# Multi-dimensional Analysis of Linguistic Features in Chinese Writing of Japanese students and Native Chinese Speakers

Qin Xu

Graduate School of Humanities
Osaka University

# Outline

1. Introduction
2. Methods
3. Factor Analysis
4. Results & Discussion

# Multi-dimensional Analysis (MDA)

**Biber (1988, 2006, 2014, *etc.*):**

- **MF/MD approach**: Multi-features / Multi-dimensional approach
- **Primary statistical tool**: **Factor analysis**

→ reduce a large number of **linguistic features** to a few **dimensions**

**explain**

**textual variation** & **linguistic variation**

# Examples of Multi-dimensional Analysis

| | | Data , Method | Results |
|---|---|---|---|
| English | Biber (1988) | 23 registers (English)<br>481 spoken and written texts<br>67 linguistic features<br>Principal factor analysis<br>→Promax rotation | Dimension 1: Involved versus Informational Production<br>Dimension 2: Narrative versus Non-narrative Concerns<br>Dimension 3: Explicit versus Situation-Dependent Reference<br>Dimension 4: Overt Expression of Persuasion<br>Dimension 5: Abstract versus Non-Abstract Information<br>Dimension 6: On-Line Informational Elaboration |
| | Biber (2006) | 10 registers (English)<br>423 spoken and written texts<br>129 linguistic features (90)<br>Principal factor analysis<br>→Promax rotation | Dimension 1: Oral vs. literate discourse<br>Dimension 2: Procedura1 vs. content-focused discourse<br>Dimension 3: Reconstructed account of events<br>Dimension 4: Teacher-centered stance |
| | Friginal & Weigle (2014) | 207 essays (L2 students)<br>72 lexico-grammatical features<br>Exploratory Factor Analysis<br>→Promax rotation | Dimension 1: Involved vs. Informational Focus<br>Dimension 2: Addressee-Focused Description vs. PersonalNarrative<br>Dimension 3: Simplified vs. Elaborated Description<br>Dimension 4: Personal Opinion vs. Impersonal Evaluation/Assessment |
| Chinese | Zhang, Z (2012) | 15 written registers (Mandarin Chinese)<br>500 random samples<br>60 linguistic features<br>Correspondence analysis | Dimension 1： Literate<br>Dimension 2： Classical<br>Dimension 3： News commentary |
| | Zhu (2015) | 16 registers (Mandarin Chinese)<br>1000 spoken and written texts<br>88 linguistic features<br>Principal factor analysis<br>→Promax rotation | Dimension 1： Interactive vs. Informational Discourse<br>Dimension 2： Literary vs. Non-literary Concern<br>Dimension 3： Colloquialized Expression with Subjective Emphases<br>Dimension 4： Situation-dependent Reference & Emotional Concern<br>Dimension 5： Persuasion and Argumentation vs. Non-persuasive and Non-argumentative Concern |

**L2 English writing**

# Text Selection

**L2 Chinese Writing** (Japanese students)

- **5 genres:** character description, narrative essay, argumentative essay, letter, diary

→ collected between 2020 and 2022

**L1 Chinese Writing** (Chinese students)

- **5 genres:** same as L2

→ collected from Chinese writing website of Chinese students

zuowen 作文网 (http://www.zuowen.com/)

# Distribution of Text Genres

| Genres of L2 Chinese writing | JP2 | JP3 | Total |
|---|---|---|---|
| 1. Character description | 183 | 111 | |
| 2. Narrative essay | 43 | 61 | |
| 3. Argumentative essay | 0 | 57 | **700** |
| 4. Letter | 82 | 0 | |
| 5. Diary | 163 | 0 | |

| Genres of L1 Chinese writing | | CHN | Total |
|---|---|---|---|
| 1. Character description | | 150 | |
| 2. Narrative essay | | 150 | |
| 3. Argumentative essay | | 150 | **750** |
| 4. Letter | | 150 | |
| 5. Diary | | 150 | |

**JP2: Intermediate level**

**JP3: Advanced level**

# Corpus Construction

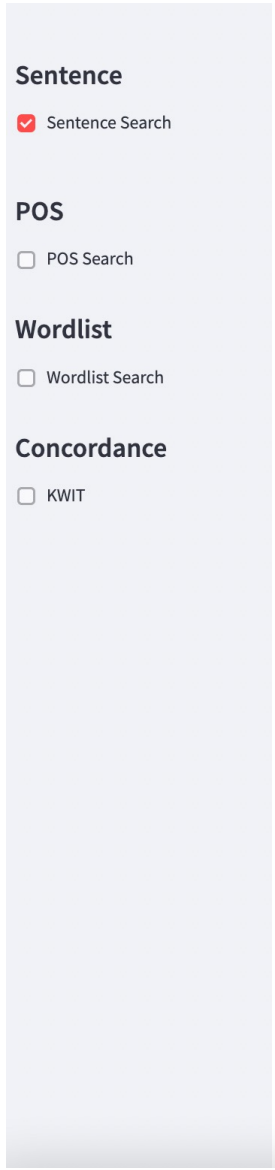
| | |
|---|---|
| **Data** | • **L1 & L2 Chinese Writing** |
| **Tools** | • **NLPLR**<br>Chinese word segmentation & part-of-speech tagging<br>• **Python & Streamlit**<br>create a website corpus through coding<br>• **GitHub & Heroku**<br>deploy and publish corpus app |
| **Functions** | • **Sentence search**<br>• **Part-of-speech search**<br>• **Wordlist search** |

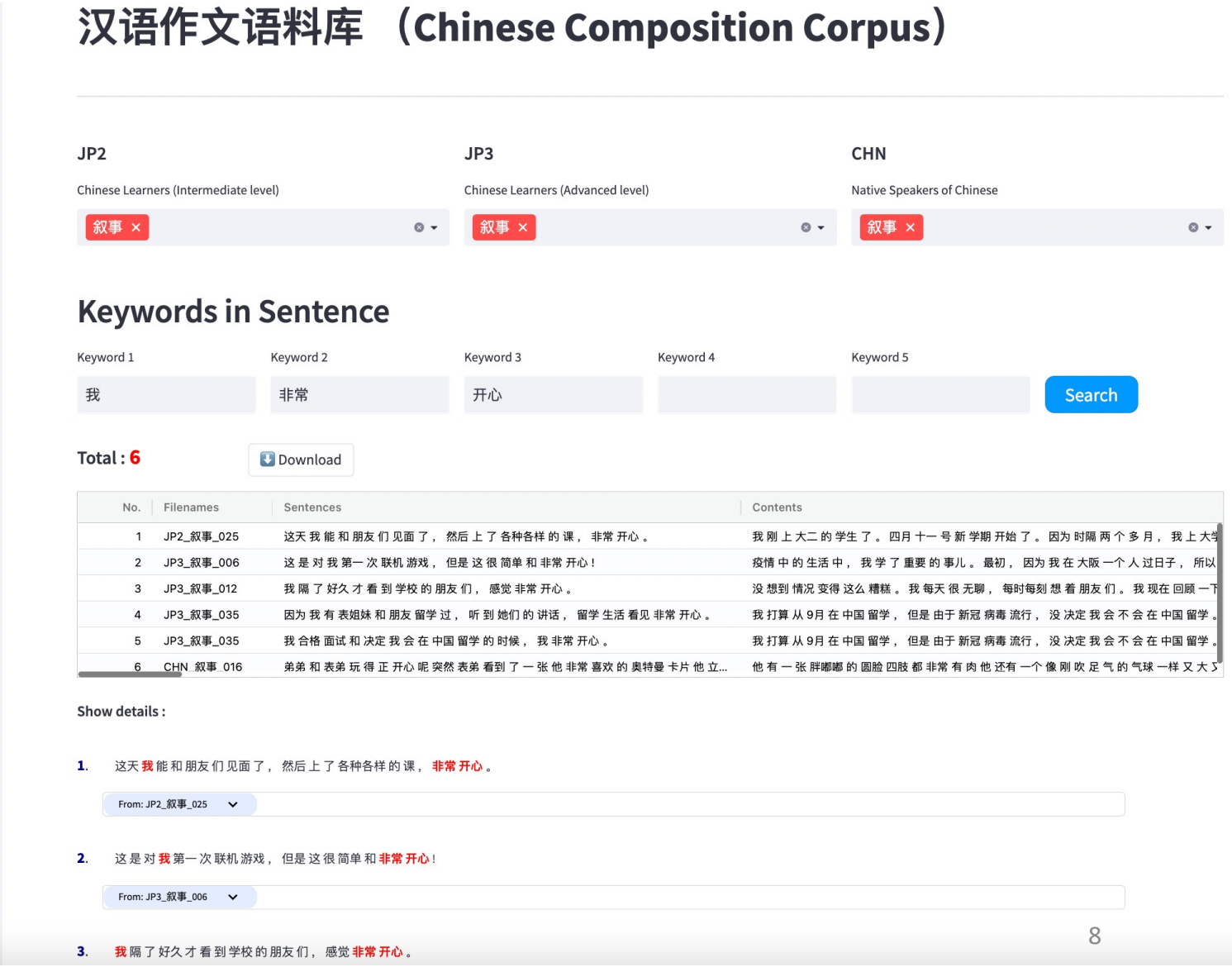
# Corpus functions

## Sentence Search

**Step:**

1. ☑ Sentence search
2. Select genres
   e.g. 叙事(Narrative writing)
3. Input keyword
4. Click "**Search**"



