# A Computerized Stylostatistical Approach to the Disputed Authorship Problem of *The Dream of the Red Chamber*

*Bing C. Chan*

## Rationale for Stylostatistical Approach

When a writer's thoughts are written down, his manner of presentation and the idiosyncracies of his style become apparent to the reader. The reader is able to make a good guess at the authenticity of a manuscript attributed to that writer. However, intuitive judgements by different individuals on a writer's style sometimes vary to a great extent, as the meaning of "style" is differently interpreted. It is generally agreed that there is a characteristic or are characteristics peculiar to an author which differentiate him/her from other authors. In this paper, the word "style" will be defined as "constant characteristics" in a writer's way of writing. Whether these "constant characteristics" are evaluated as "poor" or "good," they may still serve well for the purpose of identification of authorship.

It is commonly accepted that some features of an author's language remain constant, but the question that needs to be raised is: How much control does an author have over his own style? Evidently, if a writer wants to control and shape details of his writing, he can intentionally alter any word, and in this sense, can be said to have complete power over his language. Yet he cannot exceed his vocabulary repertory. It is believed that a speaker or writer using his native language would normally pay little or no attention to its semantic and syntactic components. The syntactic structures are not usually consciously examined as the speaker or writer expresses his ideas. In other words, grammar is generated by an unconscious faculty.[1] For instance, a writer does not consciously determine the average length

Editor's Note: In most cases, only the Wade Giles romanization system is used in *Tamkang Review*. We have retained the *pinyin* romanization system in this article so that we may not tamper with the author's computerized coding system.

of his sentences, or the frequency of certain parts of speech used in a paragraph or a chapter. These parameters are some of the stylistic elements determined by habit which form a pattern that constitutes one aspect of a writer's characteristics. Since an author is normally unaware of the nature of these subconscious measures, there is no conscious effort to control them. Even if an author discovers his own idiosyncracies and attempts to modify his style by making alterations, he is still likely to return to his habitual mode of presentation when speed and efficiency are paramount. It is this belief that underlies the published studies of several scholars who have attempted to attribute literary works of uncertain authorship to their proper authors. Studies on word frequency in a text can be traced as early as those ancient counts by Jewish scholars of the Masoretic Bible and the word counts of the Homeric text at Alexandria.[2] T. C. Mendenhall's study in 1887 on the word length of Shakespeare's selected works is often considered a pioneer article in extant.[3] Guiraud, Kirkpatrick, Lutoslavski, Herdan, and Yule all made contributions to the study of an author's vocabulary size and word frequency distributions.[4] Of these studies, the two most recent and well-known are the work of Ellegard[5] on the *Junius* letters and of Modteller and Wallace[6] on the *Federalist* papers. It is also this belief that underlies this writer's attempt to employ a computerized, stylostatistical approach to discuss the disputed authorship of *The Dream of the Red Chamber* (*Honglou Meng* 紅樓夢 ; hereafter, *The Dream*). This study was begun with no preconceptions concerning the identity of the author of *The Dream*, and it mainly seeks to provide objective evidence to support either the dual-authorship or the single-authorship theory, as indicated by the test results. The methodology adopted will use as criteria stylistic habits of writing which are more or less natural and over which the writer would have very little or no conscious control. Attention is therefore focused on grammar, vocabulary, and phrase, rather than on other more distinctive literary qualities of a writer's style. The aim is to define what is constant in a writer's method of expression.

## Pros and Cons in Stylostatistical Approach

Several attacks have been made against the application of statistical analysis of writing style. They include: (1) style is an element that is too dependent upon intuition and the sensitivity and depth of experience of the writer; (2) statistical analysis is incapable of revealing anything that is not

intuitively obvious, or disclosing anything that is relevant to the judgment of literary worth; and (3) it is sacrilegious to approach a literary work of ineffable beauty and delicate feeling through the use of formulas.

Ironically, the criticism against statistical analysis based on the argument that style is "intuitive" and too subjective and impressionistic is precisely the point that shapes the counterattack of the stylostatisticians. Their view is that the reader decipers a text by means of his own code and takes from it meanings and styles that are sometimes very far from those the writer may have consciously or unconsciously intended. The stylo-statisticians maintain that a literary forgery can occur when a capable writer deliberately imitates another writer's style, especially in fraudulent short texts. The strong point made by stylostatisticians is that the individual characteristics such as the frequency distribution of different nouns, the mean length and variability of sentences, etc., which all writers possess but of which they are mostly unconscious, are not too easy to imitate; but word length, word order, percentages of parts of speech, and other involuntary characteristics would perhaps soon reveal a pattern of identity for a specific writer. Yule strengthened this position with the following statement:

> Surely the color and flavor of a text, if I may be permitted to mix my metaphor, are determined not by the exceptional words, unless these words taken together form a large class, but in the main by common words used by the author, the words used by him over and over again . . . words used once, words used twice, words used thrice and so on.[7]

Yet one should bear in mind that in very large samples where the vocabulary available to two authors of equivalent educational background is being examined, the richness of the vocabulary of each author becomes almost identical.

The quantitative-objective approach of statistical analysis need not replace qualitative-subjective appreciation. The two can co-exist, with the former helping to objectify and clarify the latter, and to act as a tool to surmount the difficulty of appraising the mass of literary phenomena.

Although statistical methods have been shown by extensive scientific applications to be reliable and are ideally suited for stylistic analysis, one cannot say for certain that all methods are applicable to or suitable for disputed authorship problems. It should be clear that in the methods of

drawing inference about authorship of disputed works, the results take the form either that it is "more likely," or "less likely," that the same hand was responsible for the two sets of writings, or that the information gives no evidence at that point. C. B. William explained very well that:

> If two sets of writings are shown to have very different statistical characteristics, it can be concluded with high justification that they are not products of the same mind. If, on the other hand, they are found to be very similar in one or two criteria, it would not be equally justifiable to say that the same mind had produced both.[8]

Therefore, when drawing a conclusion based on statistical analysis, one has to realize, as William emphasized, that:

> The conclusion is never absolutely certain, even if all the available evidence has been studied . . . if all the available evidence has been studied, and the conclusion is not sufficiently definite, one must either accept the uncertainty, or find a new angle of approach – a new criterion.[9]

The objective in this study is merely to compare statistically the different parts of the text of *The Dream* to show vocabulary differences or similarities in the compared sections. This study is a scientific experiment. The present writer is not able to predict any outcome before the statistical tests are applied. If the hypothesis that *The Dream* was written by two different authors should issue in either a positive or a negative result, that should not be considered a complete success or failure. The statistical tests to be used here were developed previously by statisticians and linguists. These tests have been applied in languages other than English.[10]

## Divisions and Sampling

The first step was to divide *The Dream* into homogeneous sections. The first eighty chapters of the *Gengchen* Version 庚辰本[11] were divided into two parts, i.e., Chapters 1-40 (Text A) and Chapters 41-80 (Text B). Only Chapters 81-120 (Text C) of the 120-chapter *Chengyi* Version 程乙本[12] were used for comparison. Each full page of either Text A or Text B consists of approximately 30 characters. Two hundred

sixty-three pages from Text A and Text B, respectively, were selected. Each text then contributed approximately 80,000 characters. Each full page of Text C contains 240 characters. Three hundred thirty pages were chosen and thus contributed approximately 80,000 characters. As the data is enormous and is to be normalized later for statistical inference, a minor difference in the total number of characters among the three texts is not significant. All pages were chosen by adopting the numbers used in the *Table of Random Digits* by the Rand Corporation.

