# Mapping Stylistic Variation with Correspondence Analysis (以对应分析统计法图示语体变异维度)

Zhang, Zheng-sheng
(张正生)

San Diego State University
(圣地亚哥州立大学)
zzhang@sdsu.edu

**Abstract**: This article demonstrates the use of an alternative method for carrying out multi-dimensional research on stylistic (or register) variation. Correspondence Analysis (CA) is an alternative statistical procedure to Factor Analysis, which has been employed in most multi-dimensional studies on stylistic variation. The advantages of CA are ease of use and intuitive visualization in the form of "stylistic maps." The method will be illustrated with the author's work on stylistic variation in written Chinese as well as a pilot study using English.

**提要：**本文介绍一种多面向语体研究的另类统计方法。对应分析（Correspondence Analysis）是普遍使用的因子分析的一种变异形式。此法的优点是容易使用并可提供直观的语体分布图。方法的演示将以中文书面语体研究为例，以及英语语体的一个初步研究。

**Keywords:** Stylistic variation, factor analysis, correspondence analysis, stylistic map

**关键词：**语体变异、因子分析、对应分析、语体分布图

## 1. Multi-feature, Multi-dimensional Study of Stylistic/Register Variation

Even though the focus of this paper is on illustrating the use of an alternative methodology, a brief introduction is nonetheless in order on stylistic/register variation and the multi-dimensional framework for the study of such variation. For more in-depth discussion of these matters, the reader is advised to refer to a series of publications by the author listed in the references, especially Zhang (2017).

### 1.1 Stylistic/register variation

It may be safe to say that stylistic/register variation exists in all languages. Stylistic/register characteristics, such as formality and literariness, can be seen in different areas of language, most notably in the lexical domain. In English, a contrast can be seen between the colloquial *eat, buy* and their formal counterparts *dine*, *purchase*; in Chinese, there are also lexical doublets, such as 买 [buy] and 购 [purchase], 在 [at] and 于 [at]. One commonly evoked stylistic/register distinction is that of spoken vs. written or formal vs. informal. More subtly, as demonstrated by Biber, Douglas, Johansson, Leech, Conrad, & Finegan (1999), even basic syntactic categories, such as parts of speech, can be shown to have certain stylistic affinities. They show that a greater frequency of nouns and nominalized elements seems to be a hallmark of formal written texts.

### 1.2 Problem with simple binary distinctions

A single dichotomous distinction such as spoken vs. written simply cannot accommodate the complexity of stylistic variation. For example, this distinction may not align with another commonly evoked distinction of formal vs. informal. What is written in style may not always be formal, and vice versa. The same can be said about a distinction based on formality. In the case of Chinese, even though classical Chinese elements dating back to the pre-Qin era, such as 与 (“and,” which corresponds to 和 in modern Chinese), have a close affinity with the modern written style, they may not always be the most formal. As we will see later, the most formal registers actually do not have the most classical Chinese elements. In order to accommodate the intricacies of stylistic/register variation, obviously more than one distinction is required.

Past work on stylistics was mostly on features in isolation, such as word length and lexical and syntactic choices, in the absence of a broader account of how such features may figure in the overall scheme of stylistic variation. As an example of specific problems with single features, categorization based on single features can be problematic, as different features may lead to contradictory classifications. For example, word length as feature cannot be used to classify texts into written vs. non-written types. Monosyllabicity, a distinct characteristic of classical Chinese often associated with the written style, cannot be used to characterize modern written Chinese as a whole, which tends to favor disyllabic words.

A third problem is methodological. Much of previous work on stylistics was based on introspection and anecdotal evidence, without empirical and quantitative support. For example, the stylistic markings in dictionaries, mostly based on the compilers’ intuition, are neither complete nor consistent (Zhang, 2017). Introspection is also limiting. The aforementioned stylistic affinity of syntactic categories like nouns and morphological processes like nominalization will be hard to ascertain by introspection alone, without using corpora and statistical methods. Finally, without the support of quantitative information, it will be hard to go beyond simple dichotomies and entertain the possibilities of continuums, which are gradient in nature.

### 1.3 Multi-feature and multi-dimensional (MM) framework

Biber (1988) was perhaps the most influential work in introducing a multi-feature and multi-dimensional (MM) paradigm. MM-style research is different in three major ways:

a. multiple features, instead of single features, are examined simultaneously
b. multiple dimensions, instead of a single distinction, are entertained
c. more empirical, being corpus-based and quantitative in methodology

Using 67 mostly grammatical features, 23 registers from spoken and written corpora (such as press reports, official documents, and speeches) and the statistical method of Factor Analysis, Biber (1988) extracted six dimensions of register variation for English, driving home the fact that no single feature/dimension can characterize register variation:

1. Informational vs. involved production
2. Narrative vs. non-narrative
3. Explicit vs. situation-dependent reference
4. Overt expression of persuasion
5. Abstractness/nonspecific
6. Online informational elaboration

Since Biber's initial study on English, a number of MM studies of register variation have been carried out for other languages, including one on Taiwanese (Jang, 1998) and a number of them on Mandarin Chinese by the present author (Zhang, 2012, 2013, 2016, 2017, & forthcoming).

According to Zhang (forthcoming), two dimensions, rather than a single dichotomous distinction, characterize the style in modern written Chinese. In addition to the literate dimension, which is related to the common spoken vs. written (or formal vs. informal) distinction, there is a separate dimension, dubbed Alternative Diction, concerning the manner of expression. This will be explained in Section 4, where the two dimensions are presented in detail.

Although not couched in the same terms, the idea of two dimensions also seems to be implicit in Feng's (2010) analysis, which is mostly derived conceptually rather than using corpus data. His analysis separates situation (正式 [formal], associated with 现代书面语 [modern written Chinese]) from diction (庄典 [dignified and elegant], associated with 古代词语 [classical diction). Texts using more classical diction are not necessarily more formal, nor vice versa. To show the disassociation of diction from formality, Feng uses the examples of the *Yellow Emperor Epitaph* [黄帝祭文] and *Romance of the Western Chamber* [西厢记]. Although the *Yellow Emperor Epitaph* is both formal and classical, the classically-worded *Romance of the Western Chamber* is not formal at all.

## 1.4 Gathering data in MM-style variation research

Research on stylistic variation requires frequency data of the relevant linguistic features in various contexts, such as registers. Therefore, the initial steps in a MM-style study are:

a. Selecting stylistically relevant linguistic features
b. Selecting a balanced corpus with multiple types

c. Collecting frequency data of features in various types

In what follows, each of these steps will be illustrated.

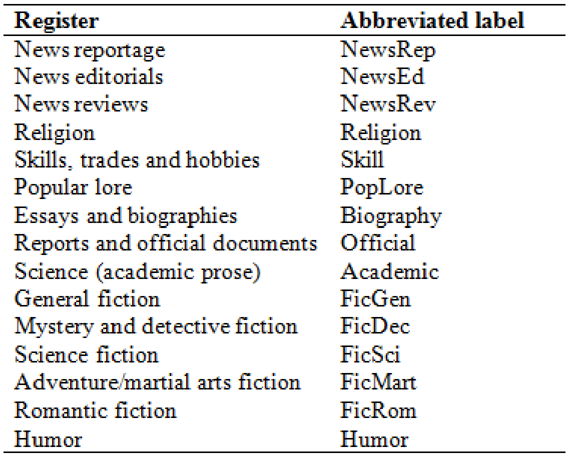
### 1.4.1 Feature selection

To avoid any *a priori* assumption, in principle any feature that is potentially relevant to stylistic variation should be included. It can be a lexical item, a phrase, a syntactic category, such as noun and verb, and even a whole construction. However, there are practical considerations that limit the number and type of features that can be included. First of all, only those that can be searched for with little or no manual work can be included. Searching for discontinuous strings such as 除了… 以外 [ in addition; except for] will require extra work. To ensure reliability, only features with high enough occurrences should be selected. This consideration favors whole parts of speech and function words. In order for the features to be visually legible on the bi-plots generated by SPSS, the number of features should not be too large either.

### 1.4.2 Balanced corpora

As the most important information in stylistic variation concerns how style varies in different contexts, the selected corpus needs to provide such variant contexts, namely multiple registers. In other words, the corpus needs to be balanced. One example of balanced corpora is the Lancaster Corpus of Mandarin Chinese (LCMC), which includes as many as 15 registers, despite its modest size of one million words (see Table 1).

**Table 1. LCMC registers**



| Register | Abbreviated label |
|---|---|
| News reportage | NewsRep |
| News editorials | NewsEd |
| News reviews | NewsRev |
| Religion | Religion |
| Skills, trades and hobbies | Skill |
| Popular lore | PopLore |
| Essays and biographies | Biography |
| Reports and official documents | Official |
| Science (academic prose) | Academic |
| General fiction | FicGen |
| Mystery and detective fiction | FicDec |
| Science fiction | FicSci |
| Adventure/martial arts fiction | FicMart |
| Romantic fiction | FicRom |
| Humor | Humor |

As can be seen, LCMC is quite finely differentiated. Some major register categories have subcategories. Fiction alone has five different subtypes and the journalistic type have three subtypes. Such fine differentiation allows us to gain insight into intraregister as well as interregister variation.