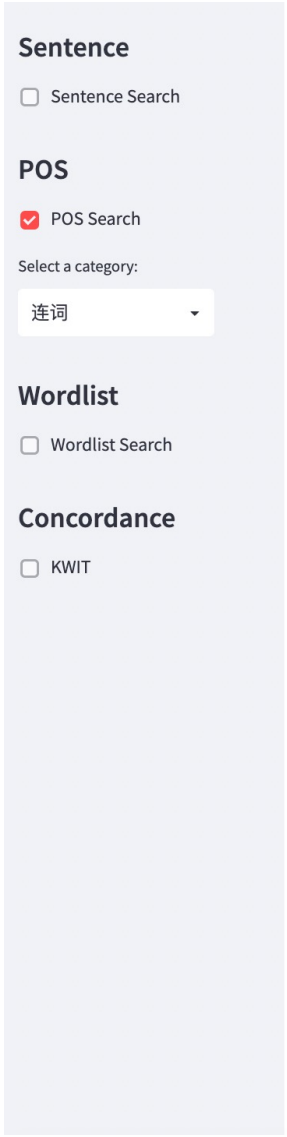



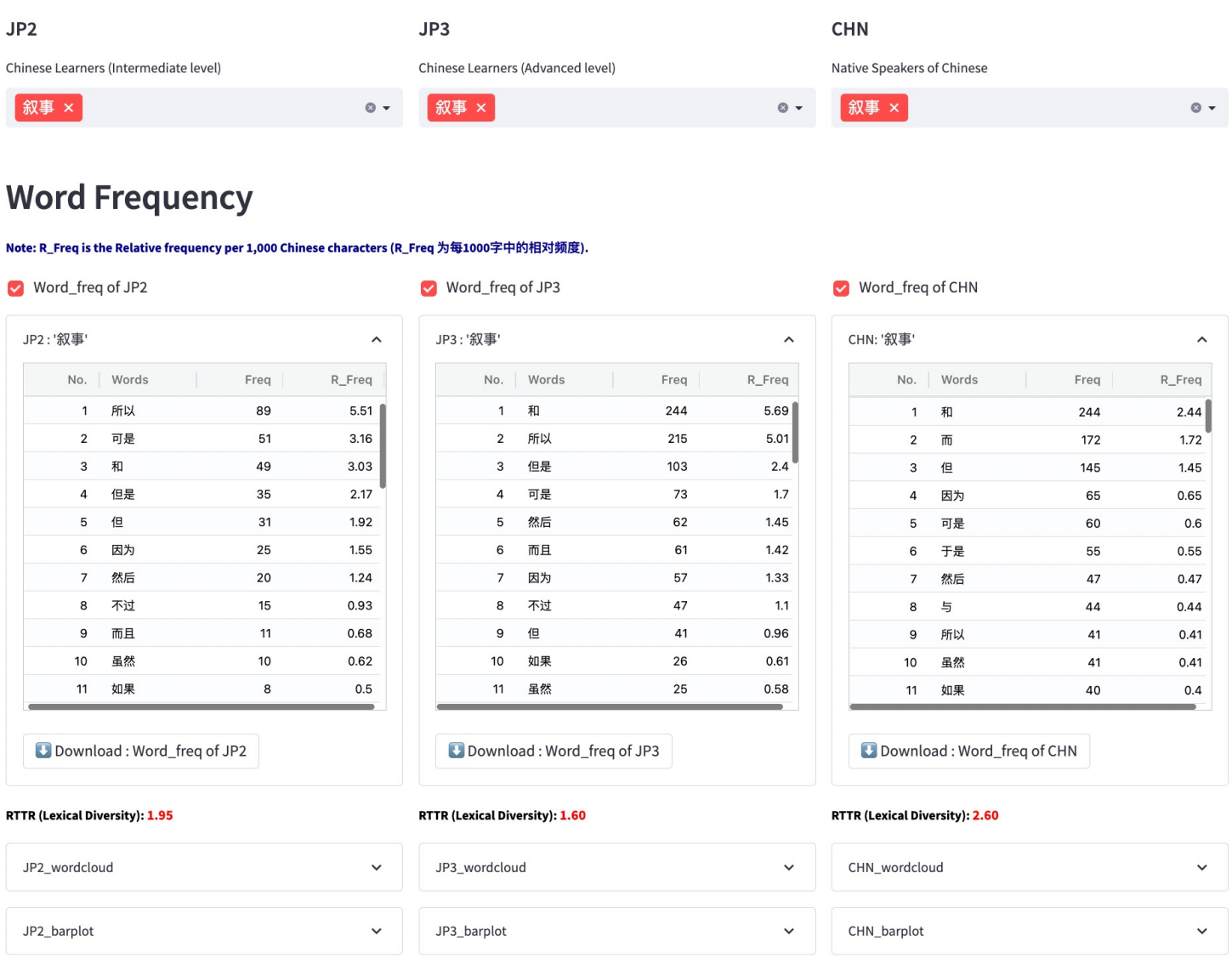
# Corpus functions

## Part-of-speech Search

**Step:**

1. ☑ POS search
2. Select a category
   e.g. 连词(conjunction)
3. Select genres
4. ☑ Word_freq of XX



# Corpus functions

## Part-of-speech search

**Step:**

1. ☑ POS search
2. Select a category
   e.g. 连词(conjunction)
3. Select genres
4. ☑ Word_freq of XX
5. Show wordcloud

Sentence

☐ Sentence Search

POS

☑ POS Search

Select a category:

连词

Wordlist

☐ Wordlist Search

Concordance

☐ KWIT

## Word Frequency

**Note: R_Freq is the Relative frequency per 1,000 Chinese characters (R_Freq 为每1000字中的相对频度).**

☑ Word_freq of JP2

JP2 : '叙事'

| No. | Words | Freq | R_Freq |
|---|---|---|---|
| 1 | 所以 | 89 | 5.51 |
| 2 | 可是 | 51 | 3.16 |
| 3 | 和 | 49 | 3.03 |
| 4 | 但是 | 35 | 2.17 |
| 5 | 但 | 31 | 1.92 |
| 6 | 因为 | 25 | 1.55 |
| 7 | 然后 | 20 | 1.24 |
| 8 | 不过 | 15 | 0.93 |
| 9 | 而且 | 11 | 0.68 |
| 10 | 虽然 | 10 | 0.62 |
| 11 | 如果 | 8 | 0.5 |

Download : Word_freq of JP2

☑ Word_freq of JP3

JP3 : '叙事'

| No. | Words | Freq | R_Freq |
|---|---|---|---|
| 1 | 和 | 244 | 5.69 |
| 2 | 所以 | 215 | 5.01 |
| 3 | 但是 | 103 | 2.4 |
| 4 | 可是 | 73 | 1.7 |
| 5 | 然后 | 62 | 1.45 |
| 6 | 而且 | 61 | 1.42 |
| 7 | 因为 | 57 | 1.33 |
| 8 | 不过 | 47 | 1.1 |
| 9 | 但 | 41 | 0.96 |
| 10 | 如果 | 26 | 0.61 |
| 11 | 虽然 | 25 | 0.58 |

Download : Word_freq of JP3

☑ Word_freq of CHN

CHN: '叙事'

| No. | Words | Freq | R_Freq |
|---|---|---|---|
| 1 | 和 | 244 | 2.44 |
| 2 | 而 | 172 | 1.72 |
| 3 | 但 | 145 | 1.45 |
| 4 | 因为 | 65 | 0.65 |
| 5 | 可是 | 60 | 0.6 |
| 6 | 于是 | 55 | 0.55 |
| 7 | 然后 | 47 | 0.47 |
| 8 | 与 | 44 | 0.44 |
| 9 | 所以 | 41 | 0.41 |
| 10 | 虽然 | 41 | 0.41 |
| 11 | 如果 | 40 | 0.4 |

Download : Word_freq of CHN

**RTTR (Lexical Diversity): 1.95**

JP2_wordcloud


**RTTR (Lexical Diversity): 1.60**

JP3_wordcloud


**RTTR (Lexical Diversity): 2.60**

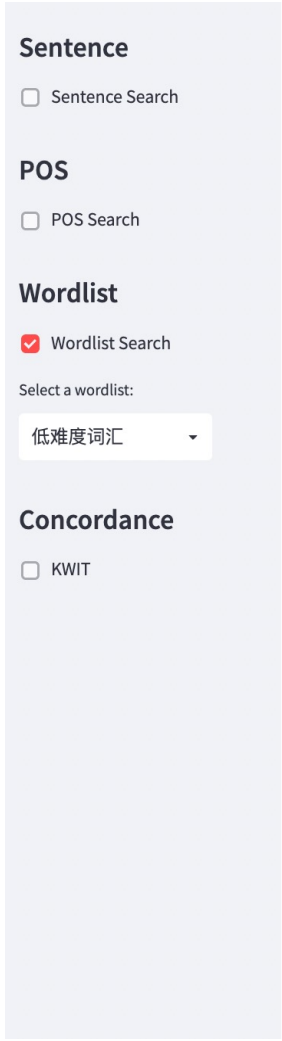
CHN_wordcloud

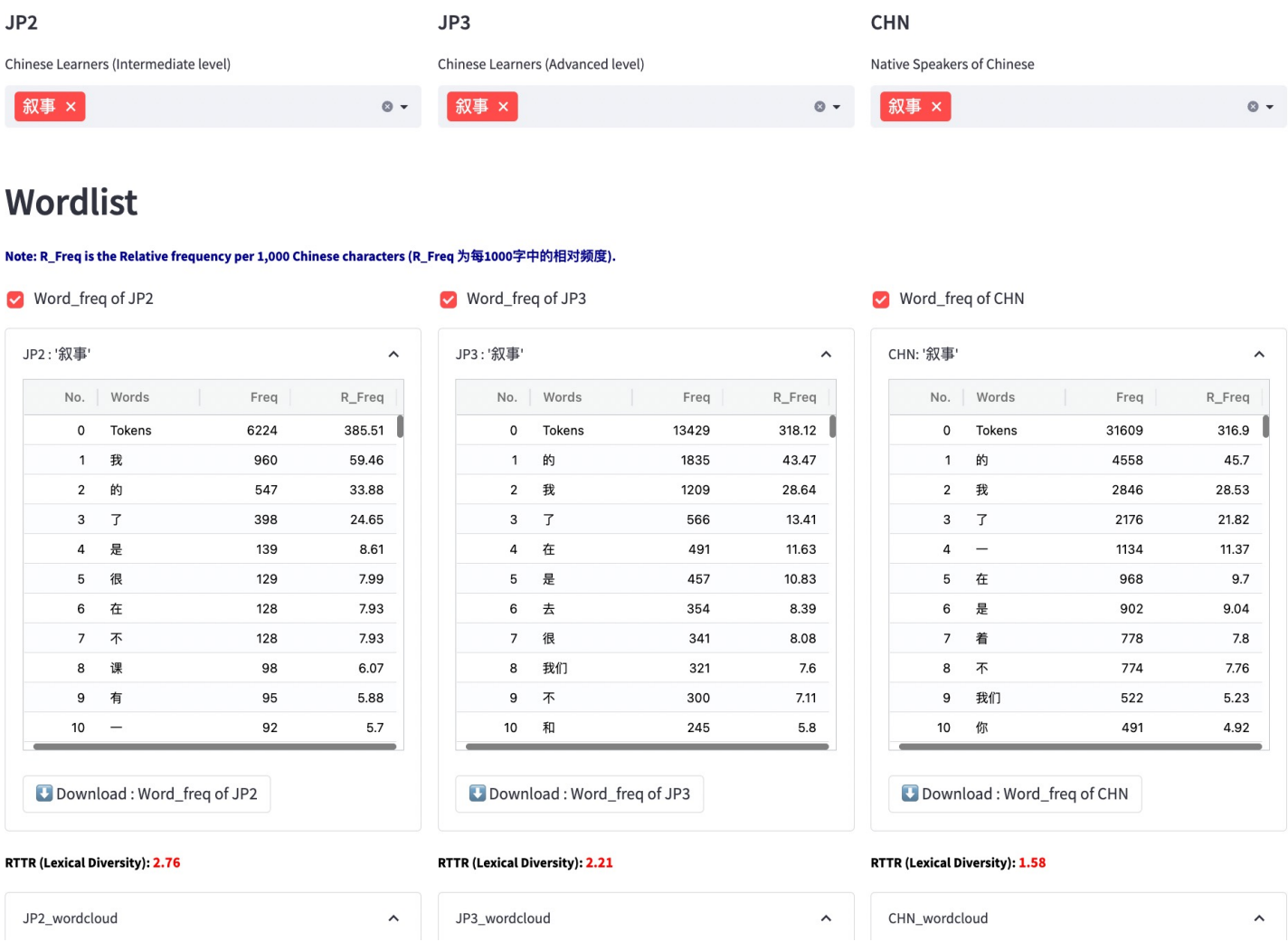
# Corpus functions

## Wordlist search

### Step:

1. ☑ Wordlist search
2. Select a wordlist
   e.g. 低难度词汇
   (Low-difficulty vocabulary)
3. Select genres
4. ☑ Word_freq of XX



