A Corpus-based Study on the Common Spoken Words of the Illiterate

Guo Qijun

Jinan University, Guangzhou, Guangdong, China

Keywords: Corpus; Oral Language; Common Words; Measurement

Abstract: Based on a self-built spoken corpus of more than 300,000 characters used by the illiterate, this paper describes the basic spoken words from the frequency of use, the number of users and the degree of usage. The statistical analysis shows that the number of the illiterate's common spoken words is about 2000. In addition, the use rate of conjunctions is low. This reflects the narrow width of the use of illiterate spoken words and the narrow logic of linguistic expression.

1. Introduction

According to the definition of the Central Committee for the Eradication of Illiteracy, "those who know less than 500 words or even nothing are called the illiterate." Compared with the study of the language of the educated, the research results of illiterate language and its language use are rare. Besides, the research is mainly conducted from the field of psychological science and neuro-cognitive science. And the discussion focuses on the language processing mechanism and brain mechanism of the special population. Researchers such as Prinsloo, Mastin, Baynham, Michael J., Xiaohua Cao, Lihe Huang, Xiujun Li and Bosen Ma all studied illiterate language. On the basis of the above researches, this paper investigates the common spoken words of the illiterate with the help of the self-built corpus.

2. Corpus Sources and Their Labeling

The data used for analysis in this paper was collected from the recordings of natural, conversational speech between 12 illiteracy people and each other or speech between the illiterate and the educated from 4 villages in Qin'an County, Gansu Province. And the recordings sum up to 45 hours, which were collected in February 2014, May 2015 and August 2015 respectively. We randomly extracted 22 hours' recordings to transcribe and got the data of 434,435 characters. Then we eliminated the spoken data of the educated and finally got the data of 319,606 characters. The contents of corpus are related to the change of village, character experience, life status, family members, crop harvest and so on, which can basically show the features of the illiterate's daily spoken language.

In this paper, the data were segmented and tagged by CorpusWordParse, then corrected artificially according to the relevant dialect dictionaries and finally we got a glossary generated by AntConc. In the glossary, the total number of the words is 193,993 and the number of the word types is 7,056.

3. Quantitative Description of the Illiterate's Common Spoken Words

According to China National Committee for Terms in Sciences and Technologies (2011), common words are those people often use in social activities, with the characteristics of high use frequency and wide range of application. Researcher Huaiqing Fu (2004) claims that the most common standard used in the determination of common words is the frequency with which the words are used in the most popular books and periodicals. In this paper, we count the number of common spoken words from the aspects of word frequency, the number of users and the degree of usage.

Before the statistics, we need to remove the person names, place names and sayings in the corpus,

DOI: 10.25236/icepms.2018.225

which has a total of 7,056 words, including 527 person names, 241 place names, 24 other proper nouns as well as 260 farming proverbs, sayings and some non-fixed collocations. Besides, we eliminate 40 nonverbal elements resulting from inserting, makeup, kibitz and incoherent thinking and finally get 5,964 entries.

3.1 Word Frequency

According to the frequency of the occurrence of words in the corpus, it is divided into 3 classes and 6 categories.

Table 1: The Frequency Distribution of 5,964 Words

	Word Freq.	Number of Word Types	Cumulative Number of Word Types	Percentage	Average Word Freq.	Cumulative Word Freq.	Percentage of Cumulative Word Freq.
High Freq.	1000⩾	35	35	0.59%	2643.4	92519	48.83%
	100-999	195	230	3.27%	278.08	146744	77.46%
Medium Freq.	10-99	1059	1289	17.76%	29.13	177595	93.74%
	4-9	1094	2383	18.34%	5.81	183947	97.09%
Low Freq.	2-3	1406	3789	23.57%	2.37	187278	98.85%
	1	2175	5964	36.47%	1	189453	100.00%

The data in the above table show that the word frequency and the number of word types are in inverse relationship: the higher the word frequency, the smaller the number of word types; the lower the word frequency, the larger the number of word types. For instance, 195 high-frequency words account for only 3.86% of the total number of word types, while 3,581 low-frequency words account for 60.04% of the total number of word types. The above data also tell us that the higher the word frequency, the higher the text coverage of the words. For example, 195 high-frequency words cover 77.46% of the entire text, while 2,175 one-frequency words cover only 1.15% of the entire text. In other words, the most common words in the corpus are these 195 high-frequency words.

3.2 The Number of Users

The number of the words used by local people is defined as the frequency with which the words appear in different people' spoken language. One relevant expression is "distribution", which means the number of occurrences of a word in multiple texts. Supposing that there are 100 texts, and a word appears in these 100 texts, then the distribution of this word is 1. If the word only appears in 10 texts, then the distribution of this word is 0.1. In this study, the recordings of 12 illiterate people are transcribed into 12 texts respectively. Considering that, we call a word common word when it is used by 12 illiterate people. In contrast, if a word is used by only 1 or 2 illiterate people, it cannot be considered as a common word. The research data are shown in the table below.