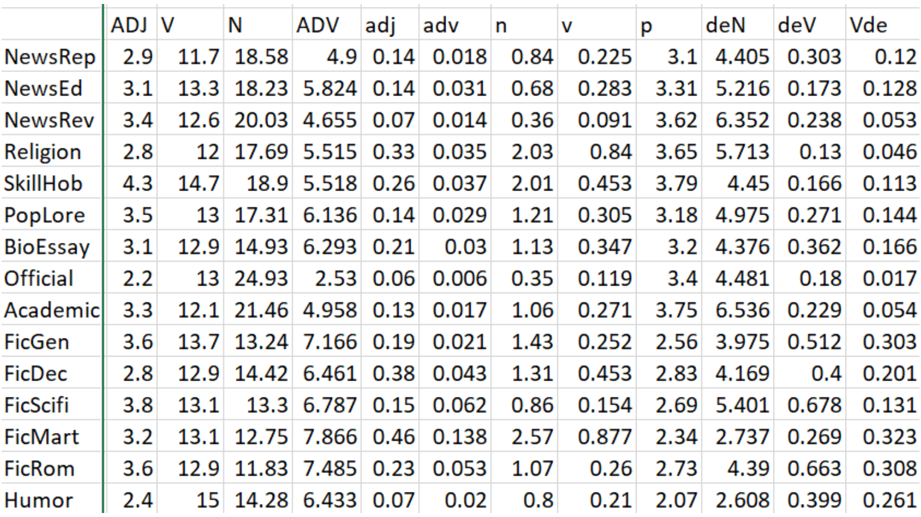
Another well-known balanced corpus is the much larger BCC corpus from Beijing Language and Cultural University, albeit with only four types: 文学 [literature], 报刊

[press], 微博 [tweets], and 科技 [science and technology]. Its large size will prove vital in studying lower frequency lexical items, such as near synonyms.

### 1.4.3 Collecting frequency data

Frequency data of features in registers can typically be obtained by searching the chosen corpus through an interface. For example, LCMC and several other similarly structured corpora (UCLA, ZCTC) are available online at the Beijing Foreign Studies University CQP website (http://111.200.194.212/cqp/), which hosts altogether 41 corpora, both Chinese and English. The BCC corpus has its own online interface, which provides rather extensive search options. The frequency information is given both in raw counts and normalized counts (typically per million words), which facilitates comparisons between registers.

The retrieved frequency data can be first organized in an Excel file, before they are entered into SPSS. Seen in Figure 1 is a partial screen capture using the LCMC corpus with its 15 registers.



| | ADJ | V | N | ADV | adj | adv | n | v | p | deN | deV | Vde |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NewsRep | 2.9 | 11.7 | 18.58 | 4.9 | 0.14 | 0.018 | 0.84 | 0.225 | 3.1 | 4.405 | 0.303 | 0.12 |
| NewsEd | 3.1 | 13.3 | 18.23 | 5.824 | 0.14 | 0.031 | 0.68 | 0.283 | 3.31 | 5.216 | 0.173 | 0.128 |
| NewsRev | 3.4 | 12.6 | 20.03 | 4.655 | 0.07 | 0.014 | 0.36 | 0.091 | 3.62 | 6.352 | 0.238 | 0.053 |
| Religion | 2.8 | 12 | 17.69 | 5.515 | 0.33 | 0.035 | 2.03 | 0.84 | 3.65 | 5.713 | 0.13 | 0.046 |
| SkillHob | 4.3 | 14.7 | 18.9 | 5.518 | 0.26 | 0.037 | 2.01 | 0.453 | 3.79 | 4.45 | 0.166 | 0.113 |
| PopLore | 3.5 | 13 | 17.31 | 6.136 | 0.14 | 0.029 | 1.21 | 0.305 | 3.18 | 4.975 | 0.271 | 0.144 |
| BioEssay | 3.1 | 12.9 | 14.93 | 6.293 | 0.21 | 0.03 | 1.13 | 0.347 | 3.2 | 4.376 | 0.362 | 0.166 |
| Official | 2.2 | 13 | 24.93 | 2.53 | 0.06 | 0.006 | 0.35 | 0.119 | 3.4 | 4.481 | 0.18 | 0.017 |
| Academic | 3.3 | 12.1 | 21.46 | 4.958 | 0.13 | 0.017 | 1.06 | 0.271 | 3.75 | 6.536 | 0.229 | 0.054 |
| FicGen | 3.6 | 13.7 | 13.24 | 7.166 | 0.19 | 0.021 | 1.43 | 0.252 | 2.56 | 3.975 | 0.512 | 0.303 |
| FicDec | 2.8 | 12.9 | 14.42 | 6.461 | 0.38 | 0.043 | 1.31 | 0.453 | 2.83 | 4.169 | 0.4 | 0.201 |
| FicScifi | 3.8 | 13.1 | 13.3 | 6.787 | 0.15 | 0.062 | 0.86 | 0.154 | 2.69 | 5.401 | 0.678 | 0.131 |
| FicMart | 3.2 | 13.1 | 12.75 | 7.866 | 0.46 | 0.138 | 2.57 | 0.877 | 2.34 | 2.737 | 0.269 | 0.323 |
| FicRom | 3.6 | 12.9 | 11.83 | 7.485 | 0.23 | 0.053 | 1.07 | 0.26 | 2.73 | 4.39 | 0.663 | 0.308 |
| Humor | 2.4 | 15 | 14.28 | 6.433 | 0.07 | 0.02 | 0.8 | 0.21 | 2.07 | 2.608 | 0.399 | 0.261 |

**Figure 1. Data Matrix in Excel file (columns= features; rows=registers)**

## 2. Two Statistical Procedures

To make sense of the large amount of frequency data gathered, statistical methods have to be employed to extract a smaller number of underlying factors or dimensions, which may reveal some meaningful patterns. This kind of procedure is generally referred to as dimension reduction. In this section, two kinds of dimension reduction procedures will be briefly described, namely Factor Analysis and Correspondence Analysis.

### 2.1 Factor Analysis

A method such as Factor Analysis is commonly used to reveal underlying factors hidden in the bewildering amount of variability. According to Wikipedia (https://en.wikipedia.org/wiki/Factor_analysis), "Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors**"** (see Gorsuch, 1983, for a classical exposition of the method).

In the field of teaching Chinese as a foreign language, Factor Analysis has been used in the study of learning strategies, such as those for learning Chinese characters (Shen, 2005) and in the study of rating criteria used in assessment of learners of Chinese (Chen, 2016).

For the non-statisticians though, Factor Analysis presents quite a steep learning curve. Correspondence Analysis (CA), which also can accomplish the reduction of large amount of data into smaller number of dimensions, provides a more user-friendly alternative.

## 2.2 Correspondence Analysis (CA)

Even though Factor Analysis has been the preferred method for dimension extraction in MM-oriented research, the present author's studies have all employed Correspondence Analysis. According to an online publication:

> In a nutshell, correspondence analysis (CA) may be defined as a special case of principal components analysis (PCA) of the rows and columns of a table, especially applicable to a cross-tabulation. …… Its primary goal is to transform a table of numerical information into a graphical display, in which each row and each column is depicted as a point (https://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/CorrespondenceAnalysis/UNESCO_IDAMS_CorrespAnal.pdf; for details of this method, refer to Greenacre, 1984).

CA has been commonly used in market research, such as brand preference, by different demographic groups. It has also been used in literary studies to uncover hidden patterns of linguistic usage, such as a writer's preferences in lexical choices. According to Gries (2015), CA is only occasionally used in corpus linguistics work. As far as the author is aware, it has not been adopted for the study of register variation using the MM model.

Correspondence Analysis is much easier to use than Factor Analysis. It is highly flexible with data requirements, the only strict data requirement being a rectangular data matrix made up of columns and rows with no negative entries. According to Tabata (2007), which also employs Correspondence Analysis instead of Principle Component Analysis (PCA) and Factor Analysis (FA) in his study of English literary authors, "one advantage CA has over PCA and FA is that PCA and FA cannot be computed on a rectangular matrix where the number of columns exceeds the number of rows." As the number of columns is most likely many times the number of rows, the data do not readily lend themselves readily to Factor Analysis without extensive re-organization. There is also no need to deal with the choice of rotation methods, which do produce different results. Like Factor Analysis, Correspondence Analysis is also available in common statistical packages such as SPSS.