Except for the beginning page of each chapter and some of the last pages of certain chapters, each page of the first eighty chapters contains 10 columns (*hang* 行), i.e., a full page. Any page that had less than 10 *hang* was omitted in the sampling process. However, Chapter 19 was excluded for the reason that it contained many pages of poetry and lyrics, and was not qualified for a sampling of a "full page." If it happened that the random number picked a "non-full" page, the next or the following full page was used. Chapters 64 and 67 were also excluded from this study as these two chapters were adapted from versions other than the *Gengchen* itself.[13]

At this point, it should be noted that out of approximately 700,000 characters in the whole novel of *The Dream*, the sample of the three texts A, B, C, contains approximately 240,000 characters, which is a little over 34 percent, an amount of material considered more than sufficient in any statistical analysis.

Each selected page then was pre-edited by marking nouns in blue, adverbs in green, adjectives in red, stative verbs in orange, and particles in different color symbols, a procedure which permitted the subsequent recovery of all instances of each class of the content vocabulary during the checking and rechecking process. The variables were then copied out and grouped categorically for coding in the *pinyin* 拼音 romanization system. It was expected that some 50,000 index cards would be needed for recording the data before any computer processing could be done.

The information was then key-punched in free format for four types of the vocabulary (the particles were not as numerous and could be done manually). A master word list of each category was produced and manually checked.

## Computerized Data and Statistical Analysis of the Vocabulary

All preliminary processing was carried out at the computer center of the

State University of New York at Stony Brook, on a Univac 1100 computer to obtain relative frequency counts and alphabetical listings, and for test processing. Computer programs, written in Fortran language, were used to instruct the computer to transform the data to formats suitable for the selected statistical tests on the massive sample. The sample size is far beyond that is needed in routine statistical analysis, and without the aid of a computer, the time needed to perform all the mechanical operations of counting, alphabetizing, and computation would have been prohibitive, if not impossible, for an individual to complete. After numerous trials and errors, the final computer output included twelve master word distribution tables and an extensive combined word frequency distribution talbe of Texts A, B, C of *The Dream* and Text D, *The Gallant Ones* (*Ernu Yingxiong Zhuan* 兒女英雄傳),[14] which was added for the purpose of testing the validity of the statistical tests used.

There are two very good reasons to choose *The Gallant Ones* for comparison. First, the novel is written in Peking dialect. *The Dream* also is essentially written in Peking dialect. This linguistic feature makes the two novels uniquely compatible for comparison, since no other novel shares this characteristic. Second, and more importantly, *The Gallant Ones* is written about one hundred years later than *The Dream*. During these one hundred years, evolution of the language must have taken place. These changes in language, no doubt, should set apart explicitly the difference in diction in both *The Gallant Ones* and *The Dream*, which were written by two different authors. If the statistical tests in this research are not able to indicate the linguistic differences of these two authors, the tests will not be sensitive enough and thus are invalid for use in a study of this kind.

## Classification of the Vocabulary

The principles used to classify the vocabulary of the novels in this study must be described. It is the belief of this researcher that determining word-functions stems from analysis of the specific context where a particular word is used. It seems unnecessary to go into particulars here, as linguistic research has proved with sufficient certainty that each language possesses its own unique grammar. It is also this researcher's conclusion that the grammatical description of words in Chinese must obviously vary from the procedures adopted in similar studies made for other languages. Categories will have to be reduced to a minimum. Therefore the words studied in the

text of *The Dream* were chosen from these five grammatical classes only: nouns, adverbs, adjectives, stative verbs, and particles.

The professional literature in Chinese in the field of Chinese grammar is less ample than one could wish, and was able to provide little guidance. This study therefore had to devise its own ways and means. All classification of grammatical elements in this study of *The Dream* was done by the present writer based on his best understanding of the Chinese language, with the aid of reference works by Wang Li 王力[15] and Chao Yuan-ren 趙元任.[16] As the classification was carried out by the present writer alone, it is understood that there is consistency throughout the whole process. Scratched notes were written to remind the writer of the easily confused parts. It would be assuming too much to state that no error has occurred, as human errors do exist without one's knowledge. But repeated checking has eliminated possible errors to a degree that even if some were found, they would not be of sufficient importance to twist the experimental result.

A general guideline for each selected grammatical category is given here for the clarification of concepts.

*Noun*: A noun is a word used for naming some person or thing. Only common nouns are considered in the present study. A common noun denotes no one person or thing in particular, but is common to any and every person or thing, such as "book," "poetry," "tree," and "flower," but not *Sishu* 四書 (The Four Books), *taoshu* 桃樹 (the peach tree), or *juhua* 菊花 (chrysanthemum). Certain nouns, even if they are common nouns, due to their contextual necessities, are discarded from this category.

Words placed together without connecting particles that blend into a compound belonging to their own or another part of speech are then treated as single words. In the term *xiangshui* 香水 (perfume), the two characters, while they maintain their relationship to each other as adjective and substantive, constitute in the general syntax of the sentence a single noun. Their individual sense and mutual relation are not destroyed, but in common use are entirely forgotton.

When classifying the nouns, the following guidelines are observed consistently within this study:

1. In the V-O (verb-object) construction, the "O" is counted as a separate noun, e.g., chi*fan* 吃飯 (to eat a meal); zou*lu* 走路 (to walk a street); chuan *yifu* 穿衣服 (to wear clothes).
2. Nouns like *haizi* 孩子 (child), *hair* 孩兒 (child), *xiao haizi* 小孩子 (child) and *haizi men* 孩子們 (children) are counted as distinct nouns

due to the computer's inability to distinguish their meanings when processing the data.

3. Nouns such as *dong jiaomen* 東角門 (the eastern side door) and *dong yuan* 東院 (the eastern court) are counted without the directional affix *dong* 東 (eastern).
4. Phrases like *nie shou nie jiao* 躡手躡腳 (stealthily) and *shengri zhi li* 生日之禮 (birthday present) are considered as having two nouns each, namely, *shou, jiao*, and *shengri, li*.
5. Agent nouns such as *yao fande* 要飯的 (beggar) and *da shifande* 打十幡的 (a moaner in a funeral) are not counted as nouns but the elements they contain are counted separately as nouns such as *fan* and *shifan*.
6. The following elements are not considered in this study. They are:
   a) Time words such as: *jinr* 今兒 (now), *muqian* 目前 (the present moment), *zheige yue* 這個月 (this month), *chun* 春 (spring), *xiawu* 下午 (afternoon), and so forth.
   b) Place words such as: *zher* 這兒 (here), *youbiar* 右邊兒 (the right), *xifang* 西方 (the west), and so forth.
   c) Proper nouns such as names of persons, places, plants, flowers, medicines, rivers, buildings, mountains and the like.
   d) Titles or designations of persons such as: *Wang taitai* 王太太 (Mrs. Wang), *Bao guniang* 寶姑娘 (Miss Bao), etc.
   e) Kinship terms such as: *fu* 父 (father), *yima* 姨媽 (aunt), *wai shengnur* 外甥女兒 (female cousin), and so forth.
   f) Master-servant designations such as: *laoye* 老爺 (master), *yatou* 丫頭 (maid), *zhuzi* 主子 (master), and so forth.
   g) Words like *ren* 人 (people, person) and *jia* 家 (family, home), are not counted because they are very contextual in *The Dream*.