Table 2: Usage Distribution of 5,964 Words

Number of Users	Number of Word Types	Cumulative Number of Word Types	Percentage	Cumulative Percentage	Total Word Freq.	Average Word Freq.	Percentage of Cumulative Word Freq.[1]
12	209	209	3.50%	3.50%	141177	675.49	74.52%
11	87	296	1.46%	4.96%	8939	102.75	79.24%
10	98	394	1.64%	6.61%	6044	61.67	82.43%
9	115	509	1.93%	8.53%	5149	44.77	85.14%
8	96	605	1.61%	10.14%	3230	33.65	86.85%
7	144	749	2.41%	12.56%	3859	26.80	88.89%
6	158	907	2.65%	15.21%	3063	19.39	90.50%
5	229	1136	3.84%	19.05%	3271	14.28	92.23%
4	297	1433	4.98%	24.03%	2955	9.95	93.79%
3	468	1901	7.85%	31.87%	3177	6.79	95.47%
2	878	2779	14.72%	46.60%	3456	3.94	97.29%
1	3185	5964	53.40%	100.00%	5133	1.61	100.00%

([1] In order to clearly show the relationship between the amount of words and the coverage of texts, the statistics excludes the words that are eliminated, so the cumulative text coverage is 1.)

Firstly, from the above table we can see that the number of users and the amount of words used are basically in inverse relationship: the larger the number of users, the smaller the amount of words; the smaller the number of users, the the larger the number of users. The data in the above table show that the words that are used by 12 illiterate people account for only 3.5% of all the words. Even when the words are used only by 2 illiterate people, they account for merely 46.6% of all the words. That is to say, there are 3,185 words appearing in the recordings of 1 illiterate people, so these words are not common words.

Secondly, the number of users is proportional to the percentage of cumulative word frequency. When the words are used by more illiterate people, the percentage of cumulative word frequency will be higher. In this study, the number of the words that are used by 12 illiterate people is 209, while its percentage of cumulative word frequency is up to 74.52%. And the number of the words that are used by only 1 illiterate people is 3,185, while its percentage of cumulative word frequency is as low as 2.71%.

Thirdly, the average word frequency is closely related to the number of users. When the words are used by more illiterate people, the average word frequency will be higher. In contrast, the average word frequency will be lower if the words are used by less illiterate people. In this study, the average word frequency of 209 words used by 12 illiterate people is 675.49, while the average word frequency of 3,185 words used by only 1 illiterate people is only 1.61.

We suppose the words that are used by 9 to 12 illiterate people are high-usage words, while those that are used by 4 to 8 illiterate people are medium-usage words and those that are used by 1 to 4 illiterate people are low-usage words. Then in the study there are 509 high-usage words accounting for 85.14% of the entire texts and 627 medium-usage words accounting for 7.09% of the entire texts. The cumulative coverage of both is up to 92.23%.

Thus, the common words in the illiterate spoken language are basically high-usage words and medium-usage words, with a total of 1136 words. Due to the limitation of the time and the number of the illiterate people that are surveyed, we assume that the words that are used by 3 or more illiterate people are common words and finally get 1,901 common words. In addition, these 1,901 common words are included in 2,175 high frequency words and intermediate frequency words according to the word frequency analysis.

3.3 The Words' Degree of Usage

The degree of usage is a method used to measure whether a certain word is commonly used or not from the following four indexes: differences in era, stylistic differences, distribution and word frequency. However, there are some differences in the calculation methods for the degree of usage. In this study, we use the method of calculating the words' degree of usage in the Modern Chinese Frequency Dictionary to investigate the degree of usage. And the formula is as follows.

$$Ui = \frac{\mathrm{Fi} * \mathrm{Di}}{\sum(\mathrm{Fj} * \mathrm{Di})}$$

In the formula above, Di represents the number of the illiterate people, while Ui represents the words' degree of usage and Fi represents the frequency of a certain word in the corpus. The numerator of this formula is the product of a certain word's frequency in the corpus and Di of this word, while the denominator represents the sum of all numerators. The following table is a Ui distribution of 5,964 words.

The table above clearly shows us that there is a close relationship among the degree of usage, average word frequency and average number of users: the larger the degree of usage, the higher the average word frequency and the larger the average number of users. For example, there are 146 words whose degree of usage≥0.001. And their average word frequency is 930.92. Their average number of users is 11.94 and their cumulative word frequency accounts for 71.74%. Compared with the high frequency words used by 12 illiterate people, there is little difference. In other words, the number of the most commonly used words in illiterate spoken language is within 200. Except for these common words, there are also some other words that are commonly used by the illiterate. For instance, there are 1,593 words whose degree of usage≥0.00001. And their average word frequency is 112.88. Their average number of users is 4 and their cumulative word frequency accounts for 94.92%. Besides, there are some words whose degree of usage≥0.000005. Their average number of users is 2.79 and their average word frequency is 5.79. These 1,976 words can be basically considered as common words used by the illiterate.