The greatest appeal of Correspondence Analysis lies in its intuitive bi-plot visualization: “Categories that are similar to each other appear close to each other in the plots. In this way, it is easy to see which categories of a variable are similar to each other or which categories of the two variables are related” (SPSS help). In other words, the closer things are together, the more alike they tend to be, as in the Chinese saying 物以类聚 [birds of a feather flock together]. This kind of intuitive visualization can help detect relationships among categories and also aid in the interpretation of dimensions. Correspondence Analysis is therefore better suited for exploration and practical applications. An example of a bi-plot is given in Figure 2, which shows the distribution of the 5 text types from COCA (Corpus of Contemporary American English, https://corpus.byu.edu/coca/ ) along the horizontal and vertical dimensions.

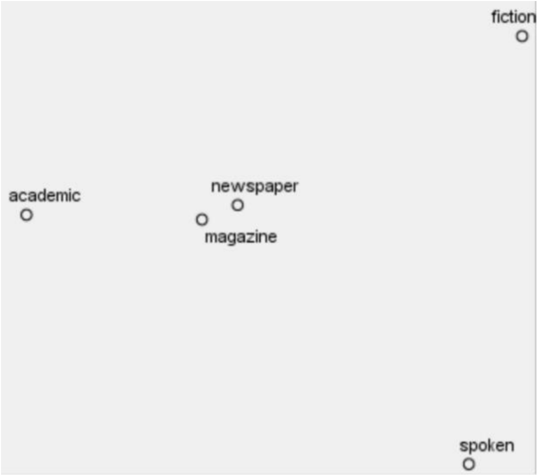


**Figure 2. Distribution of COCA text types**

*Newspaper* and *magazine* are understandably closer together than anything else, as both are journalistic registers. On the other hand, *academic* writing, *fiction* and *spoken* are farthest apart possible, academic being most different from fiction and spoken on the horizontal dimension and fiction and spoken are most distinct from each other on the vertical dimension. Details of the pilot study using COCA will be given in section 5, where the two dimensions will be interpreted.

## 3. Running Correspondence Analysis in SPSS

In this section, the technical details on applying Correspondence Analysis in SPSS will be briefly described. The major steps for extracting dimensions and generating bi-plots include the following:

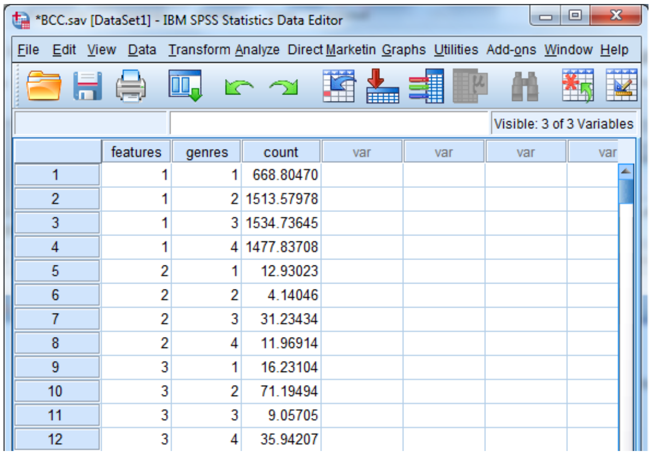
1. Coding frequency data (as shown in Figure 1) in SPSS
2. Selecting options for dimension extraction

3. Selecting options for bi-plot display

Interpreting the dimensions and plots will be done in Section 4, where sample results of the author's recent studies will be presented.

## 3.1 Data coding in SPSS

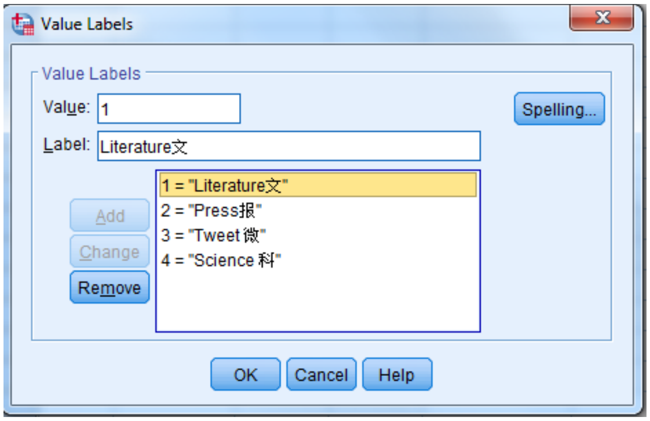
After gathering data in the form of frequency of occurrence of different features in various contexts, the next step is to create the data file in SPSS. A special note is in order on data coding, as this step is potentially the most unintuitive part of the whole process. For researchers familiar with Factor Analysis, a notable difference exists in how data are coded in the SPSS implementation of Correspondence Analysis. Instead of using variables directly to represent linguistic features (columns in the earlier Excel file) and registers (rows in the Excel file), all the linguistic features are represented with one single variable (named, for example, as Feature), which is then divided into the same number of values as the number of original variables; in the same manner, all the register features are also represented with one single variable (possibly named Register), which is likewise divided into the same number of values as the number of original registers. The values of both features (frequencies of features in registers) are coded in the third feature called Count. A screen capture showing this scheme is given in the Figure 3 below. The rows represent the values of the features *Features*, *Genres* (registers) and *Count*:


*BCC.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketin Graphs Utilities Add-ons Window Help
Visible: 3 of 3 Variables

| | features | genres | count | var | var | var | var |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 668.80470 | | | | |
| 2 | 1 | 2 | 1513.57978 | | | | |
| 3 | 1 | 3 | 1534.73645 | | | | |
| 4 | 1 | 4 | 1477.83708 | | | | |
| 5 | 2 | 1 | 12.93023 | | | | |
| 6 | 2 | 2 | 4.14046 | | | | |
| 7 | 2 | 3 | 31.23434 | | | | |
| 8 | 2 | 4 | 11.96914 | | | | |
| 9 | 3 | 1 | 16.23104 | | | | |
| 10 | 3 | 2 | 71.19494 | | | | |
| 11 | 3 | 3 | 9.05705 | | | | |
| 12 | 3 | 4 | 35.94207 | | | | |

**Figure 3. Example of coding scheme in SPSS**

The current version of SPSS (v25) supports both English and Chinese text, as seen in Figure 4, which uses a mixture of Chinese and English in the labels for the text types:



Value: 1

Label: Literature文

1 = "Literature文"
2 = "Press报"
3 = "Tweet 微"
4 = "Science 科"

Figure 4. Text in either English or Chinese

## 3.2 Applying Correspondence Analysis (SPSS)

To run the Correspondence Analysis procedure, choose the *Dimension Reduction* option from the pulldown menu *Analyze* and then choose the *Correspondence Analysis* option, as seen in Figure 5:

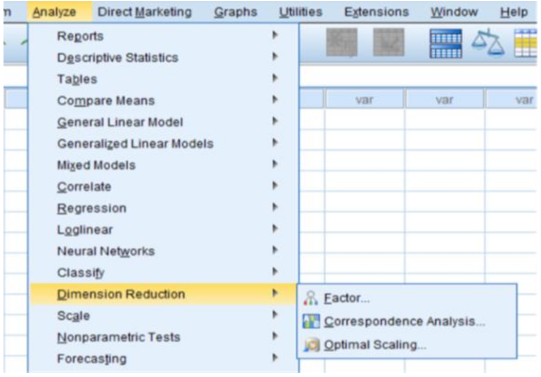
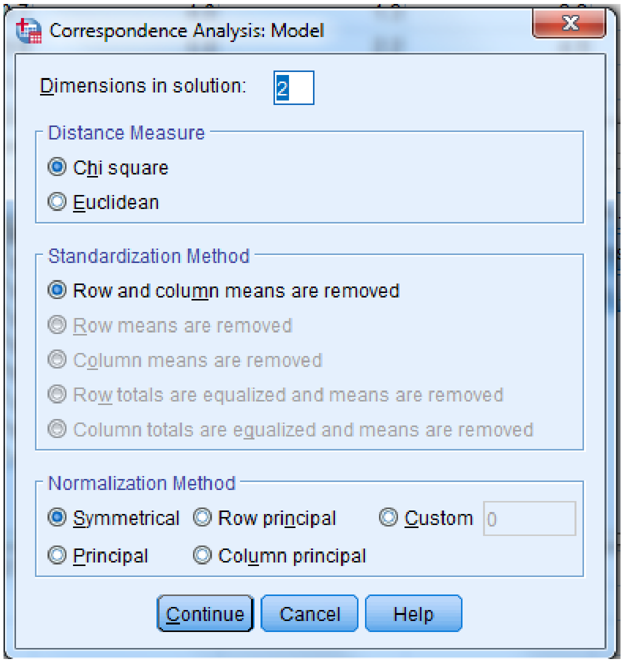


Figure 5. Correspondence Analysis in SPSS

You will then be asked to choose from the various options associated with the extraction of dimensions, as shown in Figure 6:


Correspondence Analysis: Model

Dimensions in solution: 2

Distance Measure

● Chi square
○ Euclidean

Standardization Method

● Row and column means are removed
○ Row means are removed
○ Column means are removed
○ Row totals are equalized and means are removed
○ Column totals are equalized and means are removed