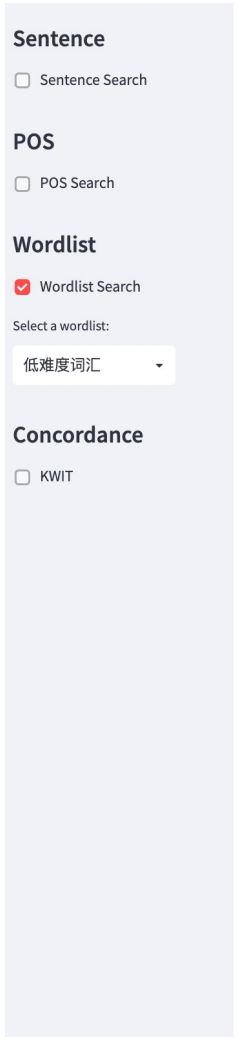



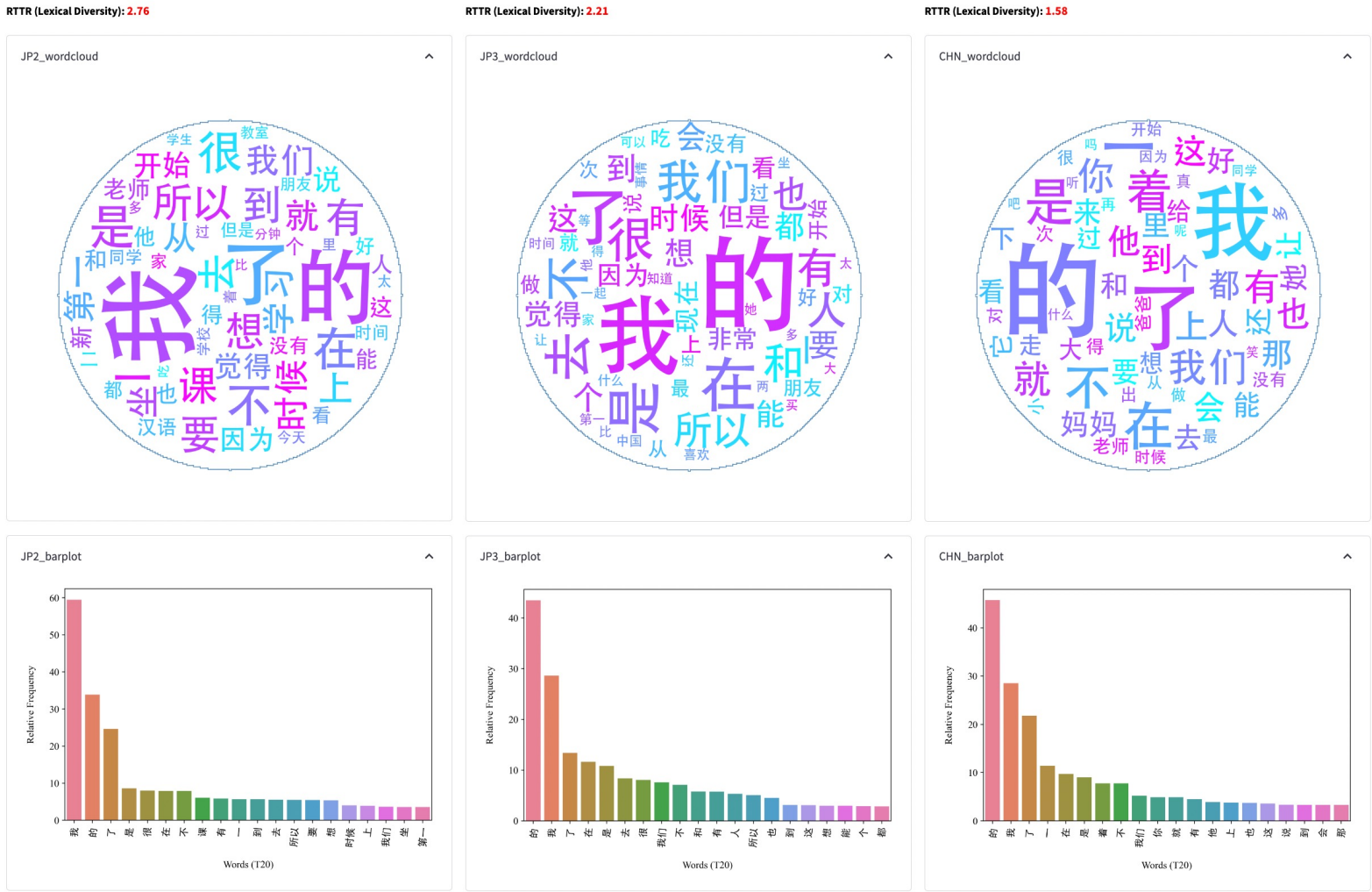
# Corpus functions

## Wordlist search

### Step:

1. ☑ Wordlist search
2. Select a wordlist
   e.g. 低难度词汇
   (Low-difficulty vocabulary)
3. Select genres
4. ☑ Word_freq of XX
5. **Show wordcloud**
6. **Show Barplot**



# Selection of Linguistic Features

'Prior to any comparison of texts, a principled decision must be made concerning the linguistic features to be used' (Biber 1988:71).

## 92 Linguistic Features

- **Previous studies on multidimensional analysis**

  Biber(1988, 2006); Zhang (2012); Liu(2019), etc.

- **Publications and research related to Chinese studies**

  Huang & Liao(2017); Feng(2000) etc.

- **L2 Chinese Vocabulary (HSK vocabulary list)**

# Part of the **92 Linguistic Features**

| ID | Linguistic Features |
|---|---|
| 1 | Noun: most commonly used |
| 2 | Noun: moderately commonly used |
| 3 | Noun: rarely used |
| 4 | Abstract noun |
| 5 | Concrete noun |
| | …… |
| 11 | Verb: most commonly used |
| 12 | Verb: moderately commonly used |
| 13 | Verb: rarely used |
| | …… |
| 23 | Adjective: most commonly used |
| 24 | Adjective: moderately commonly used |
| 25 | Adjective: rarely used |
| | …… |
| 29 | Adverb: most commonly used |
| 30 | Adverb: moderately commonly used |

| ID | Linguistic Features |
|---|---|
| 31 | Adverb: rarely used |
| | …… |
| 36 | Adverb of time |
| 37 | Adverb of degree |
| | …… |
| 40 | Less frequently used first person pronoun |
| 38 | First person pronoun: I / “我” |
| 39 | First person pronoun: we / “我们” |
| 40 | Less frequently used first person pronoun |
| 41 | Second person pronoun |
| 42 | Third person pronoun |
| 52 | Possessive affix: de / “的” |
| 53 | Adverbializer: di / “地” |
| 54 | Resultative complementizer: de/ “得” |
| 55 | Durative aspect: zhe / “着” |
| 56 | Past aspect: le/ “了” |

| ID | Linguistic Features |
|---|---|
| 57 | Experiential aspect: guo / “过” |
| | …… |
| 70 | Low-level vocabulary |
| 71 | Medium-level vocabulary |
| 72 | High-level vocabulary |
| 73 | Non-HSK vocabulary |
| | …… |
| 79 | Parallel compound sentence |
| 80 | Successive compound sentence |
| | …… |
| 87 | Purpose compound sentence |
| 88 | Turning compound sentence |
| 89 | Lexical diversity |
| 90 | Lexical density |
| 91 | Average word length |
| 92 | Average sentence length |

# Frequency Counts of Linguistic Features

## Normalized frequency:

**frequency of per 1,000 Chinese characters**

except for

| ID | Linguistic Features | |
|---|---|---|
| 89 | Lexical diversity | (types / √tokens) |
| 90 | Lexical density | (content words / tokens) |
| 91 | Average word length | (characters / tokens) |
| 92 | Average sentence length | (characters / sentences) |

# Counting Frequencies

# Python Programming

Frequency Count

- [x] Raw Frequency
- [x] Relative Frequency

## Frequency Count of 92 Linguistic Features

### Upload text file

**The text file should be in UTF-8 format and has been annotated by the NLPIR system. (请上传由NLPIR汉语分词系统进行分词后的TXT文本,TXT格式需为UTF-8。)**

Choose txt files

Drag and drop files here
Limit 200MB per file • TXT

Browse files

file_10.txt 3.1KB
file_09.txt 3.7KB
file_08.txt 4.2KB

Showing page 1 of 4

Note: Before uploading, please click here for a text example. The text to be uploaded should be in "UTF-8" format (上传前请点击此处参考TXT分词文本示例。将分词文本保存为编码格式为"UTF-8"的txt文件。)