*Adverb*: An adverb is a word used to qualify any part of speech except a noun or pronoun. Only simple adverbs are considered here. A simple adverb modifies a word or a group of words.

Adverbs qualifying verbs are derived from adjectives by repeating them with a suffix. The words *de* 地 ,的 , *zhe* 着 , *er* 兒 and *li* 裏 are the most common endings to these groups. A few examples are given: *xixide* 細細的 (slightly), *chanchangde* 常常的 (often), *siside* 私私的 (privately).

Simple and dissyllabic adjectives take the same endings without repetition, as in *anxiali* 暗下裏 (secretly) and *fengkwar* 鋒快兒 (sharply).

An adjective, repeated or not, before a verb becomes an adverb, as in

*mingming shuo* 明明說 (spoke plainly) and *xiaoxiao shuo* 悄悄說 (spoke quietly).

There are also many single, simple adjectives used as adverbs, which enter into combination with simple verbs, as in *bai fei gongfu* 白費功夫 (fruitlessly waste time) and *man qu* 慢去 (slowly going).

However, adverbs of place, direction, arrangement and time, such as *zher* 這兒 (here), *cishi* 此時 (right now), *ruhe* 如何 (how), etc., are not collected in the sample. Adverbs of manner which include a verb or noun form in the structure, such as *huan tian xi di de* 歡天喜地的 (happily), *xiao xixide* 笑嘻嘻的 (smilingly) and *tang yan mo lei de* 淌眼抹淚的 (tearfully), are not considered in the corpus, as all verbs are excluded in this study while the noun forms are marked consistently as nouns throughout the whole classification in the three texts of *The Dream*. Thus the structure *huan tian xi di de* would be marked as consisting of two nouns, i.e., *tian* 天 (sky) and *di* 地 (earth).

*Adjective*: An adjective is a word used to qualify a noun. Only the descriptive adjective in its epithet use is considered. A descriptive adjective restricts the application of a noun to such persons or things that possess the quality or state denoted by the adjective. An adjective is used as an epithet when it qualifies its noun directly, as in *hei ma* 黑馬 (black horse) and *biaozhi de nuren* 標緻的女人 (pretty woman).

When an adjective is in the "predicative use," it is not considered as an adjective in this study but is counted as a stative verb, as in *Tian heile* 天黑了 (The sky is dark) and *Neige nuren hen biaozhi* 那個女人很標緻 (That woman is very pretty).

*Stative verb*: A stative verb is different from a regular verb in that:

1. A stative verb can be preceded by a word meaning "very," such as *hen* 很, *shifen* 十分, and *feichang* 非常.
2. It does not take any object.
3. It describes a state of being and thus in meaning resembles an adjective in the predicative use in an English sentence, as in *Ta hao* 他好 (He is fine) and *Ta ren you biaozhi, yantan you shuangli* 她人又標緻, 言談又爽利 (She is pretty and her words are frank).

The word distribution tables of the four types of vocabulary from the four texts are condensed in Tables 1, 2, 3, and 4.[17]

There are two ways, according to Herdan, to ascertain objectively a resemblance or difference between frequency distributions.[18] He pointed out that the distributions may be similar because they are only chance

Table 1

Adjective – Condensed Relative Frequency Table

| Frequency | Text | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 ..... | 194 | 141 | 134 | 183 |
| 2 ..... | 34 | 30 | 23 | 41 |
| 3 ..... | 16 | 19 | 5 | 23 |
| 4 ..... | 8 | 11 | 7 | 7 |
| 5 ..... | 10 | 3 | 4 | 4 |
| 6 ..... | 4 | 3 | 2 | 4 |
| 7 ..... | 1 | 4 | 1 | 3 |
| 8 ..... | 2 | 0 | 1 | 2 |
| 9 ..... | 0 | 0 | 1 | 1 |
| 10 ..... | 0 | 1 | 0 | 2 |
| 11-20 ..... | 5 | 5 | 3 | 3 |
| 21-30 ..... | 1 | 2 | 1 | 1 |
| 31-40 ..... | 1 | 0 | 0 | 1 |
| 41-50 ..... | 0 | 1 | 0 | 0 |
| 50 & up .... | 3 | 3 | 3 | 2 |
| Total no. of words ..... | 279 | 223 | 185 | 277 |
| Total no. of occurrence.. | 864 | 874 | 528 | 818 |
| Mean ..... | 3.097 | 3.919 | 2.854 | 2.953 |
| Standard dev. | 10.518 | 14.320 | 8.030 | 11.553 |

Total number of different adjectives yielded by the four texts A, B, C, and D: 641

Table 2

Noun – Condensed Relative Frequency Table

| Frequency | Text | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 ..... | 1267 | 1157 | 1056 | 1537 |
| 2 ..... | 329 | 330 | 263 | 416 |
| 3 ..... | 144 | 138 | 95 | 166 |
| 4 ..... | 78 | 65 | 65 | 85 |
| 5 ..... | 56 | 37 | 38 | 45 |
| 6 ..... | 28 | 28 | 26 | 32 |
| 7 ..... | 29 | 21 | 19 | 25 |
| 8 ..... | 14 | 12 | 12 | 18 |
| 9 ..... | 21 | 7 | 10 | 20 |
| 10 ..... | 10 | 14 | 14 | 16 |
| 11-20 ..... | 46 | 45 | 51 | 55 |
| 21-30 ..... | 18 | 18 | 15 | 24 |
| 31-40 ..... | 8 | 4 | 9 | 7 |
| 41-50 ..... | 4 | 10 | 5 | 2 |
| 50 & up .... | 12 | 9 | 9 | 14 |
| Total no. of words ..... | 2073 | 1895 | 1687 | 2462 |
| Total no. of occurrence.. | 6482 | 5859 | 5686 | 7501 |
| Mean ..... | 3.127 | 3.092 | 3.370 | 3.047 |
| Standard dev... | 9.329 | 9.361 | 11.102 | 10.284 |