Normalization Method

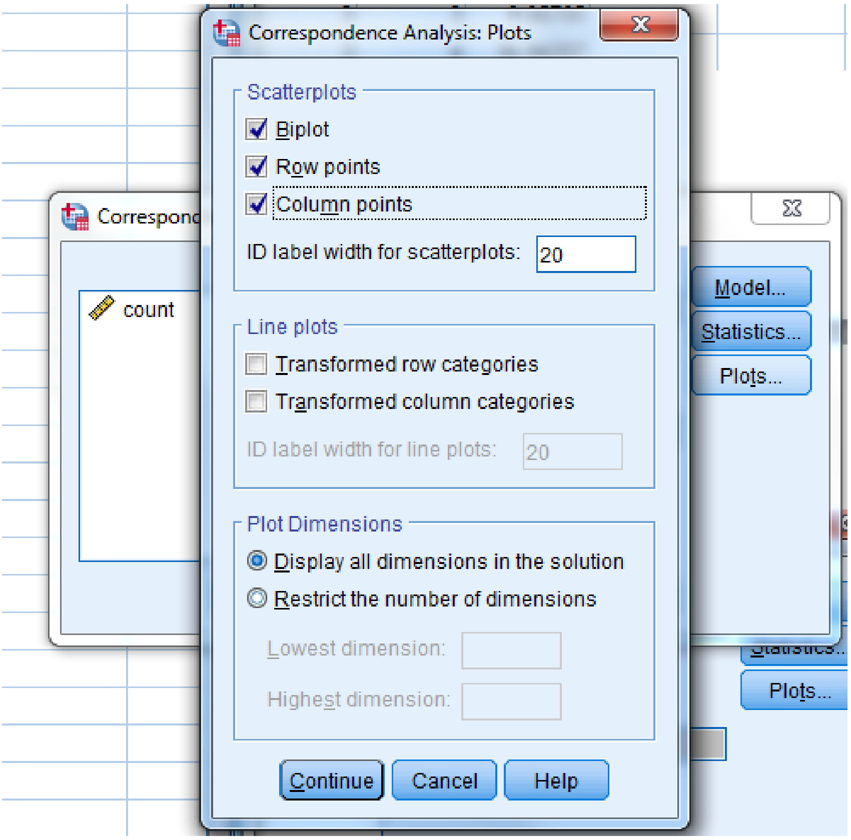
● Symmetrical ○ Row principal ○ Custom 0
○ Principal ○ Column principal

**Figure 6. Options for extraction of dimensions**

The number of dimensions in the solution can be specified by the user. The procedure can automatically generate up to N dimensions (N=number of registers minus 1). So for LCMC, 14 dimensions (15-1) will be generated and for BCC three dimensions (4-1) will be generated. While the present author has found the first 2 dimensions to be most interpretable, it is up to the reader to explore greater number of dimensions.

The default choice for Distance Measure is “Chi Square,” the default for Standardization Method is “Row and column means removed,” and for Normalization Method the default choice is “Symmetrical.” The choices “Row principal” and “Column principal” have the effect of stretching the horizontal or the vertical dimension.

There are also various options for the display of Plots, as seen in Figure 7.



Correspondence Analysis: Plots

Scatterplots

☑ Biplot

☑ Row points

☑ Column points

ID label width for scatterplots: 20

Line plots

□ Transformed row categories

□ Transformed column categories

ID label width for line plots: 20

Plot Dimensions

● Display all dimensions in the solution

○ Restrict the number of dimensions

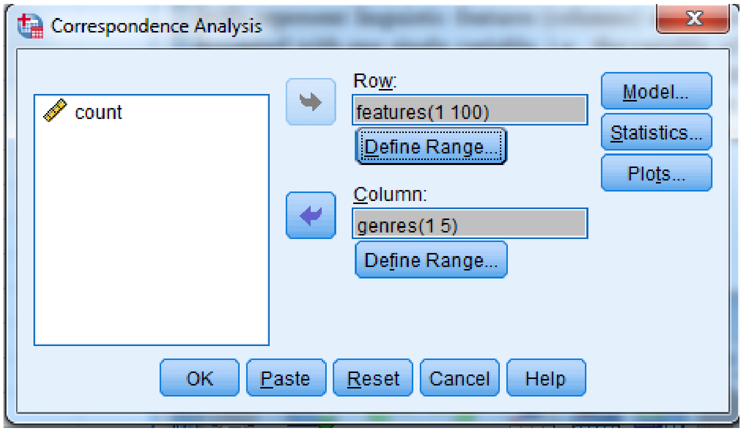
Lowest dimension:

Highest dimension:

**Figure 7. Options for plots display**

When the "Row points" and "Column points" options are checked under Scatterplots options, separate bi-plots for features and registers are generated.

Finally, when running the procedure, it is possible to select a subset of contiguously numbered features or registers with the option of "Define Range," as is shown in Figure 8. This can be used to explore the effects of excluding certain features or registers from computation.



Correspondence Analysis

count

Row: features(1 100)

Column: genres(1 5)

**Figure 8. Option for using subset of features & genres (registers)**

## 4. Illustrative Results: Mapping Stylistic Variation in Written Chinese

In this section, some sample results from two studies in Zhang (2017) will be used as an illustration. They respectively demonstrate the two stylistic dimensions in written Chinese and a more close-up look at two sets of near synonyms mapped along these two dimensions.

### 4.1 Two dimensions of written Chinese (LCMC)

To demonstrate the overall dimensions in written Chinese, the LCMC corpus with more finely differentiated registers (albeit with a small size) will be used. Most of the 50 features used are structurally related items, as they occur more frequently.

In this section, it will be shown that contrary to the commonly assumed one single formal/written vs. informal/spoken distinction, at least two dimensions of stylistic variation are found for written Chinese, which are respectively the “Literate” dimension and the “Alternative Diction” dimension. As mentioned in section 3, even though the number of dimensions can be as many as the number of registers minus 1 (15-1 for LCMC), it has been the author’s experience that only the first two dimensions have sufficiently clear interpretations.

As Correspondence Analysis is mostly used for exploratory studies, measures of statistical significance such as eigen values and scree plots will not concern us further. Suffice it to say although small in number, the total variation accounted for by the two dimensions is very high (around 80%). Furthermore, the correlation between the two dimensions is small, which means the dimensions are sufficiently independent of each other.

#### 4.1.1 Dimension 1: Literate

Dimension 1 (the horizontal dimension) is a very strong dimension, accounting for two thirds of all variation. On the bi-plot in Figure 9, where 50 linguistic features are distributed in a two dimensional space, several clear patterns can be noted:

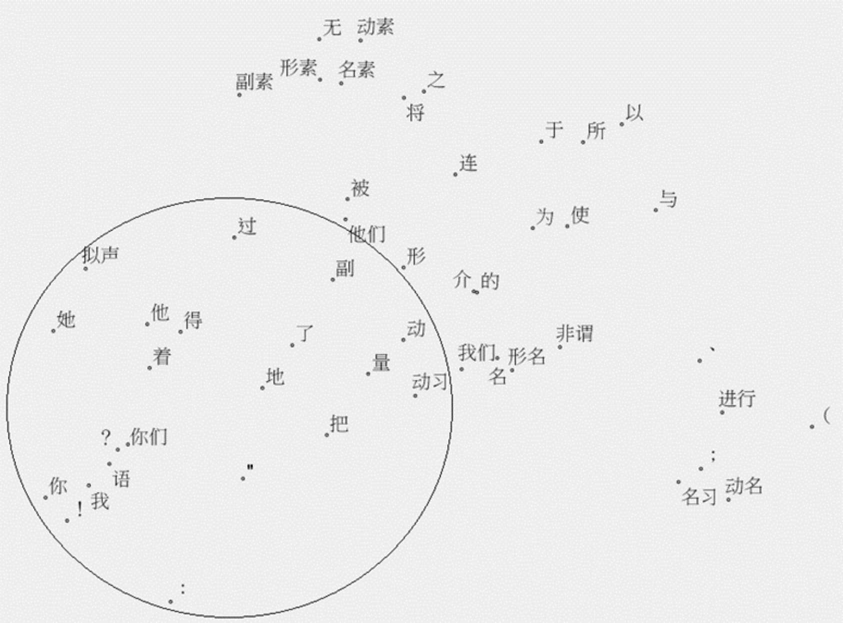


**Figure 9. Clustering of interactive/narrative features on Dimension 1**

1) There is a clear clustering of interactive/narrative features on the left (encircled), such as pronouns (我 [I]、你 [you]、她 [she]、他 [he]、你们[ you plural]), particles (语), aspectual features (了、着、过), the two verbal *de* (地、得) and measure words (量).

2) There are clear contrasts between verbal and nominal features: verbs (动) and their associated features, such as the two verbal *de* (地、得) and the aspectual markers (了、着、过) are on the left, whereas nouns (名) and associated features, such as nominalized verbs (动名) and adjectives (形名) and attributive adjectives (非谓), are on the right.