### Raw Frequency of 92 Linguistic Features

| | 01.高频名词 | 02.中频名词 | 03.低频名词 | 04.抽象名词 | 05.具象名词 | 06.心理名词 | 07.指人名词 | 08.集体名词 | 09.普通名词复数型 | 10.名词化功能词 | 11.高频动词 | 12.中频动词 | 13.低频动词 | 14.动作行为动词 | 15.心理动词 | 16.肯定性动词 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 58.0000 | 11.0000 | 4.0000 | 8.0000 | 63.0000 | 0.0000 | 38.0000 | 2.0000 | 0.0000 | 4.0000 | 84.0000 | 13.0000 | 0.0000 | 51.0000 | 3.0000 | 0.0000 |
| 1 | 58.0000 | 14.0000 | 2.0000 | 7.0000 | 64.0000 | 0.0000 | 31.0000 | 1.0000 | 0.0000 | 2.0000 | 88.0000 | 12.0000 | 2.0000 | 53.0000 | 5.0000 | 1.0000 |
| 2 | 61.0000 | 9.0000 | 7.0000 | 12.0000 | 55.0000 | 2.0000 | 32.0000 | 1.0000 | 0.0000 | 6.0000 | 66.0000 | 14.0000 | 5.0000 | 44.0000 | 7.0000 | 1.0000 |
| 3 | 65.0000 | 15.0000 | 6.0000 | 13.0000 | 76.0000 | 1.0000 | 27.0000 | 0.0000 | 0.0000 | 8.0000 | 72.0000 | 27.0000 | 4.0000 | 60.0000 | 11.0000 | 4.0000 |
| 4 | 43.0000 | 11.0000 | 5.0000 | 5.0000 | 54.0000 | 0.0000 | 19.0000 | 0.0000 | 0.0000 | 2.0000 | 89.0000 | 15.0000 | 2.0000 | 60.0000 | 11.0000 | 0.0000 |
| 5 | 75.0000 | 13.0000 | 3.0000 | 12.0000 | 74.0000 | 1.0000 | 26.0000 | 1.0000 | 0.0000 | 3.0000 | 81.0000 | 13.0000 | 1.0000 | 61.0000 | 3.0000 | 1.0000 |
| 6 | 50.0000 | 11.0000 | 2.0000 | 5.0000 | 53.0000 | 0.0000 | 24.0000 | 1.0000 | 0.0000 | 1.0000 | 54.0000 | 17.0000 | 3.0000 | 39.0000 | 9.0000 | 1.0000 |
| 7 | 66.0000 | 13.0000 | 0.0000 | 13.0000 | 63.0000 | 1.0000 | 33.0000 | 3.0000 | 0.0000 | 3.0000 | 72.0000 | 22.0000 | 4.0000 | 45.0000 | 13.0000 | 2.0000 |
| 8 | 48.0000 | 16.0000 | 4.0000 | 7.0000 | 65.0000 | 0.0000 | 25.0000 | 0.0000 | 0.0000 | 3.0000 | 78.0000 | 24.0000 | 2.0000 | 51.0000 | 8.0000 | 1.0000 |
| 9 | 30.0000 | 11.0000 | 2.0000 | 12.0000 | 25.0000 | 1.0000 | 6.0000 | 0.0000 | 2.0000 | 5.0000 | 66.0000 | 13.0000 | 1.0000 | 39.0000 | 8.0000 | 1.0000 |

⬇️Download : Raw Frequency of 92 Linguistic Features

### Relative Frequency of 92 Linguistic Features

Relative Frequency is the relative frequency per 1,000 Chinese characters (标准化频率为每1000字中的相对频率).

| | Filenames | 01.高频名词 | 02.中频名词 | 03.低频名词 | 04.抽象名词 | 05.具象名词 | 06.心理名词 | 07.指人名词 | 08.集体名词 | 09.普通名词复数型 | 10.名词化功能词 | 11.高频动词 | 12.中频动词 | 13.低频动词 | 14.动作行为动词 | 15.心理动词 | 16.肯 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | file_01.txt | 106.2271 | 20.1465 | 7.3260 | 14.6520 | 115.3846 | 0.0000 | 69.5971 | 3.6630 | 0.0000 | 7.3260 | 153.8462 | 23.8095 | 0.0000 | 93.4066 | 5.4945 | |
| 1 | file_02.txt | 113.9489 | 27.5049 | 3.9293 | 13.7525 | 125.7367 | 0.0000 | 60.9037 | 1.9646 | 0.0000 | 3.9293 | 172.8880 | 23.5756 | 3.9293 | 104.1257 | 9.8232 | |
| 2 | file_03.txt | 98.2287 | 14.4928 | 11.2721 | 19.3237 | 88.5668 | 3.2206 | 51.5298 | 1.6103 | 0.0000 | 9.6618 | 106.2802 | 22.5443 | 8.0515 | 70.8535 | 11.2721 | |
| 3 | file_04.txt | 104.0000 | 24.0000 | 9.6000 | 20.8000 | 121.6000 | 1.6000 | 43.2000 | 0.0000 | 0.0000 | 12.8000 | 115.2000 | 43.2000 | 6.4000 | 96.0000 | 17.6000 | |
| 4 | file_05.txt | 80.8271 | 20.6767 | 9.3985 | 9.3985 | 101.5038 | 0.0000 | 35.7143 | 0.0000 | 0.0000 | 3.7594 | 167.2932 | 28.1955 | 3.7594 | 112.7820 | 20.6767 | |
| 5 | file_06.txt | 123.7624 | 21.4521 | 4.9505 | 19.8020 | 122.1122 | 1.6502 | 42.9043 | 1.6502 | 0.0000 | 4.9505 | 133.6634 | 21.4521 | 1.6502 | 100.6601 | 4.9505 | |
| 6 | file_07.txt | 89.7666 | 19.7487 | 3.5907 | 8.9767 | 95.1526 | 0.0000 | 43.0880 | 1.7953 | 0.0000 | 1.7953 | 96.9479 | 30.5206 | 5.3860 | 70.0180 | 16.1580 | |
| 7 | file_08.txt | 106.9692 | 21.0697 | 0.0000 | 21.0697 | 102.1070 | 1.6207 | 53.4846 | 4.8622 | 0.0000 | 4.8622 | 116.6937 | 35.6564 | 6.4830 | 72.9335 | 21.0697 | |
| 8 | file_09.txt | 82.0513 | 27.3504 | 6.8376 | 11.9658 | 111.1111 | 0.0000 | 42.7350 | 0.0000 | 0.0000 | 5.1282 | 133.3333 | 41.0256 | 3.4188 | 87.1795 | 13.6752 | |
| 9 | file_10.txt | 67.5676 | 24.7748 | 4.5045 | 27.0270 | 56.3063 | 2.2523 | 13.5135 | 0.0000 | 4.5045 | 11.2613 | 148.6486 | 29.2793 | 2.2523 | 87.8378 | 18.0180 | |

⬇️Download : Relative Frequency of 92 Linguistic Features

# Data pre-processing

**Data**： 1450 texts & 92 Linguistic Features. → 1450 rows × 92 columns

**Tools**： Factor_analyzer of Python

## Data pre-processing

- Remove **NaN** columns after converting to Z-score (result of dividing by 0)
  → 1450 rows × 84 columns
- Remove the rows containing **outliers** (|Z-score| ≥ 5) in each group
  → 1314 rows × 84 columns

```
Data = df_Data[df_Data.groupby('Subcorpus').
               apply(lambda x: np.abs(x-x.mean())/x.std() < 5).all(axis=1)]

print(f'Total :  {Data.shape[0]} rows × {Data.shape[1]} columns')
```

```
Total :  1314 rows × 84 columns
```

# KMO and Bartlett's Test

```
# KMO and Bartlett's Test
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all, kmo_model = calculate_kmo(Data)

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, p_value = calculate_bartlett_sphericity(Data)

print("     KMO :",round(kmo_model, 3), "\n p_value :", p_value,
      "\nchi_square :",round(chi_square_value, 3))
```

```
     KMO : 0.512
 p_value : 0.0
chi_square : 54520.289
```

**Kaiser(1974):**

0.90 = marvelous
0.80 = meritorious
0.70 = middling
0.60 = mediocre
0.50 = miserable
below 0.50 = unacceptable

**Biber (2006: 182-183) :**

1) Some features were dropped because they overlapped to a large extent with other features.
2) Features were dropped because they were extremely rare.
3) Some features were dropped because they shared little variance with the overall factorial structure.

**→ Remove variables with low communalities.**

# Remove variables with low communalities

```
fa = FactorAnalyzer(rotation='promax',method='minres')
fa.fit(Data)

communalities = pd.DataFrame(fa.get_communalities(), index=list(Data.columns))
features_comm = list(communalities[communalities[0] > 0.30].index)

new_Data = Data[features_comm]
```

# KMO and Bartlett's test again

```
# KMO and Bartlett's Test
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all, kmo_model = calculate_kmo(new_Data)

from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, p_value = calculate_bartlett_sphericity(new_Data)

print("     KMO :", round(kmo_model, 3), "\n p_value :", p_value,
      "\nchi_square :", round(chi_square_value, 3))
```

```
     KMO : 0.771
 p_value : 0.0
chi_square : 13685.163
```

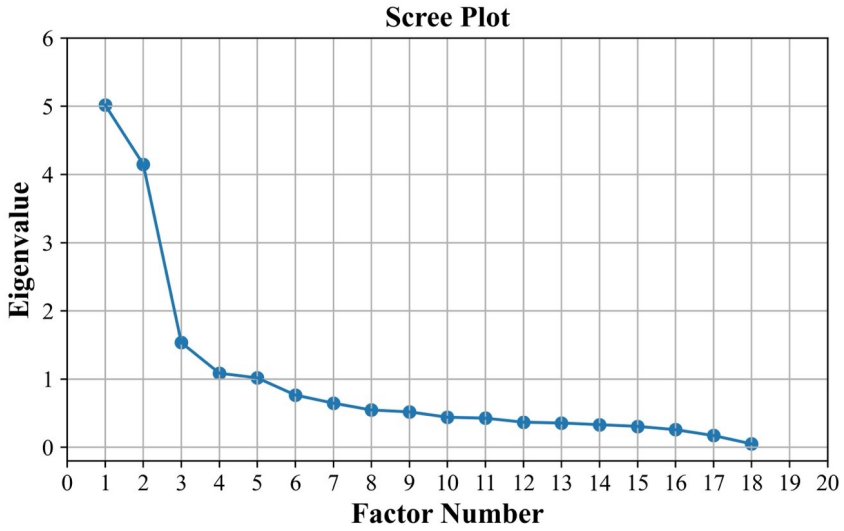
# Determine the number of factors

```
# Check Eigenvalues
EigenValue, value = fa.get_eigenvalues()

# Highlight the values if they are greater than 1.
def highlightEigenvalue(x):
    return ['background-color: yellow' if v > 1 else '' for v in x]

df_eigen = pd.DataFrame({'Factor': range(1, len(EigenValue) + 1), 'Eigenvalue': EigenValue})
df_eigen.style.apply(highlightEigenvalue, subset = ['Eigenvalue'])
```

| | Factor | Eigenvalue |
|---|---|---|
| 0 | 1 | 5.016026 |
| 1 | 2 | 4.147064 |
| 2 | 3 | 1.535661 |
| 3 | 4 | 1.087326 |
| 4 | 5 | 1.017840 |
| 5 | 6 | 0.766440 |
| 6 | 7 | 0.647815 |
| 7 | 8 | 0.547122 |
| 8 | 9 | 0.519536 |
| 9 | 10 | 0.441441 |
| 10 | 11 | 0.426966 |
| 11 | 12 | 0.368614 |
| 12 | 13 | 0.356418 |
| 13 | 14 | 0.332324 |
| 14 | 15 | 0.307517 |
| 15 | 16 | 0.258691 |
| 16 | 17 | 0.172528 |
| 17 | 18 | 0.050672 |