Total number of different nouns yielded by the four texts A, B, C, and D: 5376

Table 3

Adverb – Condensed Relative Frequency Table

| Frequency | Text | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 ..... | 262 | 188 | 198 | 227 |
| 2 ..... | 62 | 51 | 77 | 63 |
| 3 ..... | 35 | 14 | 45 | 37 |
| 4 ..... | 17 | 35 | 14 | 17 |
| 5 ..... | 18 | 18 | 16 | 16 |
| 6 ..... | 16 | 7 | 18 | 12 |
| 7 ..... | 7 | 10 | 14 | 4 |
| 8 ..... | 10 | 9 | 7 | 9 |
| 9 ..... | 11 | 9 | 6 | 6 |
| 10 ..... | 5 | 3 | 6 | 4 |
| 11-20 ..... | 29 | 32 | 34 | 19 |
| 21-30 ..... | 17 | 13 | 13 | 6 |
| 31-40 ..... | 8 | 12 | 11 | 6 |
| 41-50 ..... | 8 | 7 | 5 | 5 |
| 50 & up .... | 23 | 25 | 16 | 14 |
| Total no. of words ..... | 528 | 432 | 481 | 445 |
| Total no. of occurrence.. | 6763 | 6888 | 5763 | 5534 |
| Mean ..... | 12.809 | 15.944 | 11.981 | 12.436 |
| Standard dev .. | 50.751 | 59.331 | 45.948 | 52.490 |

Total number of different adverbs yielded by the four texts A, B, C, and D: 1004

Table 4

Stative Verb – Condensed Relative Frequency Table

| Frequency | Text | | | |
|---|---|---|---|---|
| | A | B | C | D |
| 1 ..... | 302 | 285 | 225 | 172 |
| 2 ..... | 91 | 67 | 55 | 46 |
| 3 ..... | 31 | 25 | 29 | 18 |
| 4 ..... | 14 | 20 | 16 | 9 |
| 5 ..... | 9 | 15 | 12 | 9 |
| 6 ..... | 12 | 7 | 5 | 4 |
| 7 ..... | 8 | 9 | 7 | 6 |
| 8 ..... | 4 | 5 | 3 | 3 |
| 9 ..... | 5 | 1 | 4 | 3 |
| 10 ..... | 4 | 4 | 3 | 1 |
| 11-20 ..... | 6 | 12 | 8 | 3 |
| 21-30 ..... | 3 | 3 | 3 | 0 |
| 31-40 ..... | 1 | 0 | 2 | 0 |
| 41-50 ..... | 0 | 1 | 0 | 1 |
| 50 & up .... | 1 | 1 | 1 | 0 |
| Total no. of words ..... | 491 | 455 | 373 | 275 |
| Total no. of occurrence.. | 1257 | 1305 | 1162 | 606 |
| Mean ..... | 2.560 | 2.868 | 3.115 | 2.204 |
| Standard dev .. | 7.311 | 9.035 | 10.518 | 10.104 |

Total number of different stative verbs yielded by four texts A, B, C, and D: 1049

variations of one and the same basic distribution, or they may be similar because they have something in common. The former possibility suggests a significance test such as the "chi-square" test or "t-test," which reveals whether the difference between the two distributions are such that they can be attributed to chance only. The latter possibility suggests a correlation test, which reveals whether the two distributions are related to some extent, or discloses the amount of variation that is common to both.[19]

As the variables in this research do not have a normal distribution, it is not appropriate to use statistical tests based on normal distribution. For this reason, the chi-square test and the t-test for testing the significance of difference were not performed. Statistical measurements of association which indicate degrees of correlation are more appropriate than tests of significant differences.

Guided by the principles of Herdan, the correlations between the four texts through frequency of use of words were computed with the following formula. The data information and test results are recorded in the following tables; only the calculation of the correlation coefficient of adverb is cited as an illustration.[20]

$$r_{xy} = \frac{\Sigma x_i y_i - N\bar{x}\bar{y}}{N\sigma_x\sigma_y}$$

Table 5

Adverb – Correlation Coefficient r

| Text | Mean | Std. de. | $\Sigma XY$[21] | r |
|---|---|---|---|---|
| A | 12.809 | 50.751 | AB=1514307 | ab=0.4331 |
| B | 15.944 | 59.331 | AC=1218173 | ac=0.4545 |
| C | 11.981 | 45.948 | AD=1313869 | ad=0.4314 |
| D | 12.436 | 52.490 | BD=1295531 | bd=0.3507 |
| | | | BC=1265324 | bc=0.3992 |

N = 1004

The above results were obtained by using the formula to calculate the correlation coefficient between the two texts. Thus,

$$r_{ab} = \frac{1514307 - 1004 \times 12.809 \times 15.944}{1004 \times 50.751 \times 59.331}$$

$$= 0.43308$$

The test results recorded in the last column of Table 5 indicate that for the adverb distribution, the correlation coefficient for AC (0.4545) is stronger than those for AB (0.4331) and/or AD (0.4314), and that the correlation coefficient for BC (0.3922) is stronger than that for BD (0.3507). Text A and Text B have a closer correlation to Text C than to Text D. Total figures for the correlation test results of adjectives, nouns, and stative verbs are condensed in Table 6.

Table 6

Correlation Between Texts Through Frequency Use of Words

| Vocab / Pair of attributes | Adjective | Noun | Stative Verb |
|---|---|---|---|
| AB | 0.3208 | 0.3127 | 0.3530 |
| AC | 0.3322 | 0.2760 | 0.3048 |
| AD | 0.2005 | 0.2512 | 0.0340 |
| BC | 0.2583 | 0.2329 | 0.3084 |
| BD | 0.2119 | 0.1939 | 0.0381 |

The data shows clearly that the values of the correlation coefficient of the overall vocabulary of the four categories of the Texts A, B, C, and D, yield to the same tendency, i.e., there is between Text C and Text A or B a closer and stronger correlation than that between Text D and Text A or B.

For describing the strength of association between two variables, a measure called the "phi coefficient" is available.[22] It is directly related to Karl Pearson's chi-square statistic but while chi-square has the undesirable property of increasing with increases in sample size, phi ($\phi$) does not have that property and is therefore suggested as a statistic to employ in measuring meaningful differences or correlations.[23] To use phi coefficient, it is necessary to calculate the vocabulary overlaps between authors. Data

collected from *The Dream* and *The Gallant Ones* is arranged in the following Tables 7, 8, 9, and 10 for further illustration.

Table 7

Adjective – Vocabulary Overlaps Between Texts (Authors)

| Text | Size of vocab. | Overlaps | N |
|---|---|---|---|
| A | 279 | AB: 99 | 641 |
| B | 223 | AC: 74 | 641 |
| C | 185 | AD: 91 | 641 |
| D | 277 | BC: 69 | 641 |
| | | BD: 79 | 641 |

Table 8

Adverb – Vocabulary Overlaps Between Texts (Authors)

| Text | Size of vocab. | Overlaps | N |
|---|---|---|---|
| A | 529 | AB: 287 | 1004 |
| B | 432 | AC: 276 | 1004 |
| C | 481 | AD: 224 | 1004 |
| D | 445 | BC: 235 | 1004 |
| | | BD: 183 | 1004 |

Table 9

Noun – Vocabulary Overlaps Between Texts (Authors)

| Text | Size of vocab. | Overlaps | N |
|---|---|---|---|
| A | 2073 | AB: 778 | 5376 |
| B | 1895 | AC: 677 | 5376 |
| C | 1687 | AD: 716 | 5376 |
| D | 2462 | BC: 612 | 5376 |
| | | BD: 688 | 5376 |

Table 10

Stative Verb – Vocabulary Overlaps Between Texts (Authors)

| Text | Size of vocab. | Overlaps | N |
|---|---|---|---|
| A | 491 | AB: 183 | 1049 |
| B | 455 | AC: 149 | 1049 |
| C | 373 | AD: 144 | 1049 |
| D | 275 | BC: 126 | 1049 |
| | | BD: 125 | 1049 |

The numerical data in Tables 7, 8, 9, and 10 has been used to form five Association Tables for each vocabulary category. This totals twenty Association Tables for the four types of vocabulary. To illustrate this point, one Association Table for the adjectives between Text A and Text B is given below in Table 11.