3) There is a clear contrast between verbs in general (动) and light verbs: verbs in general are very centrally located while light verbs such as *jinxing* (进行) are very peripheral, being almost at the extreme right of the plot. There is a similar contrast between adjectives in general and attributive adjectives: adjectives (形) are fairly centrally located while attributive adjectives (非谓) are to the right of it. The contrast between the three *de* (得、地、的) is also quite astounding. The nominal 的 is clearly to the right of the verbal 得 and 地, as highlighted in Figure 10:

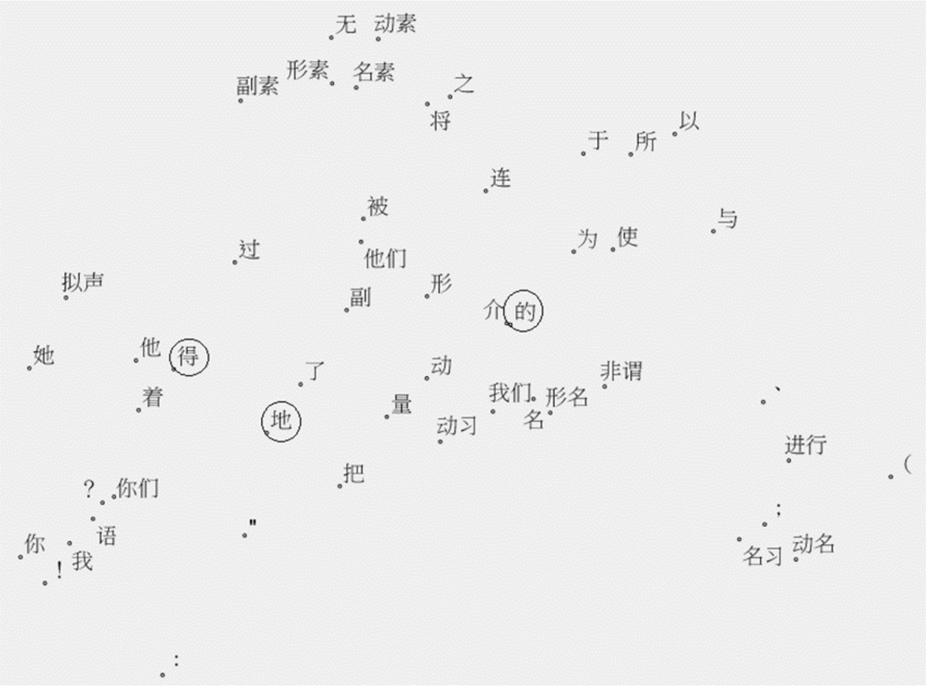


**Figure 10. 得、地、的 on Dimension 1**

4) The contrast between the two kinds of punctuation marks, shown in Figure 11, is no less astounding: question, exclamation, colon and quotation marks are all on the left, whereas parenthesis, semi-colon, and Chinese-style pause marks (、) are all on the right.

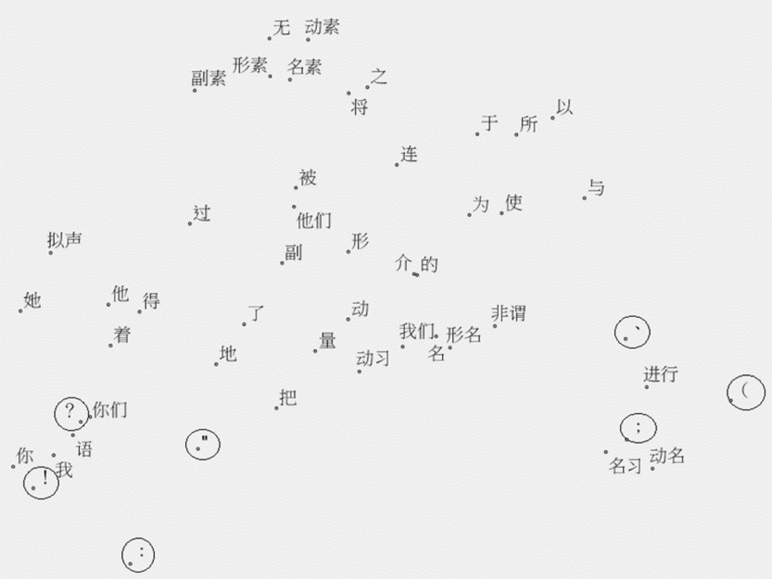


**Figure 11. Two kinds of punctuation marks on Dimension 1**

The distributional patterns are strikingly reminiscent of the commonly evoked "spoken vs. written" distinction, even though the LCMC corpus is exclusively written. It seems the same parameters that distinguish spoken and written styles are also at work here, such as the degree of interactivity, narrativity, nominal vs. verbal and unplanned vs. pre-planned, and so on. We will dub Dimension 1 the "Literate" dimension.

### 4.1.2 Dimension 2: Alternative Diction

While Dimension 1 seems to evoke the commonly used “spoken vs. written” distinction, the existence of the second dimension should be particularly worthy of note. Obviously, this is only possible with the multi-dimensional framework. But substantively, what can the second dimension be?

On the plot in Figure 12, one immediately notices the clear clustering of classical Chinese elements (encircled) on the upper half of the bi-plot.

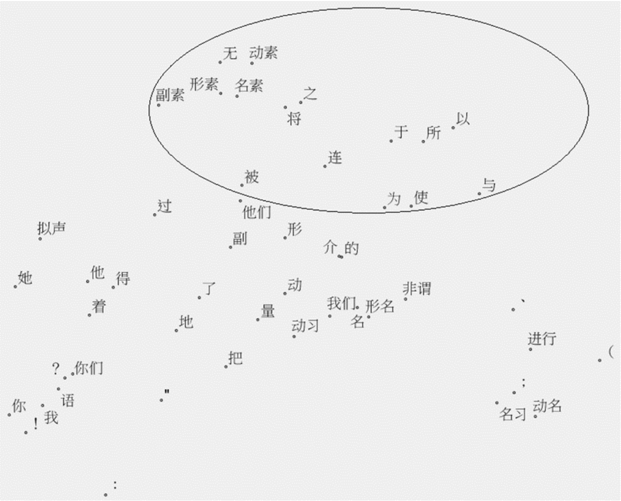


**Figure 12. Clustering of classical elements on Dimension 2**

These classical elements are of two kinds: individual items (为、以、所、与、于、之、将、无、使), located towards the bottom right of the encircled area, and four class features (名素、动素、形素、副素), located towards the top left part of the encircled area. These class features may require some explanation. The 素 in their labels means that these are bound morphemes that can only be part of compound words in modern Chinese but can still occur as standalone words in classical and classical-flavored texts. An example can be given from the LCMC corpus: 堤下鸡鸣，鸟叫，犬吠 [under the dike, roosters crow, birds chirp, and dogs bark]. The three underlined syllables are such morphemes. 鸣 [chirp] is verbal (动素); 犬 [dog] is a nominal morpheme (名素), and; 吠 [bark] is again verbal.

It is worth noting that of the two kinds, the classes of bound morphemes are more positive on this dimension than the individual words, although they are less literate on the “Literate” dimension. This may mean that they are less integrated into modern written Chinese. They also seem to be more likely to be content words (实词) rather than function words (虚词).

The contrast between classical words and their non-classical counterparts can also be seen in the two minimal contrastive pairs of 将 vs. 把 and 之 vs. 的, given in Figure

13. The classical 将 and 之 are both north of their non-classical counterparts 把 and 的 respectively.

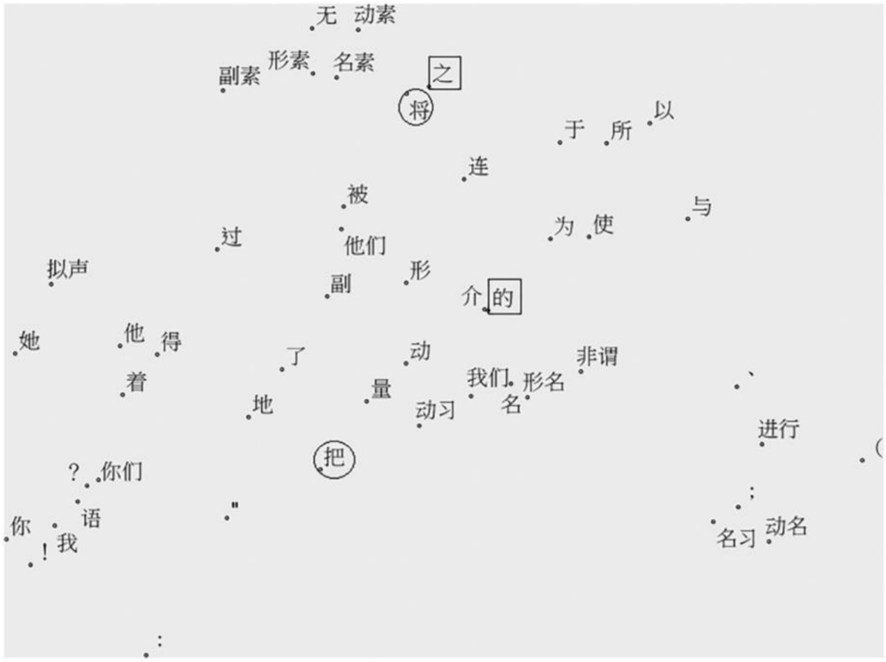


**Figure 13. 将 vs. 把 and 之 vs. 的 on Dimension 2**

Even though "classical" seems to suggest itself readily as the interpretation for Dimension 2, "Alternative Diction" may in fact be more accurate (Zhang, 2017). This is because classical elements are not the only ones with this distribution; some non-classical elements also are similarly distributed. These non-classical items include literary elements, dialectal elements, internet neologisms and other non-canonical forms, whose presence can be seen in larger corpora such as BCC.