→ optimal:
4 – factor solution

# Factorial Structure

```
# Factor analysis with 'promax' rotation and 'minres' method.
factor_number = 4
fa = FactorAnalyzer(n_factors = factor_number, rotation = 'promax', method = 'minres')
fa.fit(new_Data)

# Loading factors
fac_loadings = pd.DataFrame(fa.loadings_,
                            columns = ['FAC{}'.format(i) for i in range(1, factor_number+1)],
                            index = new_Data.columns)

# Highlight the values if they are greater than 0.3.
def highlightLoadings(x):
    return ['background-color: yellow' if abs(v) > 0.3 else '' for v in x]

fac_loading_matrix=fac_loadings.style.apply(highlightLoadings)
fac_loading_matrix
```

# Cumulative Variance

```
# Explained variance
idx = ['SS Loadings', 'Proportion Variance', 'Cumulative Variance']
df_variance = pd.DataFrame(data = fa.get_factor_variance(),
                           index = idx,
                           columns = ['FAC{}'.format(i)
                                      for i in range(1, factor_number+1)])
df_variance
```

| | FAC1 | FAC2 | FAC3 | FAC4 |
|---|---|---|---|---|
| SS Loadings | 3.614584 | 3.095772 | 2.268486 | 2.056467 |
| Proportion Variance | 0.200810 | 0.171987 | 0.126027 | 0.114248 |
| Cumulative Variance | 0.200810 | 0.372798 | 0.498824 | 0.613073 |

61.3%

| | FAC1 | FAC2 | FAC3 | FAC4 |
|---|---|---|---|---|
| Verb: most commonly used | -0.174105 | 0.013148 | 0.098985 | 0.539323 |
| Action verb | 0.131870 | 0.015472 | -0.039234 | 0.916157 |
| Directional verb | 0.130745 | -0.091794 | -0.049437 | 0.853418 |
| Adverb: most commonly used | 0.154552 | -0.547757 | 1.174826 | -0.021973 |
| Adverb of time | 0.193677 | 0.171381 | 0.709916 | 0.041387 |
| Adverb of degree | -0.201113 | -0.754177 | 0.157277 | 0.045434 |
| First person pronoun: I | -0.853183 | 0.054474 | -0.131128 | -0.130865 |
| Durative aspect: zhe | -0.021426 | 0.624432 | 0.029299 | 0.017067 |
| Conjunction | -0.055221 | -0.596186 | 0.035952 | -0.053222 |
| Modal particle | -0.208098 | 0.235601 | 0.180528 | 0.217695 |
| Adverbial hedge | -0.063486 | -0.840374 | 0.281723 | 0.061472 |
| Disyllabic word used in disyllabic template | 0.464959 | -0.083011 | 0.026919 | -0.174268 |
| Low-level vocabulary | -1.005067 | 0.065500 | 0.145831 | -0.143110 |
| High-level vocabulary | 0.613895 | 0.277319 | 0.016917 | -0.165214 |
| Non-HSK Vocabulary | 0.556531 | 0.344564 | -0.079418 | 0.182605 |
| Lexical diversity | 0.580975 | 0.389995 | 0.126566 | -0.004576 |
| Average word length | 0.434289 | -0.518553 | -0.347096 | 0.017766 |
| Average sentence length | 0.479686 | 0.225569 | 0.222237 | -0.102111 |

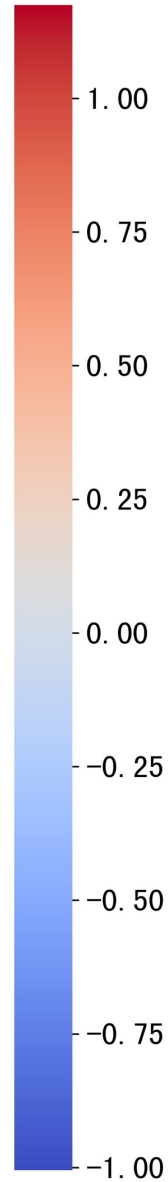
**HSK Vocabulary**

**Total: 5000words**

- HSK 6 ⇨ 2500 words
- HSK 5 ⇨ 1300 words
- HSK 4 ⇨ 600 words
- HSK 3 ⇨ 300 words
- HSK 2 ⇨ 150 words
- HSK 1 ⇨ 150 words

**Matrix of Factor Loadings**

| | FAC1 | FAC2 | FAC3 | FAC4 |
|---|---|---|---|---|
| Verb: most commonly used | -0.17 | 0.01 | 0.10 | 0.54 |
| Action verb | 0.13 | 0.02 | -0.04 | 0.92 |
| Directional verb | 0.13 | -0.09 | -0.05 | 0.85 |
| Adverb: most commonly used | 0.15 | -0.55 | 1.17 | -0.02 |
| Adverb of time | 0.19 | 0.17 | 0.71 | 0.04 |
| Adverb of degree | -0.20 | -0.75 | 0.16 | 0.05 |
| First person pronoun: I | -0.85 | 0.05 | -0.13 | -0.13 |
| Durative aspect: zhe | -0.02 | 0.62 | 0.03 | 0.02 |
| Conjunction | -0.06 | -0.60 | 0.04 | -0.05 |
| Modal particle | -0.21 | 0.24 | 0.18 | 0.22 |
| Adverbial hedge | -0.06 | -0.84 | 0.28 | 0.06 |
| Disyllabic word used in disyllabic template | 0.46 | -0.08 | 0.03 | -0.17 |
| Low-level vocabulary | -1.01 | 0.07 | 0.15 | -0.14 |
| High-level vocabulary | 0.61 | 0.28 | 0.02 | -0.17 |
| Non-HSK Vocabulary | 0.56 | 0.34 | -0.08 | 0.18 |
| Lexical diversity | 0.58 | 0.39 | 0.13 | -0.00 |
| Average word length | 0.43 | -0.52 | -0.35 | 0.02 |
| Average sentence length | 0.48 | 0.23 | 0.22 | -0.10 |




**HSK 1-2 Vocabulary** ←

**HSK 5-6 Vocabulary** ←

**Biber (1988:87) : [L]oadings having an absolute value less than .30 are generally excluded as unimportant even if they are statistically significant.**

## Dimension 1

| | |
|---|---|
| High-level vocabulary | 0.61 |
| Lexical diversity | 0.58 |
| Non-HSK Vocabulary | 0.56 |
| Average sentence length | 0.48 |
| Disyllabic word used in disyllabic template | 0.46 |
| (Average word length | 0.43 ) |
| …… | |
| First person pronoun: I/“我” | -0.85 |
| Low-level vocabulary | -1.01 |

## Dimension 2

| | |
|---|---|
| Durative aspect: zhe/“着” | 0.62 |
| (Lexical diversity | 0.39 ) |
| (Non-HSK Vocabulary | 0.34 ) |
| …… | |
| Average word length | -0.52 |
| (Adverb: most commonly used | -0.55 ) |
| Conjunction | -0.60 |
| Adverb of degree | -0.75 |
| Adverbial hedge | -0.84 |

## Dimension 3

| | |
|---|---|
| Adverb: most commonly used | 1.17 |
| Adverb of time | 0.71 |
| …… | |
| (Average word length | -0.35 ) |

## Dimension 4

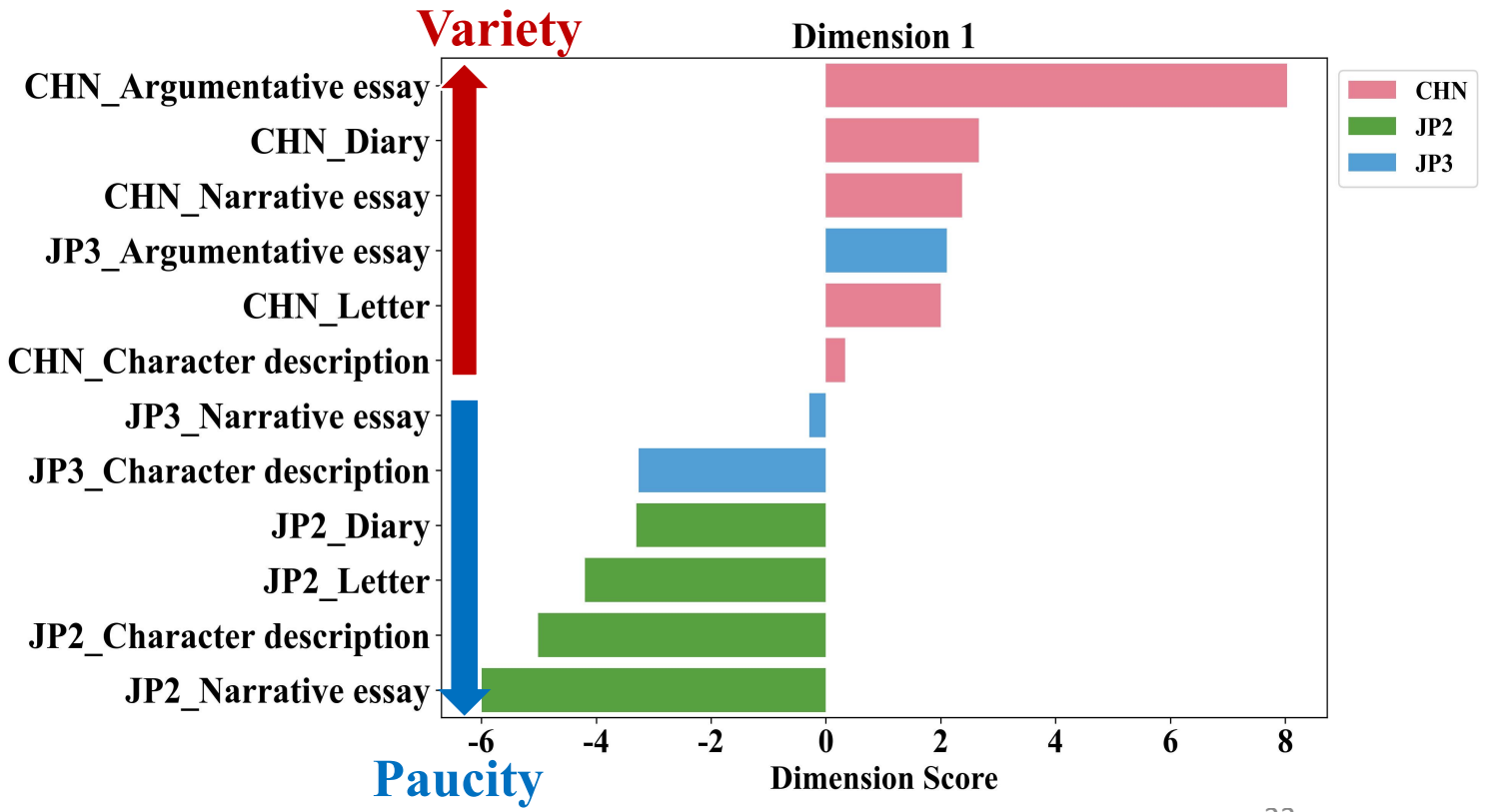
| | |
|---|---|
| Action verb | 0.92 |
| Directional verb | 0.85 |
| Verb: most commonly used | 0.54 |