Table 3: The Degree of Usage of 5,964 Words

Degree of Usage	Number of Word Types	Cumulative Number of Word Types	Cumulative Percentage	Average Number of Users	Word Freq.	Percentage	Average Word Freq.
⩾0.001	146	146	2.45%	11.94	135915	71.74%	930.92
⩾0.0005	95	241	4.04%	11.26	11836	6.25%	125.59
⩾0.0001	375	616	10.33%	9.09	18729	9.89%	44.6
⩾0.00005	225	841	14.10%	6.6	5072	2.68%	22.57
⩾0.00001	752	1593	26.71%	4.29	8280	4.37%	11.01
⩾0.000005	383	1976	33.13%	2.79	2216	1.17%	5.79
⩾0.000001	1235	3211	53.84%	1.72	4074	2.15%	3.3
⩾0.0000005	2753	5964	100.00%	1	3331	1.76%	1.21

Although these 1,976 words account for only 33.13% of the whole, which has a total of 5,964 words, they make the greatest contribution to the illiterate spoken language. Instead, the other 3,988 words make little contribution to the illiterate spoken language despite their accounting for 66.87% of the whole.

3.4 The Distribution of Word Types

The following table shows the distribution of word types for 1,976 common words.

Table 4: Common Words and Their Distribution of Word Types

	Word Types	Number of Common Words	Percentage	Average Word Frequency
Content Words	Noun	656	33.20%	36.54
	Verb	634	32.09%	69.41
	Adjective	215	10.88%	31.86
	Numeral	53	2.68%	99.85
	Quantifier	81	4.10%	71.98
	Pronoun	110	5.57%	268.28
Function Words	Adverb	116	5.87%	162.52
	Preposition	23	1.16%	241.61
	Conjunction	16	0.81%	29.43
	Auxiliary	43	2.18%	842.26
	Interjection	17	0.86%	293.17
	Onomatopoetic Words	12	0.61%	40.5

From the table above we can see that the number of nouns is the largest, accounting for 33.2% of the common words and followed by verbs and adjectives. These three word types are the basic components of the illiterate spoken language. Of the 1,976 common words, these three word types account for 76.16%. In comparison, the number of function words is less. Auxiliary, preposition, conjunction and interjection account for only 5.62% of the 1,976 common words.

When it comes to the average word frequency, we can see that the average word frequency of the three major word types is lower, followed by numeral, quantifier and conjunction. The average word frequency is higher in several word types, such as auxiliary, pronoun, preposition and interjection. There are some changes between the ratio of each word type to common words and the ratio of each word type to the whole. Besides, content words have great changes while function words have little changes.

Hui Wang (2011) made statistical description of the basic words in Singapore's spoken Chinese, in which conjunction accounts for 1.4%. It is 0.59% higher than that of the illiterate spoken language, which indicates that the ratio of conjunctions is lower in the illiterate spoken language. Ochs Keenan & Elinor (1979) divide spoken language into unplanned spoken language and planned spoken language. They also points out that we should avoid using subordinate clauses in unplanned oral language. The illiterate spoken language is more likely to be considered as unplanned spoken language. Thus the number of conjunctions in the illiterate spoken language is lower and the logic of linguistic expression is weaker.

4. Conclusions

In the above analysis, we can draw a few assumptions.

Firstly, the common words used by the illiterate in their daily spoken language are as few as approximately 2,000 words. Hui Wang's survey in 2011 shows that there are about 2,550 basic words in modern Chinese spoken language. There are 3,000 words in Basic Words in Mandarin Chinese while the Syllabus of Chinese Words contains 3,051 words, including 1,033 first-degree words and 2,018 second-degree words. Thus, the number of the common words used in the illiterate spoken language is lower than those used in the spoken language of the educated.

Secondly, there is a low ratio in using conjunctions in the illiterate spoken language, in which the logic of linguistic expression is weaker.

References

[1] Ochs, E. Planned and unplanned discourse, in Syntax and semantics, vol. 12: Discourse and syntax, ed. by T. Givon[M]. New York: Academic Press, 1979.

[2] Prinsloo, Mastin & Baynham, Michael J. (eds.). Literacy Studies Vol. 1-5[M]. London: Sage Publishing. 2013.

[3] Cao Xiaohua.ect. The Behavior and Brain Function of Word Processing in Illiterate Subjects, Advances in Psychological Science,2009,17(05):917-922.

[4] Li Xiujun .ect. Brain Function of Chinese Character Font Size and Phonological Processing between Literate and Illiterate Subjects: An FMRI Study, Journal of Psychological Science 2013,36(6).

[5] Ma bosen. A Contrastive Study of the Distributional Patterns of Anaphoric Reference to Persons in Spontaneous Conversation between Illiterates and Literates. Journal of Foreign Languages,2007 (03).

[6] Ma bosen. Personal reference strategies by the illiterate and literate: A contrastive Study. Contemporary Linguistics.2009, (01).

[7] Ma bosen. Anaphora correction in Chinese natural conversation. Journal of PLA University of Foreign Languages ,2014(01) .

[8] Wang hui. On the basic vocabulary of spoken Chinese. Studies of the Chinese Language ,2011(05).