Table 11

Association Table for Adjectives Between Texts A and B

| Text | B | Non-B | Totals |
|---|---|---|---|
| A | 99 | 180 | 279 |
| Non-A | 124 | 238 | 362 |
| Totals | 223 | 418 | 641 |

The general form of the table is:

Table 12

Association Table

| | | X | | |
|---|---|---|---|---|
| | | 1 | 0 | Totals |
| Y | 1 | $f_1$ | $f_2$ | $N_1$ |
| | 0 | $f_3$ | $f_4$ | $N_0$ |
| Totals | | $n_1$ | $n_0$ | N |

and the formula for the phi coefficient is:[24]

$$\phi = \frac{f_1 f_4 - f_2 f_3}{\sqrt{n_1 n_0 N_1 N_0}} \quad \text{with standard error} \quad \frac{1 - \phi^2}{\sqrt{N}}$$

Thus,

$$\phi_{ab} = \frac{99 \times 238 - 124 \times 180}{\sqrt{223 \times 418 \times 362 \times 279}} = 0.0128$$

A positive phi means that within the given sample there is a positive association between the presence of a word in the vocabulary of Text A and its presence in the vocabulary of Text B. A negative phi means that within the given sample there is a negative association between the presence of a word in the vocabulary of Text A and its presence in the vocabulary of Text B. In other words, as the frequency of a word increas in one text, its corresponding frequency in the other text decreases.

The phi coefficients which measure the correlation between authors with regard to their vocabulary overlaps are summarized in the following Tables 13, 14, 15, and 16.

Table 13

Adjective – Phi Coefficient

| Pair of attributes | Value of phi coefficient |
|---|---|
| AB | +0.0128 |
| AC | –0.0451 |
| AD | –0.1878 |
| BC | +0.0333 |
| BD | –0.1147 |

AB and BC both have positive associations, while AD and BD both have negative associations. Although here AC also has a negative association, its negative value (–0.0451) is smaller than that of AD (–0.1878) and the latter differs from zero by nearly five standard errors, thus admitting a significant negative association.

Table 14
Adverb – Phi Coefficient

| Pair of attributes | Value of phi coefficient |
|---|---|
| AB | +0.2300 |
| AC | +0.0883 |
| AD | –0.0410 |
| BC | +0.1126 |
| BD | –0.0343 |

Again, AB, BC, and AC have a positive association, while AD and BD both have a negative association.

Table 15
Noun – Phi Coefficient

| Pair of attributes | Value of phi coefficient |
|---|---|
| AB | +0.0255 |
| AC | +0.0142 |
| AD | –0.1254 |
| BC | +0.0093 |
| BD | –0.0967 |

Positive associations in pairs AB, AC and BC and negative associations in pairs AD and BD are again indicated.

Table 16
Stative Verb – Phi Coefficient

| Pair of attributes | Value of phi coefficient |
|---|---|
| AB | –0.1207 |
| AC | –0.1219 |
| AD | –0.0159 |
| BC | –0.0549 |
| BD | +0.0674 |

The result of the phi coefficient for stative verbs shows four out of five are negative. It offers no help for differentiating authorship in view of the phi values found in the other vocabulary category.

Another test, which Yule calls the "association coefficient Q," is a little different from the phi coefficient.[25] His Schematic Association Table is copied in Table 17 for illustration.

Table 17

Yule's Schematic Association Table

| Attribute | Attribute | | Totals |
|---|---|---|---|
| | B | $\beta$ | |
| A | AB | A$\beta$ | A |
| $\alpha$ | $\alpha\beta$ | $\alpha$B | $\alpha$ |
| Totals | B | $\beta$ | N |

The formula for the association coefficient Q is:[26]

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Obtaining data from Tables 7, 8, 9, and 10, values of the association coefficient Q are computed in Table 18.

Table 18

Value of Association Coefficient Q

| Pair of attributes | Adj. | Noun | Adv. | Stat. vb. |
|---|---|---|---|---|
| AB | +0.0271 | +0.0808 | +0.4593 | –0.2459 |
| AC | –0.1017 | +0.0481 | +0.1793 | –0.2529 |
| AD | –0.3717 | –0.3601 | –0.0845 | –0.0368 |
| BC | +0.0768 | +0.0327 | +0.2251 | –0.1169 |
| BD | –0.2426 | –0.2916 | –0.0698 | +0.0941 |

As summarized in Table 19, each pair of attributes has been given twelve correlation coefficient tests. Within these twelve tests, AB has 10 positive correlations and 2 negative correlations; AC has 8 positive correlations and 4 negative correlations. AD has 4 smaller positive correlations than AC and AB, but has 8 negative correlations, some of which are considerably high. BC has 10 positive correlations and 2 negative correlations; BD has 6 positive correlations and 6 negative correlations.

Table 19

Association of Attributes

| Pair of attributes | No. of positive correlations | No. of negative correlations | Total test |
|---|---|---|---|
| AB | 10 | 2 | 12 |
| AC | 8 | 4 | 12 |
| AD | 4 | 8 | 12 |
| BC | 10 | 2 | 12 |
| BD | 6 | 6 | 12 |

These associations are somewhat striking, bringing out very clearly the likeness/closeness between the vocabularies of Texts A, B, and C (pairs AB, AC, BC) of *The Dream*, and the unlikeness, or comparative unlikeness, of the pairs AD and BD. From the test results for vocabulary association between Texts D and A or Texts D and B, one may conclude at the very least that it appears highly unlikely that Text D was written by the author of Texts A and B, but Text C cannot be excluded. It is therefore suggested here that there is, on the whole, a comparatively close resemblance between the writing of the acknowledged author of *The Dream* (the first eighty chapters) and the last forty chapters, but a much lower degree of resemblance between the first eighty chapters of *The Dream* and *The Gallant Ones*.

One may ask, could this not be attributed to a careful writer of the last forty chapters, who consciously imitated the style of the first eighty chapters? The possibility exists. However, if there were such a good forger, he would not only have to have had a parallel education, but he would also have to have been equally imaginative and equally experienced in social life, to say nothing of having to have spent a tremendous amount of time to make all the frequency counts of the grammatical forms used by the author of the

first eighty chapters. Then he would have to have distributed an equal proportion of these forms in each paragraph, each page, and each chapter in order to reach resemblance in vocabulary distribution.