# Interpretation of Dimensions

**Biber (1988:87) : In the interpretation of each factor, greater attention is given to those features with the largest loadings.**

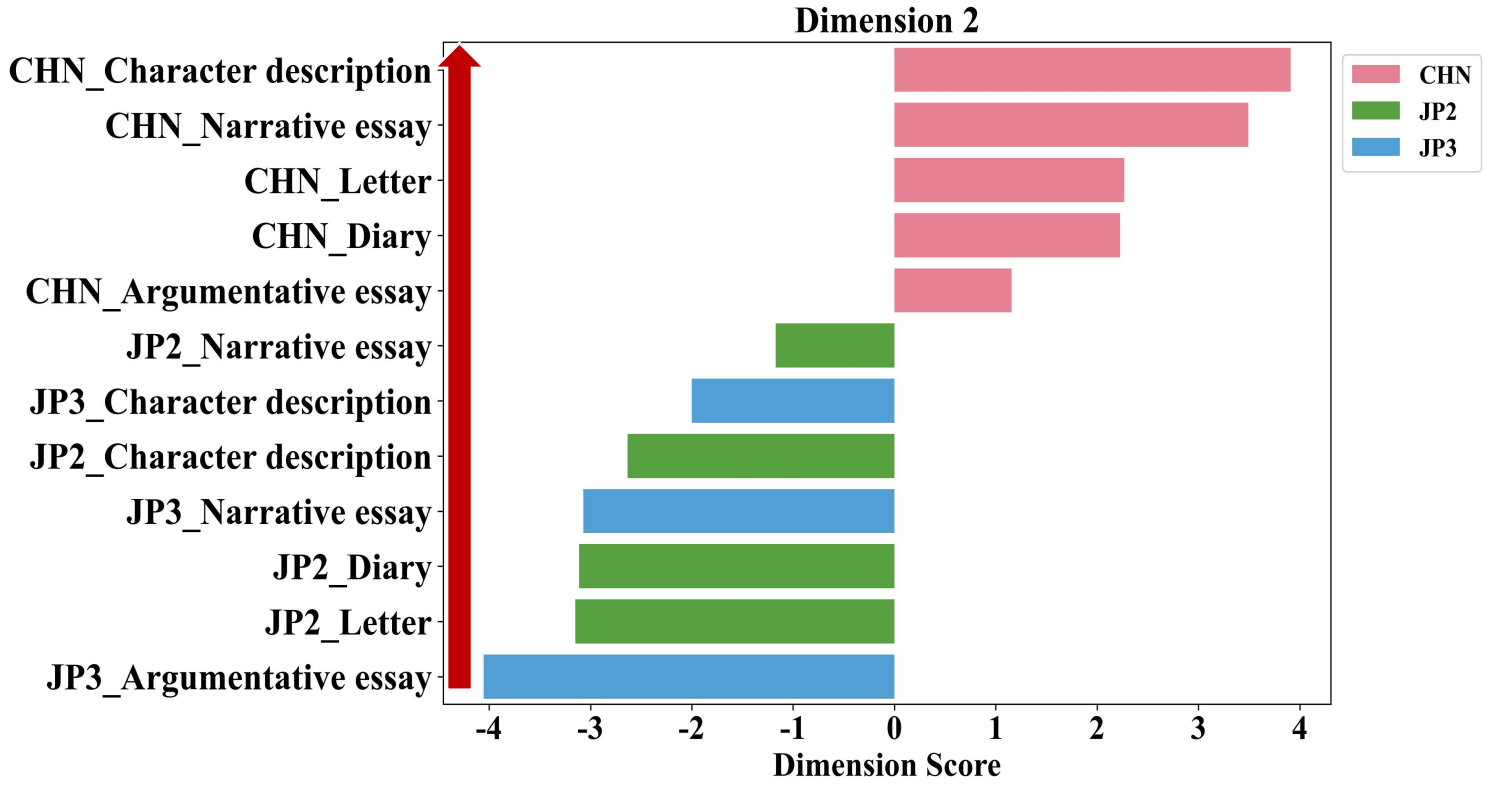
## Dimension 1: Variety vs. Paucity of Vocabulary Output

| Dimension 1 | |
|---|---|
| High-level vocabulary | 0.61 |
| Lexical diversity | 0.58 |
| Non-HSK Vocabulary | 0.56 |
| Average sentence length | 0.48 |
| Disyllabic word used in disyllabic template | 0.46 |
| (Average word length | 0.43 ) |
| …… | |
| First person pronoun: I/“我” | -0.85 |
| Low-level vocabulary | -1.01 |