The term Alternative Diction may require some explanation, as it is quite unlike the familiar notions of formality and writtenness. The best examples to illustrate Alternative Diction are the minimal pairs presented in Figure 13, which are identical in meaning but distinct in word choice resulting from a stylistic contrast unrelated to the dimension of Literateness. 将 is an alternative word choice to the synonymous 把, and 之 is an alternative word choice to the synonymous 的. The stylistic difference between 将 vs. 把 and 之vs. 的 cannot be attributed to formality or literateness on the basis of distributional evidence shown in Figure 13.

"Alternative Diction" is even better motivated when we go beyond Chinese. In Section 5, a two-dimensional analysis will also be presented for English, which shows remarkable similarity to Chinese. But needless to say, the second dimension in English, also interpretable as "Alternative Diction," cannot be classical for obvious reasons.

Compared with Dimension 1, the second dimension is much weaker in that it only accounts for less than one-sixth of the total variation (two-thirds accounted for by the first dimension). This statistical information may well be indicative of the relative importance of the two dimensions.

### 4.1.3 Distribution of the 15 registers in LCMC

The distribution of the 15 registers of LCMC is given in Figure 14. The distribution pattern provides additional support for our interpretations of the two dimensions.

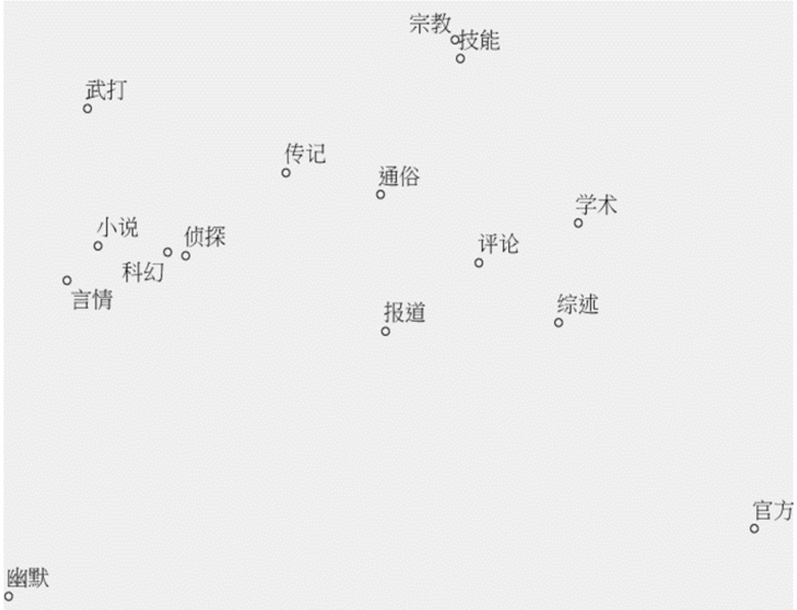


Figure 14. Distribution of *LCMC* registers

On Dimension 1, official documents (官方), academic writing (学术), and one of the three news types (综述) are the right most on this dimension, whereas all five types of fictional writing (小说、言情、科幻、侦探、武打) and humor (幽默) are on the opposite side, with news reports (报道), hobbies and skills (技能), religion (宗教), popular lore (通俗), and biography (传记) in between. This seems to concur with both our intuition and the distribution of features seen earlier.

On Dimension 2, the text types most alternative in diction are hobbies and skills, religion, and martial art novels. The least alternative are official documents and humor. In the middle are all types of fiction, all news types, and academic writing. While this may be initially surprising, it is in fact independently collaborated. Tao (1999) observed that in hobbies/skills texts, such as recipes, the classical 将 is more frequent than the synonymous modern 把, which is found more often in formal texts such as political commentaries. It is also fairly reasonable to accept that highly conventionalized official documents and academic writing are not as alternative in diction as martial arts novels.

Having two dimensions allows a text type to have different stylistic values on the two dimensions. For example, official documents rank the highest on Dimension 1 but almost the lowest on Dimension 2; official documents and humor, which are very close to each other on Dimension 2, actually occur at the opposite ends of Dimension 1. Finally, martial arts fiction, which is sandwiched between general and romantic fiction on Dimension 1, is actually quite far from the other fiction subtypes on Dimension 2. The addition of the second dimension allows us to avoid some of the quandaries that the single spoken vs. written distinction forces on us. For example, the separation of the two dimensions makes it possible to account for the fact that text types having more classical Chinese elements, such as hobbies and martial arts fiction, are not necessarily more literate, and vice versa.

### 4.2 Mapping near synonyms

One of the practical applications of Correspondence Analysis is clearer explication of near synonyms. Near synonyms are a major source of difficulty in learning Chinese. It is also difficult to describe their differences in a clear and objective manner. Reference works such as dictionaries can be vague and equivocal. Correspondence Analysis can provide a more fine-grained and empirically-based picture of the differences between them.

Presented in this section is a study based on Zhang (forthcoming). For this study, the same methodological procedure is used as in Section 4 above. Naturally, as the objective here is to show the differences between near synonyms, sets of near synonyms will need to be added to the feature set. Another difference is that the larger BCC corpus is used, which is more suitable for the study of lexical features due to its greater size.

One set of near synonyms, 妇女、女性、女子、女士 and 女人, will first be used for illustration. As shown in Figure 15, these five near synonyms for “women” are distributed along the horizontal dimension in a fairly astounding manner, from most literate to least: 妇女→女性→女子→女士→女人 (the orientation of the bi-plot is flipped from earlier bi-plots using the LCMC corpus). While the plot offers a more gradient picture, which introspection cannot hope to do, it seems in fact quite intuitive. Another striking fact is that 女子 differs from the other four in being more neutral on the literate dimension but stronger on the alternative diction dimension. While this is hard to come by from intuition alone, it too seems to agree with our intuition, as it is more commonly found in classically-inflected texts, such as martial arts novels or theatrical literature, in the classical style.

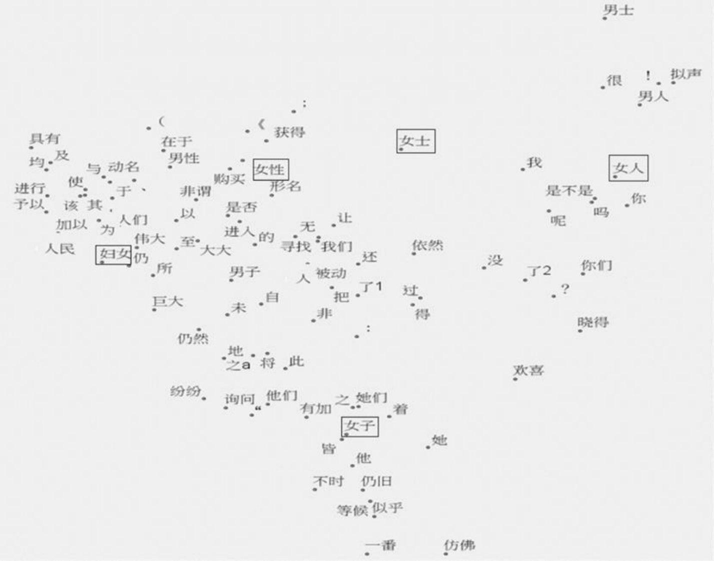


**Figure 15. Five near synonyms (left: + Literate; bottom: +Alternative Diction)**

The second set of synonyms for illustration is the homophonous pair 作 and 做 [do/make]. This pair has been the bane of Chinese language users, as they seem very similar and are sometimes used interchangeably. Although Lü (1980), Teng (1996), Wang (2005), Yang and Jia (2003), and G. Zhang (2010) all agree that the two are different in collocation, 作 being more abstract (better rendered as “doing”) but 做 more specific and concrete (better rendered as “making”), their judgements nevertheless differ. While Lü (1980) considers 作 to be more classical in flavor, G. Zhang (2010) considers both 作 and 做 to be neutral stylistically.

The use of corpus data and Correspondence Analysis helps resolve the difference in subjective judgement. The difference between the two shows up quite clearly on the bi-plot in Figure 16, as 作 lies at the more literate end of the horizontal dimension, consistent with Lü’s observation that it is more abstract than 做. As noted by Biber (1998), abstractness is also associated with literateness.

