/2004	商周出版 (Business Weekly Publications)	224	80514	53675
7	偷天換日 *Deception*	Deception	黃河 (Huang, Ho)	01/03/2004	商周出版 (Business Weekly Publications)	264	89236	58798
8	皇陵天眼 *Imperial Mausoleum*	Imperial	景旭楓 (Ching, Hsiu-feng)	30/05/2005	滾石文化 (Rock Publisher)	448	252229	168359
9	無伴奏安魂曲 *Requiem without an Obbligato*	Requiem	成英姝 (Cheng, Ying-shu)	01/11/2000	時報出版 (China Times Publishing Co.)	176	62741	42011
10	第四象限 The Fourth Quadrant	Fourth	天地無限 (Tian Ti Wu Hsian)	15/02/2002	皇冠 (Crown Culture Corporation)	256	90545	58946
11	Saltimbocca，跳進嘴裡 Saltimbocca, Jump into the Mouth	Saltimbocca	張國立 (Chang, Kuo-li)	01/11/2000	時報出版 (China Times Publishing Co.)	144	46095	31543
12	疑惑與誘惑 Confusion and Temptation	Temptation	裴在美 (Pei, Tsai-mei)	01/11/2000	時報出版 (China Times Publishing Co.)	232	69823	46982