Besides the four types of vocabulary, the non-contextual particles, or *xuzi* 虛字, of the four texts are also selected for comparison. The selection of the particles is based on the principle that these particles are not required in a sentence either for their grammatical function or semantic necessity. For instance, the *de* in the following sentences:

> Zhei shi haode xuesheng sushe 這是好的學生宿舍
> (These are good student dormitories); and
> Zhei shi hao xueshengde sushe 這是好學生的宿舍
> (These are good student's dormitories).

could not be used as samples, as these two *des* play very important roles in the meaning of the two sentences. If the *de* in the above sentences were taken away, the two sentences, which had different meanings before, would then convey the same meaning. Therefore the responsibility of *de* cannot be neglected in the syntax of contexts like these, as it is not syntactically and grammatically free. However, the first *de* in the sentence 這本新的書是我的 (This new book is mine) could be sampled, as without it the meaning and syntax of this sentence would not be affected. The second *de* in the above sentence would not be sampled, as one can see that without it the sentence would make no sense because it would indicate that "this new book is me" or "I am this new book." The following are some examples of the *de* which can be sampled:

> 劉姥姥——的領會 (*Gengchen*, Chap. 41)
> 像到了天宮裡的一樣 (*Gengchen*, Chap. 41)
> 你一個月十兩銀子的月錢，比我們多兩倍銀子(*Gengchen*, Chap. 43)
> 老太太離了鴛鴦，飯也吃不下去的 (*Gengchen*, Chap. 46)

In all these sentences, each *de* can be omitted without affecting the grammatical structure and semantic nuances of the sentences that contain it. The frequency distribution of the these particles is recorded in Table 20.

A comparison of these particles shows that within the given samples there are twenty-four distinct non-contextual particles yielded by the four Texts A, B, C, and D. There are fifteen distinct non-contextual particles in the sample of Text A, eleven in Text B, fourteen in Text C, and twenty-two

Table 20

Frequency Distribution of Particles

| Particles | Text A | Text B | Text C | Text D |
|---|---|---|---|---|
| de 的 | 656 | 379 | 506 | 192 |
| le 了 | 432 | 265 | 332 | 214 |
| yi 矣 | 10 | 1 | 0 | 3 |
| eryi 而已 | 12 | 5 | 1 | 0 |
| bale 罷了 | 9 | 6 | 7 | 3 |
| er 耳 | 2 | 2 | 1 | 0 |
| ye 也 | 9 | 1 | 0 | 22 |
| ne 呢 | 228 | 116 | 242 | 180 |
| zai 哉 | 4 | 0 | 0 | 1 |
| zi 子 | 83 | 151 | 54 | 78 |
| r 兒 | 236 | 296 | 393 | 691 |
| ba 吧 | 90 | 79 | 125 | 196 |
| a 啊 | 2 | 0 | 16 | 18 |
| ya 呀 | 1 | 0 | 12 | 54 |
| lie 咧 | 2 | 0 | 4 | 26 |
| na 哪 | 0 | 0 | 2 | 15 |
| balie 罷咧 | 0 | 0 | 6 | 1 |
| le(l) 哩 | 0 | 0 | 0 | 3 |
| po 叵波 | 0 | 0 | 0 | 4 |
| wa 哇 | 0 | 0 | 0 | 9 |
| he 呵 | 0 | 0 | 0 | 1 |
| hu 乎 | 0 | 0 | 0 | 1 |
| yue 喲 | 0 | 0 | 0 | 1 |
| wai 喂 | 0 | 0 | 0 | 1 |

in Text D. This preliminary comparison brings out a wider range of particles used in Text D than in the other three texts. If one arranges the figures in a row by descending frequency in the order D A C B, this shows immediately that C is placed between A and B, whereas Text D is outside of the A and B range. Using the frequency of nine and above as a cut-off point, there are ten non-contextual particles in Text A, six in Text B, eight in Text C, and twelve in Text D. Again, the relative frequency for Text C is in between A and B whereas the relative frequency for Text D is outside of the A and B range. Table 20 shows that there are in Text D at least nine non-contextual particles that did not appear in Text A and B at all, but there are in Text C only two non-contextual particles (*na* 哪 and *balie* 罷咧 ) that have zero frequency in Texts A and B.

The main reason that the distribution of the non-contextual particles in Text D is so different from that in the other three texts is that it is an undisputed fact that Text D was written by a different author about one hundred years later. In addition, the evolution or development of the language during that one hundred years would also play an important role in contributing to the difference.

One must be aware of the fact that since these non-contextual particles are grammatically and contextually free, the insertion of such particles in sentences is basically due to the author's idiosyncrasy. Hence one would assume that if works of approximately the same length on the same subject are by the same author, the relative frequency distribution of such particles in these works would resemble each other. However, when looking at the difference of the relative frequency distributions one cannot help wondering what causes the difference. A possible explanation is that even when grammatically and contextually free in a sentence, some of the particles are still tied by the expressive mode or manner of a narrator or a speaker, particularly in fiction. In other words, it is not what was said, but how it was said, that makes the difference. For instance, suppose a police officer returned home and found a stranger in his bedroom opening a dresser. If the policeman asks, "What are you doing?" the question accuses the intruder of wrongdoing, thus arousing a sense of guilt in the intruder. If the police officer were to ask his wife the same question when she was opening a dresser, under normal circumstance, he would not be accusing her of any wrongdoing. This can be illustrated in Chinese by the following questions:

"Ni gan shenma?" 你幹什麼？ (What are you doing?)

and

"Ni gan shenma ne?" 你幹什麼呢？ (What are you doing?).

Although both questions have the same meaning, the particle *ne*呢 in the second question is context-free and syntax-free. Moreover, it softens the expressive mood of the sentence. The fact that the selection of particles in this study did not take into consideration the mood of the narrator and speaker in the novel probably would cause a slight fluctuation of the relative frequency in the particles, but the fluctuation would not be so great that it would affect the test result.

Another important fact to remember is that Text A and Text B were in manuscript form. The interference of the copyists cannot be ignored, although no one knows the extent of the interference. As for Text C (Chapters 81-120), Gao E 高鶚 and Cheng Weiyuan 程偉元 modified the collected text because some of the chapters were found segmentary and the content disrupted. The interference of their editorship is clear, if the assumption is made that these chapters were not fabricated by Gao E. In fact, a look at the list of the particles would lead one to infer that a few particles, like *a*_啊, *ya* 呀, *na* 哪, *lie* 咧, and *balie* 罷咧, occur more often in Text C than in Texts A and B. This is evidence of colloquialization of the text by either the author or editor.[27] The evidence of colloquialization is supported by the decreasing occurrence of the classical particle *yi* 矣, which occurs ten times in Text A and decreases to a frequency of zero in Text C. Three other classical particles have changes in frequency. They are *zai*哉, from four times in Text A to zero in Texts B and C; *ye* 也, from nine times in Text A to zero in Text C; and *eryi* 而已, from twelve times in Text A to once in Text C. More obvious evidence of colloquialization of the language of the novel can be seen in the decrease in occurrence of the classical attribute *zhi* 之, which occurs 242 times in Text A, 197 times in Text B, and only fifty-nine times in Text C.