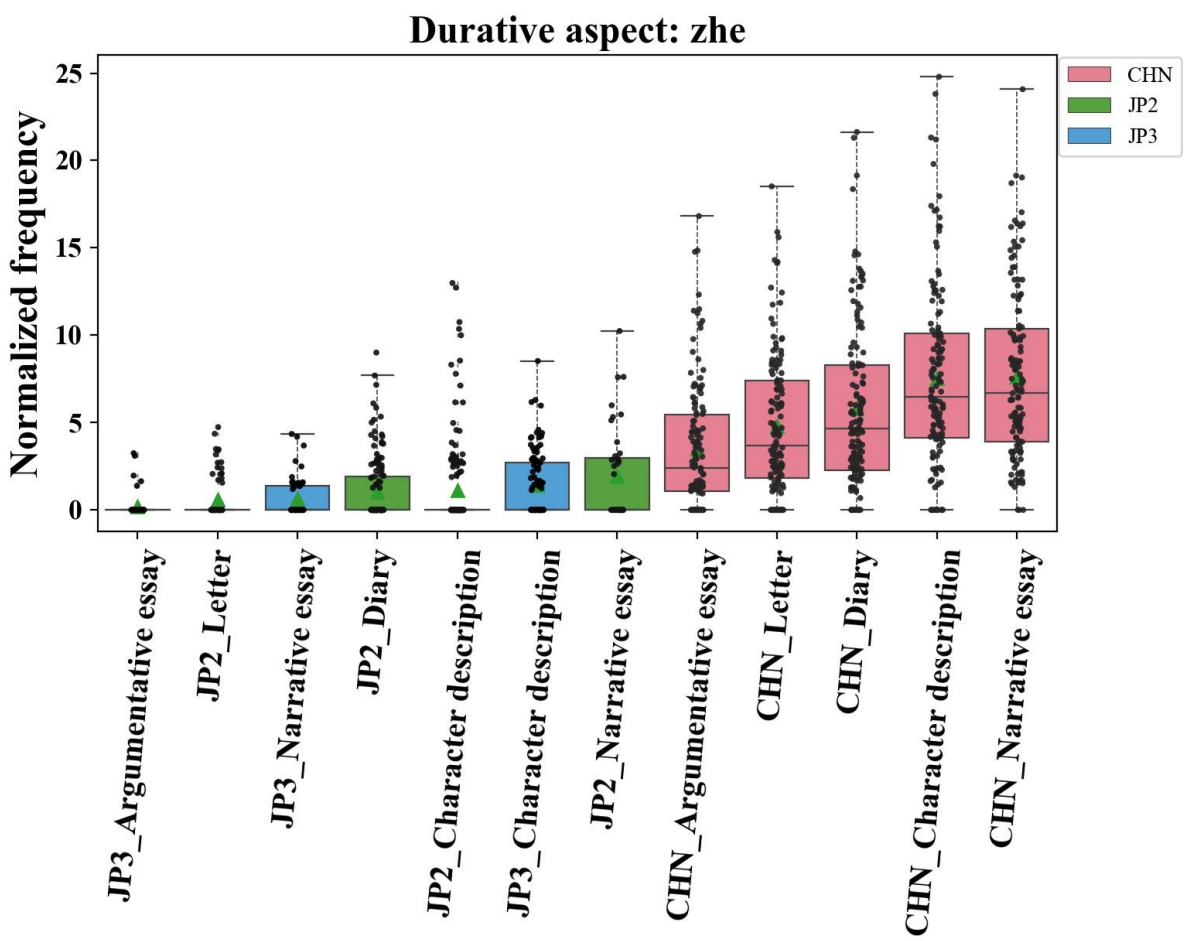


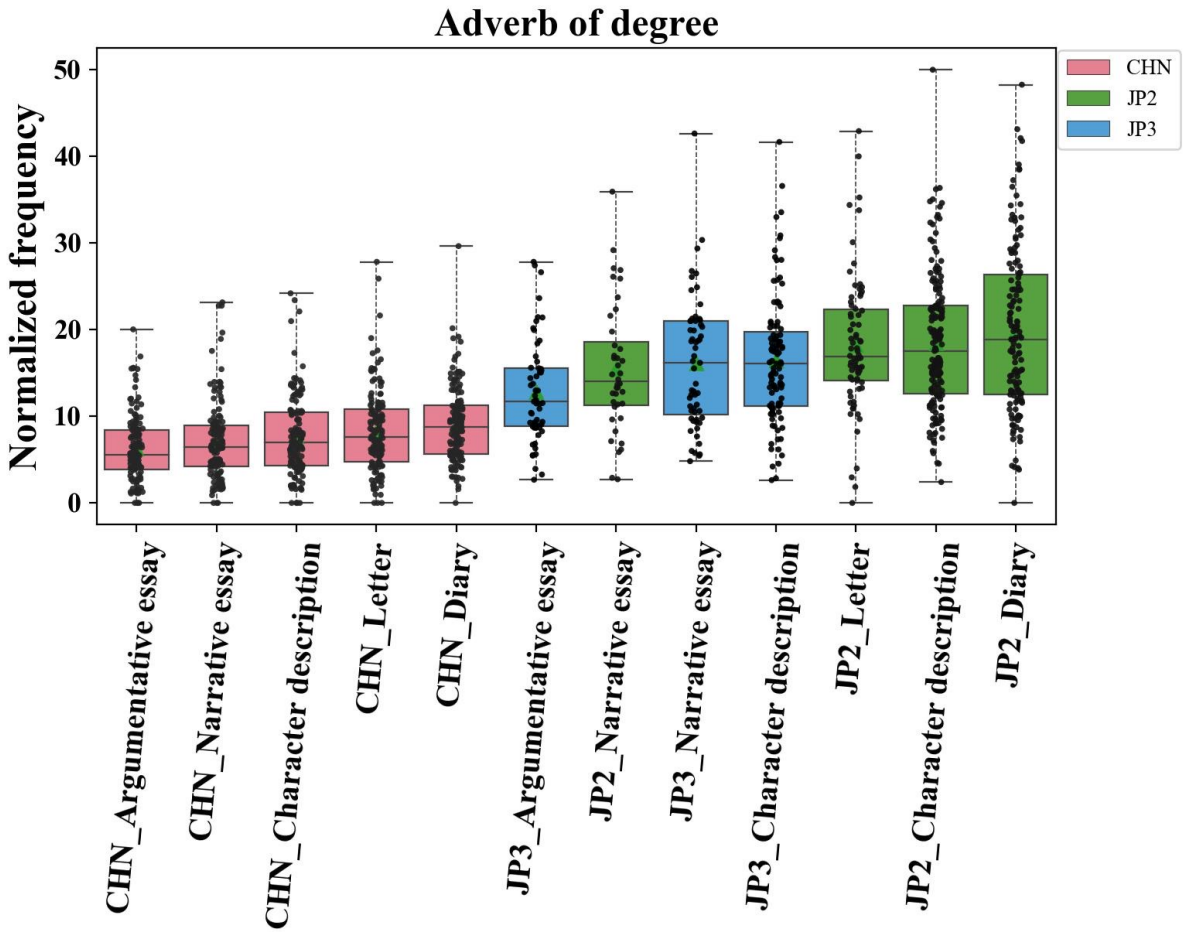
# Dimension 2: Situational Description

| Dimension 2 | |
|---|---|
| Durative aspect: zhe/“着” | 0.62 |
| (Lexical diversity | 0.39） |
| (Non-HSK Vocabulary | 0.34） |
| …… | |
| Average word length | -0.52 |
| (Adverb: most commonly used | -0.55） |
| Conjunction | -0.60 |
| Adverb of degree | -0.75 |
| Adverbial hedge | -0.84 |



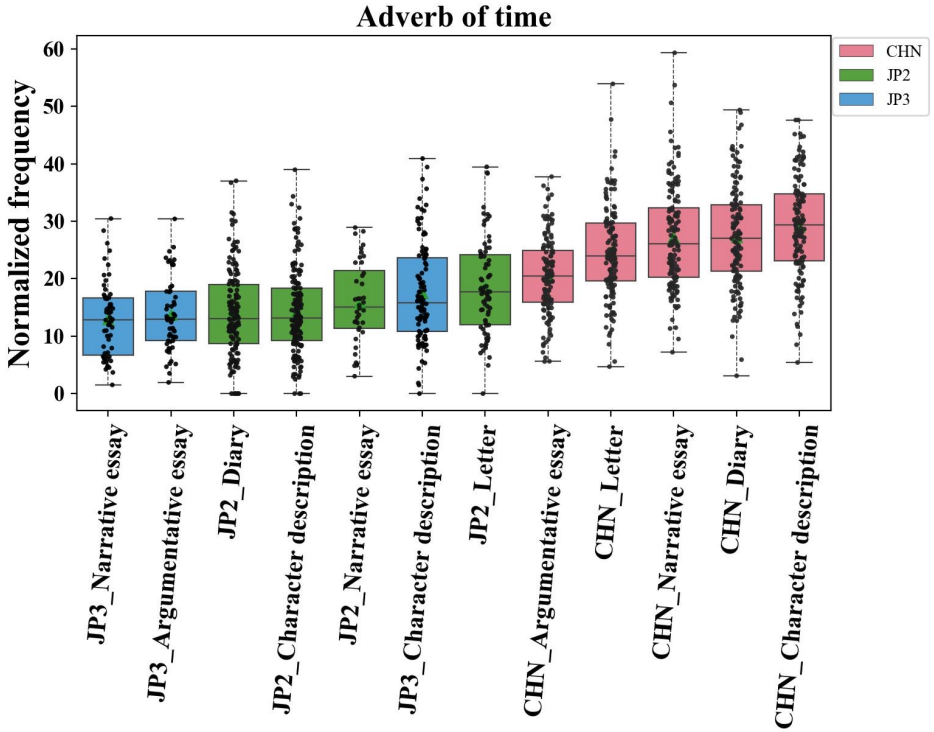
# Dimension 2: Situational Description

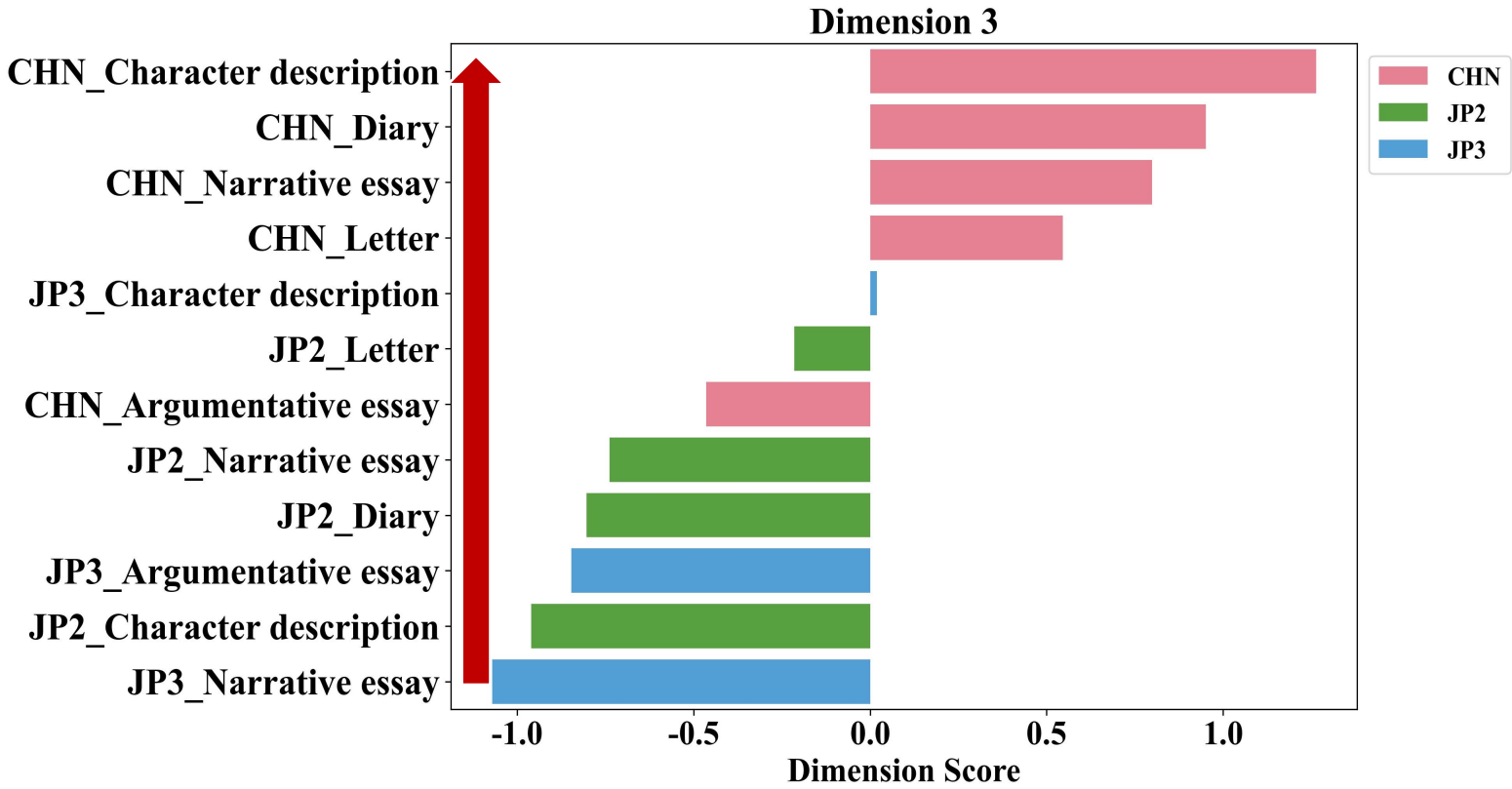


# Dimension 3: Adverbial Modification

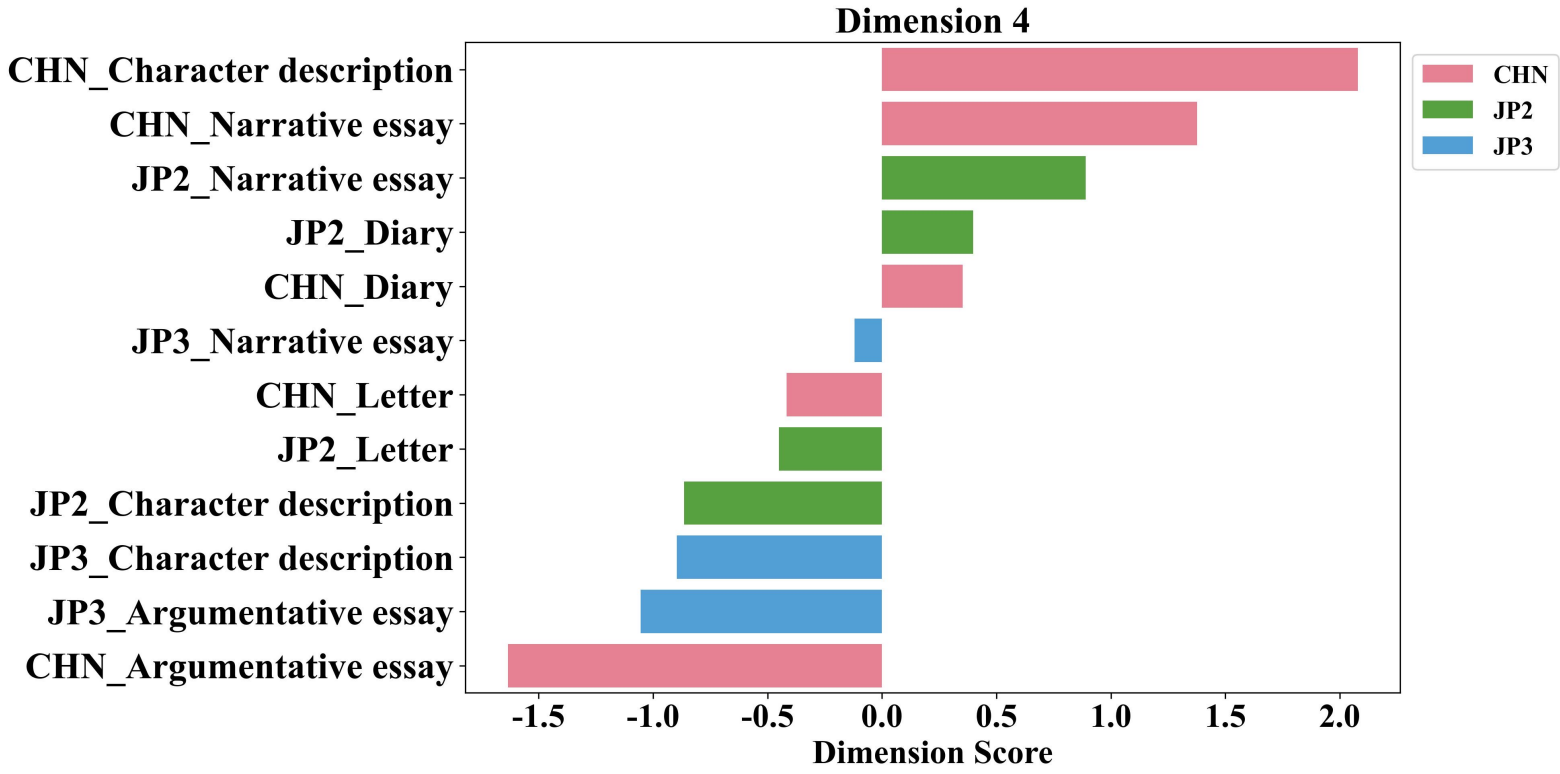
| Dimension 3 | |
|---|---|
| Adverb: most commonly used | 1.17 |
| Adverb of time | 0.71 |