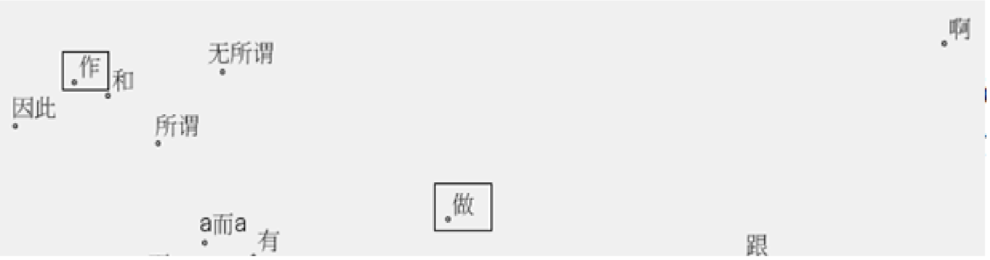


**Figure 16. Partial bi-plot contrasting 作 and 做**

## 5. An English Example

It will be shown in this section (based on Zhang, 2017) that the same methodology can be applied to English, too. Furthermore, the two dimensional analysis of Chinese gets cross-linguistic support.

For this pilot study, the large COCA corpus (The Corpus of Contemporary American English) was used (https://corpus.byu.edu/coca/). According to the developer of the corpus, Mark Davis of Brigham Young University: “The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English and the only large and balanced corpus of American English.” It includes 520 million words with 5 types of texts: Spoken, Fiction, Magazine, Newspaper, and Academic. For this study, 88 lexical and grammatical features were chosen based on their potential effect on stylistic variation.

Given the 5 types of texts, 4 dimensions (5 minus 1) can be automatically generated. But as in the case of Chinese, only the first two dimensions seem clearly interpretable. Striking similarities between Chinese and English are found.

### 5.1 Dimension 1: Literate

Figure 17 shows the distribution of the 88 features. Given the distribution patterns, it seems reasonable to assume that this primary, horizontal dimension can be interpreted the same way as in Chinese, i.e., as one of literateness. A number of literate features are found on the left (encircled). Passives, both the by-variety and the one without *by*, are left of center. Nominal suffixes, such as *-ity*, *-tion* and *-ness*, also lean toward the left. Also found here are literate and formal words and expressions like *upon*, *whom*, *thus*, *hitherto*, *demise*, *due to*, *of the opinion*, *e.g.*, *i.e.*, and *etc.* Parenthesis, hyphen and semi-colon, all associated with carefully crafted texts, also appear in this region. In contrast, the right side of the plot is populated by features such as personal pronouns, colloquial expressions (*absolutely*, *kind of*, *a couple of*), contracted forms, interjections, and punctuation marks such as quotation marks, colons, questions, and exclamations.

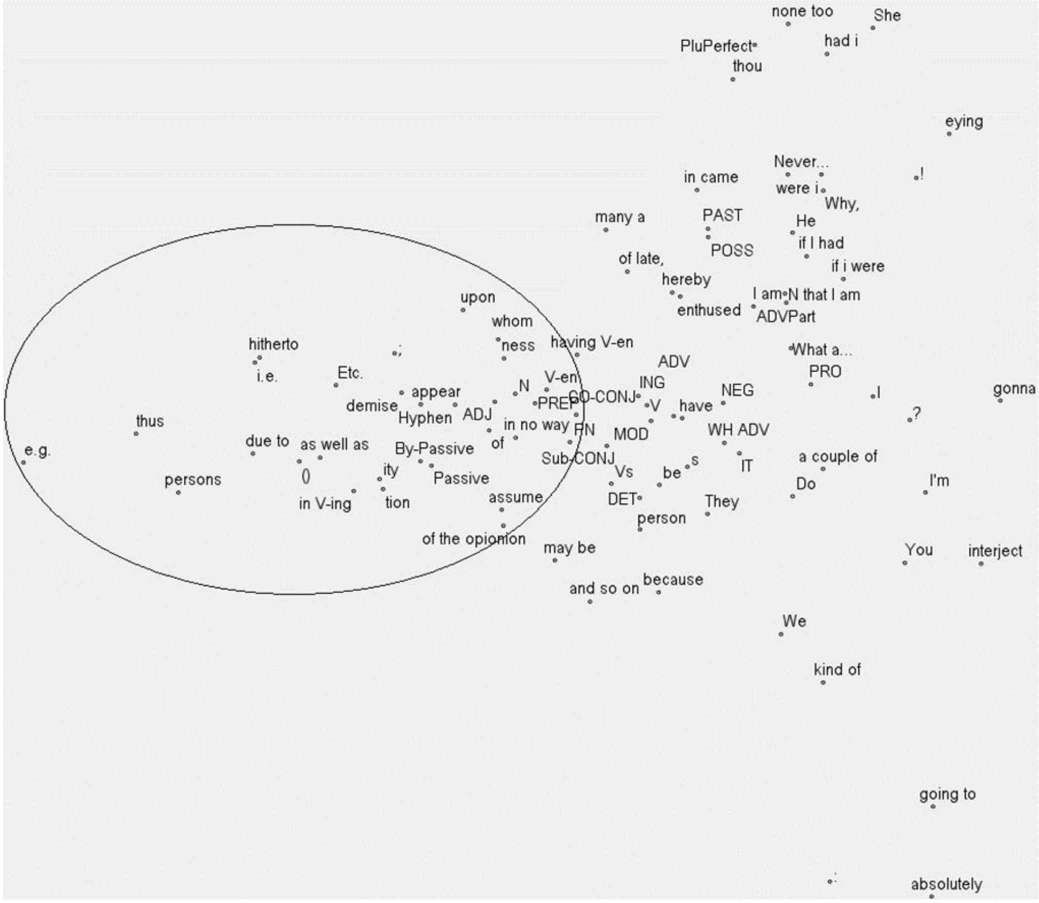


**Figure 17. Clustering of Literate features (encircled)**

There are some clear minimal contrasts between synonymous items such as *etc*. versus *and so on*, *I'm* versus *I am,* and *because* versus *due to*. Most noteworthy is the distinction between the singular *person* and the plural *persons*, located far apart on this dimension: while the singular is stylistically more neutral, the plural evokes the flavor of legalese, as in: "persons of known heart conditions should refrain from using the spa." The stylistically equivalent plural counterpart of *person* is thus not *persons* but more likely *people*.

## 5.2 Dimension 2: Alternative Diction

On the secondary, vertical dimension, there are some distributional facts that clearly support the interpretation of this dimension as "Alternative Diction." A number of words, expressions, and constructions are found at the top (encircled) of Figure 18. In addition to literary-sounding lexical items, such as *thou*, *hereby*, *enthuse*, *eying*, there are also constructions such as *none too+adj*. (e.g. *none too pleased about the prospects of meeting the family*), *of late*, *what a + noun* (e.g. *what a wonderful morning*), *many a + noun* (e.g. *many a thing you know you'd like to tell her*), *noun + that I am* (e.g. *fool that I am*). A feature that assumes extreme positive value on this dimension is the pluperfect construction, as used in *when I got there, he had already left*. While the simpler *when I got there, he already left* may now be more frequently used in speech, the pluperfect can still be found in literary texts.