So far the similarity of the frequency distribution of the particles in the three text samples of *The Dream* has been judged by inspection only. It may, however, be desirable to have at one's disposal a statistical analysis in order to give an objective view of the correlation between the particle distribution in the three texts of *The Dream* and *The Gallant Ones*. Related information for the performance and result of the correlation coefficient test is organized in Table 21.

Correlations between Texts A and B, for instance, are calculated according to Herdan's formula.[28]

$$r_{AB} = \frac{\text{Overlap between A and B}}{\sqrt{\text{vocab. size of A x vocab. size of B}}}$$

Table 21

Correlation Coefficient of Overlaps in Particles of the Four Texts

| Text | Size of vocab. | Pair of attri. | Overlaps | N | r |
|---|---|---|---|---|---|
| A | 15 | AB | 11 | 24 | 0.8563 |
| B | 11 | AC | 12 | 24 | 0.8281 |
| C | 14 | AD | 13 | 24 | 0.7156 |
| D | 22 | BC | 9 | 24 | 0.7253 |
| | | BD | 9 | 24 | 0.5785 |

The correlation coefficient of the non-contextual particles between Text C and Text A or Text B is accordingly higher than that between Text D and Text A or Text B. Once again the results of this test support the previous test results, i.e., the resemblance of Texts A and B of *The Dream* to Text C of *The Dream* is much closer than their resemblance to text D.

The phi coefficient and the coefficient of association Q of each pair of attributes are also computed by information arranged in the following manner (Table 22).

Table 22

Particles – Phi Coefficient Index Between Texts A and B

| Text | B | Non-B | Total |
|---|---|---|---|
| A | 11 | 4 | 15 |
| Non-A | 0 | 9 | 9 |
| Total | 11 | 13 | 24 |

$$\phi_{ab} = \frac{11 \times 9 - 4 \times 0}{\sqrt{11 \times 13 \times 9 \times 15}} = 0.7125$$

$$SE_{ab} = \frac{1 - 0.7125^2}{\sqrt{24}} = 0.1004$$

The phi coefficient of attribute AB differs from zero by more than two standard errors (SE). Thus it admits the conclusion of a significant association between this pair of texts. This means that the overlaps of particles in such a pair can be regarded as implying a real correlation. Table 23 shows the result of the computation of the phi coefficient and the coefficient of association Q of all the paired attributes.

Table 23

Particles – Associations

| Pair of attri. | Value of Q | Value of Phi |
|---|---|---|
| AB | +1 | +0.7125 |
| AC | +0.8666 | +0.5674 |
| AD | –1 | –0.2335 |
| BC | +0.7560 | +0.4381 |
| BD | –1 | –0.0757 |

The results of the association figures are inevitably high. The striking fact is that pairs AB, AC, and BC all have high positive associations whereas pairs AD and BD have negative associations. This seems to indicate that the association method for presence or absence of non-contextual particles from the sample is likely to be a more useful method of investigation of works of disputed authorship than correlation between number of occurrences.

Adding the results of these last two correlation tests to the figures in Table 19, one gets the following information in Table 24.

Now Texts, A, B, and C (pairs AC, BC) have a total of twenty-four positive correlations, whereas Texts A, B, and D (pairs AD, BD) have twelve; Texts A, B, and C have six negative correlations, whereas Texts A, B, and D

Table 24

Association of Attributes

| Pair of attributes | No. of positive correlation | No. of negative correlation | Total test |
|---|---|---|---|
| AB | 13 | 2 | 15 |
| AC | 11 | 4 | 15 |
| AD | 5 | 10 | 15 |
| BC | 13 | 2 | 15 |
| BD | 7 | 8 | 15 |

have nineteen. The test results show that the three texts of *The Dream* have a high positive correlation in their vocabulary distributions. Using Text D, a text undisputedly known to be written by a different author, to test the validity of the statistical tests in this study serves the purpose well, as the same tests performed on Text D indicate a strong negative association to the undisputed text of *The Dream*, Chapters 1 to 80.

This study has not neglected the possibility that minor changes in the text since the date of its origin may have been caused by scribes and editors. Nevertheless, the test results indicate that in spite of such possible changes, deletions, or additions, an overall style has been retained, as measured by the linguistic variables examined.

## Result of the Experiment

The computerized data and statistical analyses used in this research provide a more detailed vocabulary analysis as an approach to the problem of the disputed authorship of *The Dream* than has hitherto been carried out. Results of this analysis suggest a rejection of the popular view that *The Dream* is of dual authorship. These results invalidate the view that Chapters 81-120 were completely authored by Gao E. This study supports the likelihood that *The Dream of the Red Chamber* is substantially a work of single authorship.

## Notes

1. The word "unconscious" is used here to mean the status of being unaware of something because of one's personal or social habit. It does not have any meaning related to "subconscious."
2. Robert Stanley Wachal, "Linguistic Evidence, Statistical Inference, and Disputed Authorship," Diss. University of Wisconsin, Madison, 1966, p. 16.
3. Thomas Corwin Mendenhall, "The Characteristic Curves of Composition," *Science*, IV (1887), 237-49.
4. Pierre Guiraud, *Les Caracteres Statistiques du Vocabulaire* (Paris, 1954).
   E. A. Kirkpatrick, "Number of Words in Ordinary Vocabulary," *Science*, XVIII (1891), 107-08.
   Wincenty Lutoslavski, *The Origin and Growth of Plato's Logic with an Account of Plato's Style and of the Chronology of His Writings* (London, 1905).
   Gustav Herdan, *Language as Choice and Chance* (Groningen, 1956).
   Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge, 1944).
5. Alvar Ellegard, *A Statistical Method for Determining Authorship: the Junius Letters, 1769-1772* (Goteborg: n.p., 1962).
6. F. Mosteller and D. L. Wallace, *Inference of Authorship: the Federalist* (Reading: Addison-Wesley Publishing Co., 1964).
7. Yule, p. 177.
8. C. B. William, *Style and Vocabulary: Numerical Study* (London: University College London Press, 1969), p. 16.
9. Ibid., p. 77.
10. For more studies on stylostatistics, see *An Annotated Bibliography of Statistical Stylistics*, ed. by Richard W. Bailey and Lubomir Dolezel (Ann Arbor: Michigan Slavic Publications, 1968).
11. For a detailed discussion of the *Gengchen* Version and its new reprinted edition, see Feng Qiyong 馮其庸 *Lun Gengchen Ben* 論庚辰本 (A Discussion on the *Gengchen* Version) (Shanghai: Shanghai Wenyi Chubanshe 上海文藝出版社, 1978). The *Gengchen* Version used for this study is a 1959 photo-reprinted copy by Wenyuan Chubanshe 文淵出版社 of Taipei, under the title of *Guben Honglou Meng* 古本紅樓夢.
12. The edition of the *Chengyi* Version used in this study is a photo-reprinted copy of 1962 by Qingshi Shanzhuang Chubanshe 青石山莊出版社 of Taipei, entitled *Yingyin Qianlong Renzi Nian Mu Huozi Ben Bainian Hui Honglou Meng* 影印乾隆壬子年木活字本百廿回紅樓夢. The original copy of this photo-reprinted edition is owned by Hu Tianlie 胡天獵.
13. See note 8.
14. Wen Kang 文康, *Ernu Yingxiong Zhuan* 兒女英雄傳 (The Gallant Ones) edited by Miao Tianhua 繆天華 (Taipei: Sanmin Shuju Youxian Gongsi 三民書局有限公司, 1976).
15. Wang Li 王力, *Zhongguo Xiandai Yufa* 中國現代語法 (Modern Chinese Morphology) (Hong Kong: Zhonghua Shuju 中華書局, 1959).
16. Y. R. Chao 趙元任, *A Grammar of Spoken Chinese* (Berkeley and Los Angeles: University of California Press, 1965).
17. For detail of analysis and verification of data, see Bing-cho Chan, "The Authorship of *The Dream of the Red Chamber*," Diss. University of Wisconsin, 1980.
18. G. Herdan, *The Advanced Theory of Language as Choice and Chance* (Berlin and