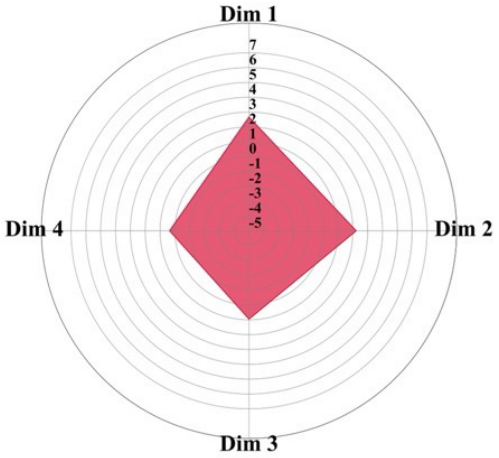
# Dimension 4: Action Behavior Description

| Dimension 4 | |
|---|---|
| Action verb | 0.92 |
| Directional verb | 0.85 |
| Verb: most commonly used | 0.54 |

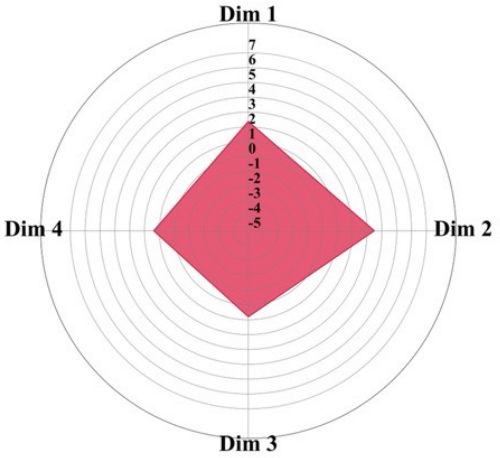


# L1 Genres : Radar Chart

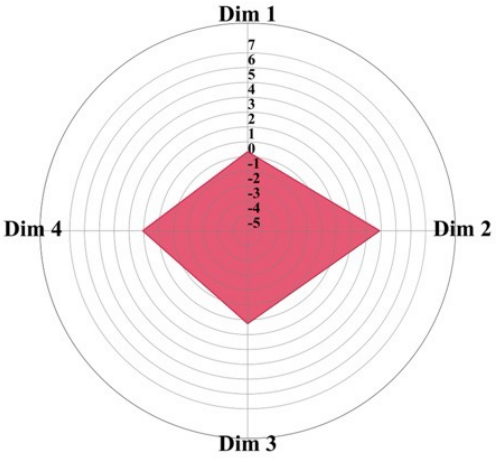
**Diary**



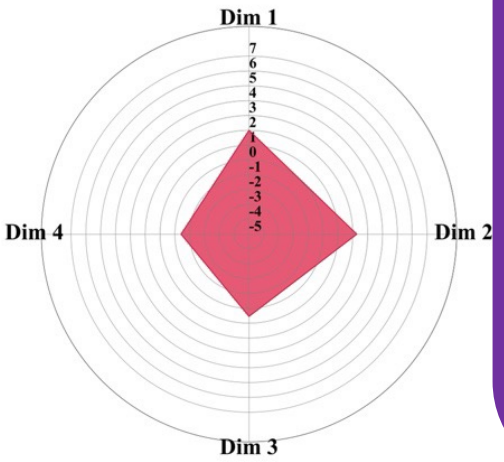

**Narrative essay**



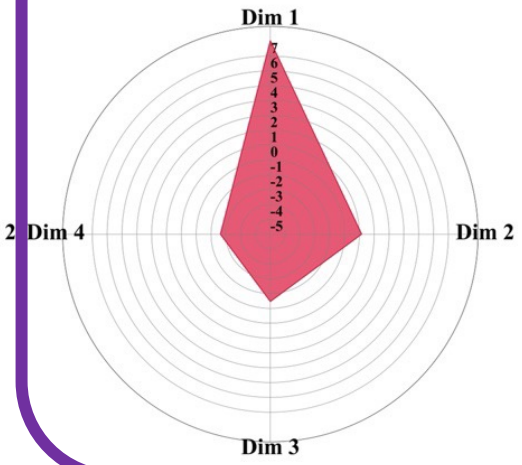

**Character description**




**Letter**




**Argumentative essay**




## Multiple Comparisons

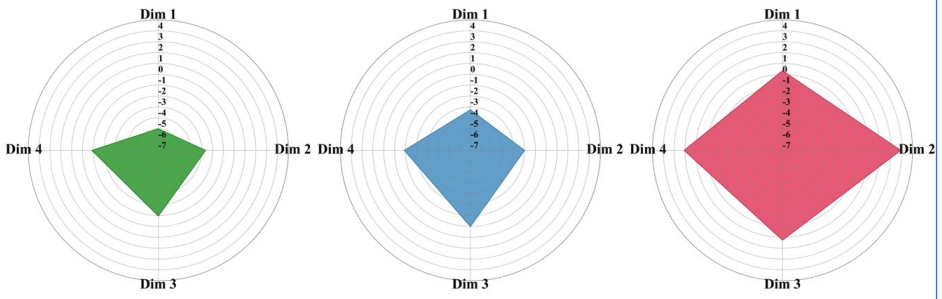
**The most special genre:**

**Argumentative essay**

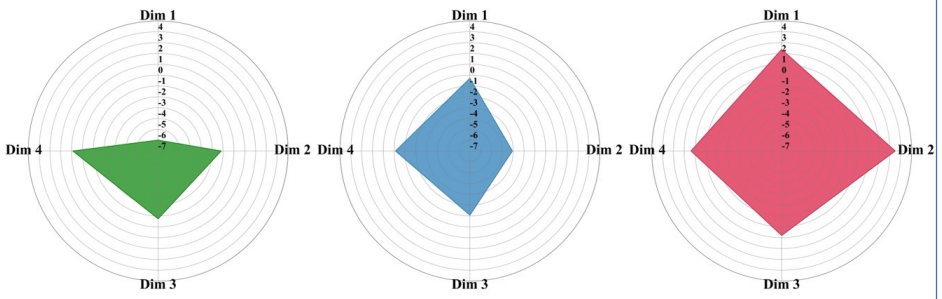
- **Dim 1 : scored highest**
- **Dim 2~4 : scored lowest**

# L2 vs. L1 Genres : Radar Chart

## Character description

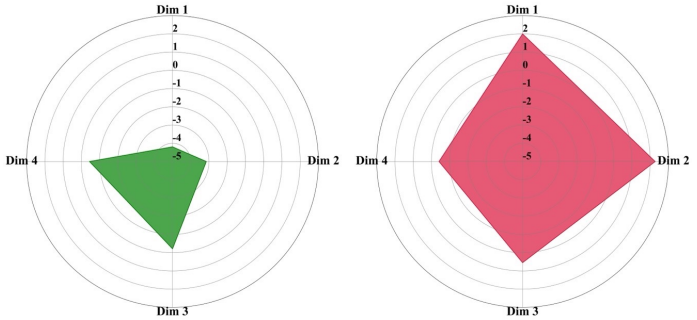



## Narrative essay

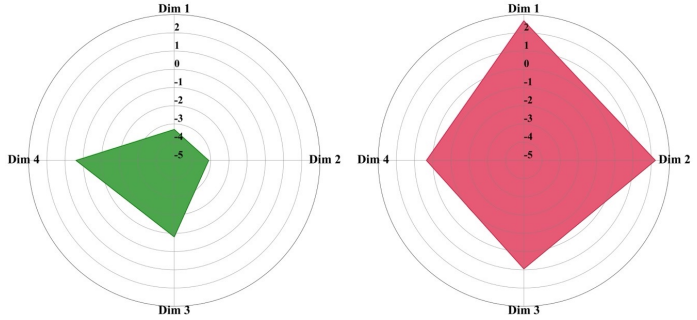



JP2 JP3 CHN

## Letter

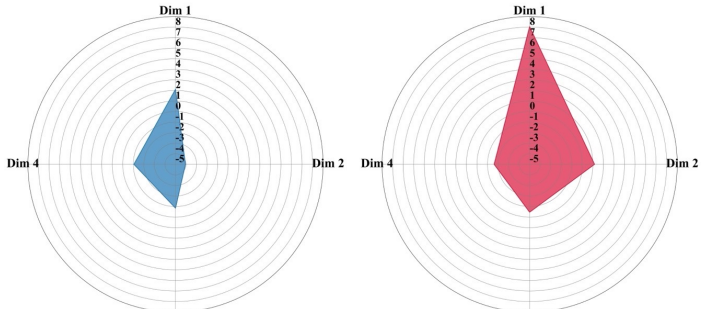



## Diary




JP2 CHN

## Argumentative essay




JP3 CHN

## Multiple Comparisons

**L2 vs. L1 Genres :**

- **Dim 1~3 :**
  **varied significantly**

# Conclusions

- L2 & L1 Chinese Writing Corpus
- Tools for counting the frequency of linguistic features
- Data pre-processing & Factor Analysis
- Successful application of MDA on L2 Chinese writing

# References


- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins Publishing.
- Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. Languages in contrast, 14(1), 7-34.
- Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing*, 26, 80-95.
- Goulart, L. (2021). Register variation in L1 and L2 student writing: A multidimensional analysis. *Register Studies*, 3(1), 115-143.
- Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 8(2), 183-207.
- Hardy, J. A., & Friginal, E. (2016). Genre variation in student writing: A multi-dimensional analysis. *Journal of English for Academic Purposes*, 22, 119-131.
- Pan, F. (2018). A multidimensional analysis of L1–L2 differences across three advanced levels. *Southern African Linguistics and Applied Language Studies*, 36(2), 117-131.
- Zhang, Z. S. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, 8(1), 209-240.

# Thanks for Listening.

# Q & A