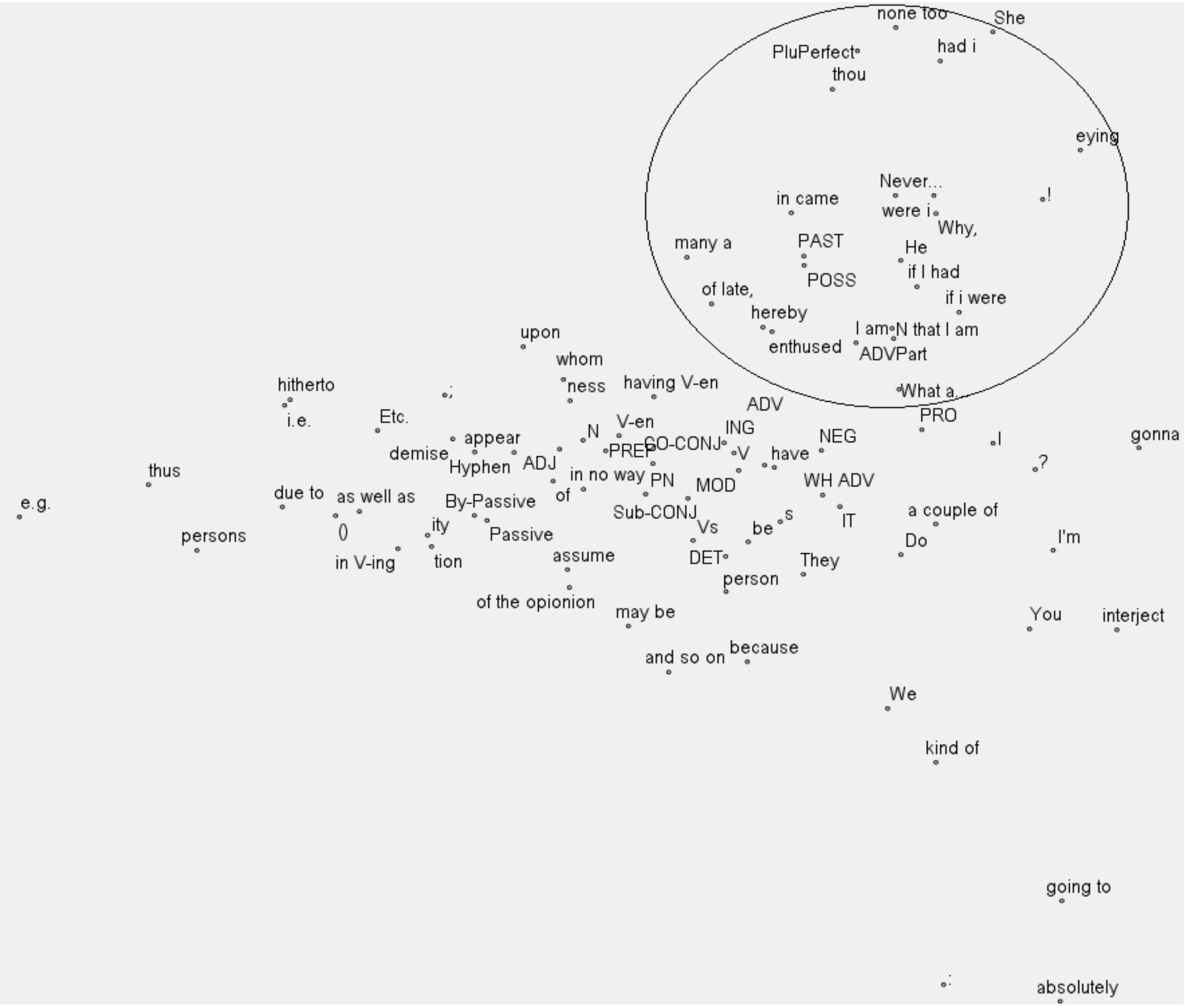


**Figure 18. Clustering of Alternative Diction features (encircled)**

Also found in this area is the non-canonical, inverted construction that is associated with literary usages (Green, 1982). They include *never + inverted clause (e.g. never have I been so insulted!)*, *were I* (alternative to *if I were*), *had I* (alternative to *if I had*), *in came* (alternative to *came in*).

The distribution of the 5 COCA text types, already seen in Figure 2 and repeated in Figure 19, provides clear support for the interpretations of the two dimensions: academic writing is the most literate while fiction is the most alternative in diction.

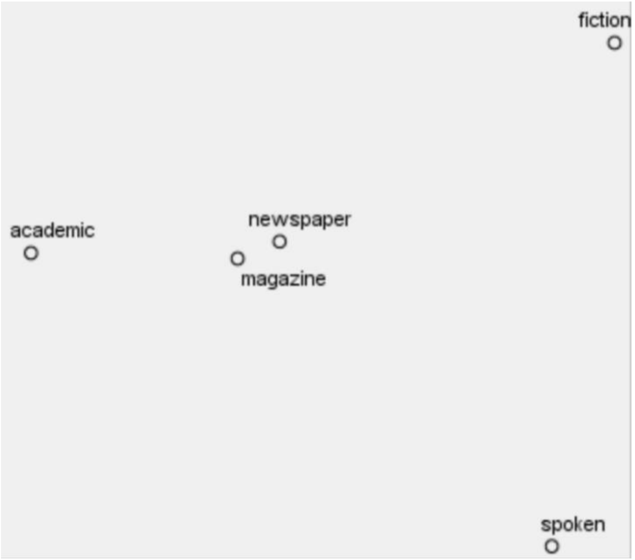


**Figure 19. Distribution of 5 *COCA* text types**

## 6. Theoretical and Practical Implications

As has been demonstrated, Correspondence Analysis is a more intuitive and easier to use alternative for carrying out multi-dimensional research on stylistic variation. But the wider use of this tool should have theoretical and practical significance beyond the obvious methodological advantages.

Theoretically, the easier application of Correspondence Analysis may encourage more researchers to conduct corpus-based multi-dimensional studies of stylistic variation. On stylistic matters, one no longer has to rely solely on introspection, which tends to be grossly simplistic. Not only can the multi-dimensional framework present a broader and more fine-grained picture of stylistic variation than simple dichotomies, it also allows for the possibility of gradient continuums, which binary distinctions do not.

Pedagogically, two-dimensional “stylistic maps” can obviously be helpful as well. Instead of relying solely on the admittedly useful heuristic notion of spoken vs. written, learners can see in a visually intuitive manner much finer shades of stylistic differences. As seen in section 4.2, near synonyms such as 妇女、女性、女子、女士 and 女人 can be shown to be clearly differentiated along not just one but two dimensions.

Another important application, especially in upper-level Chinese classes, is in the area of classical Chinese elements. A better understanding of the role of these elements in modern written Chinese will no doubt lead to more judicious instruction. We have shown that contrary to a common assumption, they are actually not exclusively associated with formal written texts. This has important pedagogical implications for issues concerning priority-setting, selection, and sequencing, which have not been sufficiently addressed. For example, how important are classical elements to overall proficiency in written Chinese? What kind of classical elements should be prioritized and introduced first? What kind of texts should be chosen that contain such priority elements? Corpus-based, multi-dimensional research using Correspondence Analysis should have a direct impact in this area.

## References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Pearson Education.

Chen, G. Y. (2016). Examining rating criteria used to assess U.S. college students' Chinese oral performance. *Journal of Chinese Language Teachers' Association*, 51(3), 286-311.

Feng, S. (2010). The mechanism of register and its grammatical properties. *Zhongguo Yuwen*, 5, 400-412. [冯胜利. (2010). 论语体的机制及其语法属性. *中国语文*, 5, 400-412.]

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Green, G. (1982). Colloquial and literary uses of inversions. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 119-154). Norwood, NJ: ABLEX Publishing Corporation.

Greenacre, M. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Gries, S. T. (2015). Quantitative designs and statistical techniques. In D. Biber, & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 50-71). Cambridge: Cambridge University Press.

Jang, S. C. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study* (unpublished doctoral dissertation). University of Hawaii, Hawaii.

Lü, S. (1980). *Eight hundred words in contemporary Chinese*. Beijing, China: Commercial Press. [吕叔湘. (1980). *现代汉语八百词*. 北京: 商务印书馆.]

Shen, H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System*, 33(1), 49-68.

Tabata, T. (2007). A statistical study of superlatives in Dickens and Smollett: A case study in corpus stylistics. *Digital Humanities 2007 Conference Abstracts* (pp. 1-5). Retrieved from https://pdfs.semanticscholar.org/23f7/e3a6ca295129bf18a4a3ca619038957a6c35.pdf

Tao, H. Y. (1999). On the grammatical significance of register distinctions. *Contemporary Linguistics*, 3, 15-24. [陶红印. (1999). 试论语体分类的语法学意义. *当代语言学*, 3, 15-24.]

Teng, S. H. (1996). *Chinese synonyms usage dictionary*. Beijing, China: Beijing Language Institute Press. [鄧守信. (1996). *汉英汉语常用近义词用法词典*. 北京: 北京语言学院出版社.]

Wang, H. (2005). *A dictionary of Chinese synonyms*. Beijing, China: Beijing Language and Culture University Press. [王还. (2005). *汉语近义词典*. 北京: 北京语言大学出版社.]

Yang, J., & Jia, Y. (2005). *1700 groups of frequently used Chinese synonyms*. Beijing, China: Beijing Language and Culture University Press. [杨寄洲, &贾永芬. (2005). *1700对近义词语用法对比*. 北京: 北京语言大学出版社.]

Zhang, G. Q. (2010). *Using Chinese synonyms*. Cambridge: Cambridge University Press.

Zhang, Z. S. (forthcoming). Visualizing stylistic differences in Chinese synonyms. In X. Lu, & B. Chen (Eds.). *Computational and corpus approaches to Chinese language learning*. Singapore: Springer.

Zhang, Z. S. (2017). *Dimensions of variation in written Chinese*. Oxford: Routledge.

Zhang, Z. S. (2016). A multi-dimensional corpus study of mixed compounds in Chinese. In H. Tao (Ed). *Integrating Chinese linguistic research and language teaching and learning* (pp. 215-238). Amsterdam, Holland: John Benjamins Publishing Company.

Zhang, Z. S. (2013). The classical elements in written Chinese: A multidimensional quantitative study. *Chinese Language and Discourse*, *4*(2), 157-180.

Zhang, Z. S. (2012). A corpus study of variation in written Chinese. In E. Csomay (Ed.), *Corpus linguistics and linguistic theory: Contemporary perspectives on discourse and corpora* 8–1 (2012) (pp. 209-240). Berlin, Germany: Walter de Gruyter.