Heidelberg: Springer, 1966), p. 41.

19. The Correlation Coefficient test is used here as a comparison index in the measure of correlation. The Product-moment index can be used as an indication of the degree of parallelism between rates of usage, and as index of similarity and dissimilarity.
20. To verify these figures, see Chan, Appendices 2 & 3.
21. Chan, Appendices 2 & 3.
22. Leonard A. Marascuilo, *Statistical Methods for Behavioral Science Research* (New York: McGraw-Hill, 1971), pp. 402-12.
23. Ibid.
24. Herdan, p. 156.
25. Yule, pp. 164-65.
26. Ibid.
27. P'an Ch'ung-k'uei 潘重規, *Honglou Meng Xin Bian* 紅樓夢新辨 (Taipei: Wen Shi Zhe Chubanshe文史哲出版社, 1964), pp. 225-30.
28. Herdan, p. 153.

# Appendix 1

The word distribution table for text A adjective*

| Col. 1 No Occurrences | Col. 2 Frequency | Col. 3 FX | Col. 4 SF FR Bottom | Col. 5 SFX FR Bottom | Col. 6 Normalized SF | Col. 7 Normalized SFX |
|---|---|---|---|---|---|---|
| 1 | 194 | 194 | 279 | 864 | 10000. | 10000. |
| 2 | 34 | 65 | 85 | 670 | 3047. | 7755. |
| 3 | 16 | 48 | 51 | 602 | 1828. | 6968. |
| 4 | 8 | 32 | 35 | 554 | 1254. | 6412. |
| 5 | 10 | 50 | 27 | 522 | 968. | 6042. |
| 6 | 4 | 24 | 17 | 478 | 609. | 5463. |
| 7 | 1 | 7 | 13 | 448 | 466. | 5185. |
| 8 | 2 | 16 | 12 | 441 | 430. | 5104. |
| 9 | 0 | 0 | 10 | 425 | 358. | 4919. |
| 10 | 0 | 0 | 10 | 425 | 358. | 4919. |
| 11 | 1 | 11 | 10 | 425 | 358. | 4919. |
| 12 | 1 | 12 | 9 | 414 | 323. | 4792. |
| 13 | 2 | 26 | 8 | 408 | 287. | 4653. |
| 14 | 0 | 0 | 6 | 376 | 215. | 4352. |
| 15 | 0 | 0 | 6 | 376 | 215. | 4352. |
| 16 | 0 | 0 | 6 | 376 | 215. | 4352. |
| 17 | 0 | 0 | 6 | 376 | 215. | 4352. |
| 18 | 0 | 0 | 6 | 376 | 215. | 4352. |
| 19 | 1 | 19 | 6 | 376 | 215. | 4352. |
| 20 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 21 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 22 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 23 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 24 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 25 | 0 | 0 | 5 | 357 | 179. | 4132. |
| 26 | 1 | 26 | 5 | 357 | 179. | 4132. |
| 27 | 0 | 0 | 4 | 331 | 143. | 3831. |
| 28 | 0 | 0 | 4 | 331 | 143. | 3831. |
| 29 | 0 | 0 | 4 | 331 | 143. | 3831. |
| 30 | 0 | 0 | 4 | 331 | 143. | 3831. |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 117 | 1 | 117 | 1 | 117 | 36. | 1354. |

S0 = 279
S1 = 864
S2 = 33.540
= 437.725
= 3.097

Variance = 110.625
Standard Deviation = 10.518
= 3.396
Percentage of Adjective Occurring once only = 69.53
Vocabulary per 1000 occurrences = 3.22

*For full table listing, see Bing-cho Chan, "The Authorship of *The Dream of the Red Chamber*: A Computerized Statistical Study of Its Vocabulary," Diss. University of Wisconsin – Madison 1980, pp. 148-150.

## Appendix 2

| Stative verb* | FRFQ: | Text A | Text B | Text C | Text D |
|---|---|---|---|---|---|
| ai 矮 | | 0 | 2 | 0 | 0 |
| aiqie 哀切 | | 0 | 0 | 1 | 0 |
| aitong 哀痛 | | 1 | 0 | 0 | 0 |
| aixi 愛惜 | | 0 | 0 | 2 | 0 |
| aiyan 礙眼 | | 0 | 1 | 0 | 0 |
| an 安 | | 3 | 4 | 5 | 4 |
| anding 安定 | | 1 | 0 | 0 | 0 |
| andun 安頓 | | 0 | 0 | 1 | 0 |
| aneen 安份 | | 1 | 0 | 0 | 0 |
| angzang 骯髒 | | 0 | 0 | 0 | 1 |
| anjing 安靜 | | 2 | 2 | 6 | 0 |
| anlian 諳練 | | 0 | 0 | 1 | 0 |
| anwen 安穩 | | 2 | 0 | 1 | 1 |
| anxin 安心 | | 0 | 1 | 0 | 0 |
| anyi 安易 | | 0 | 0 | 2 | 0 |
| anzo 暗 | | 1 | 1 | 0 | 0 |
| aoman 傲慢 | | 1 | 0 | 0 | 1 |
| aonao 懊惱 | | 1 | 1 | 0 | 0 |
| badao 霸道 | | 0 | 1 | 0 | 2 |
| bai 白 | | 0 | 1 | 0 | 3 |
| bao 鮑 | | 0 | 1 | 1 | 2 |
| baozt 薄 | | 1 | 2 | 1 | 2 |
| . | | . | . | . | . |
| . | | . | . | . | . |
| . | | . | . | . | . |
| zu 是 | | 1 | 2 | 0 | 0 |
| zungui 尊貴 | | 1 | 1 | 1 | 3 |
| zunzhong 尊重 | | 0 | 0 | 1 | 0 |
| zuoqiang 作戕 | | 1 | 0 | 0 | 0 |

Total number of different stative verbs yielded by Texts A, B, C, D is 1049
number of stative verbs that are shared commonly in Texts A, B, C, D is 64

*Chan, pp. 489-536.