International Journal of Linguistics Studies
ISSN: 2754-2599
DOI: 10.32996/ijls
Journal Homepage: www.al-kindipublisher.com/index.php/ijls

IJLS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Quantitative Approach to the Study on Length-Frequency Relationship of English Patterns

Cheng Yi
School of International Studies, Zhejiang University, Hangzhou, China
Corresponding Author: Cheng Yi, **E-mail**: 22005041@zju.edu.cn

| ABSTRACT
As a master of lexicology and corpus-driven linguistic studies, pattern grammar shares the features of concern on collocation and approximation to natural language, which provides an optimization idea for the extensively used phrase structure grammar and dependency grammar in natural language processing. This research adopts the methods of quantitative linguistics, taking patterns as the research focus to explore the length-frequency relationship of patterns, verify Zipf law and synergic linguistic model in respect of patterns, and analyze the usage characteristics of patterns in different registers. The results have demonstrated that Zipf's law is also applicable to patterns, and the synergic linguistic model shows a preferable goodness of fit on the description of the pattern length-frequency relations. This study provides bidirectional contributions to pattern grammar and the synergic relationship between length and frequency. In addition, it offers a refinement method for the application of patterns in different registers.

| KEYWORDS

Pattern grammar, length, frequency, register

| ARTICLE INFORMATION

ACCEPTED: 20 March 2023 **PUBLISHED:** 31 March 2023 **DOI:** 10.32996/ijls.2023.3.2.1

1. Introduction

In lexical research, length and frequency are two elementary properties of words. Accordingly, the length-frequency relations from the perspective of quantitative linguistics have been tapped by many scholars. The existing studies have verified the universality of the Zipfian principle (1949) on word length and frequency, as well as the appropriate mathematical model to portray it. And the Zipfian theory has also been extended from word length to lexical level. However, previous research did not investigate discontinuous word strings, and the meaningful lexical bundles may be too fixed to reflect universality.

As a master of lexicology and corpus-driven linguistic studies, pattern grammar shares the features of concern on collocation and approximation to natural language. Patterns incorporate the transitivity of verbs and the continuation of verbs, nouns, and adjectives with their complement clauses and prepositional phrases. It is the intersection of grammar and lexis as well as a unit of meaning. Patterns can be regarded as a language unit consisting of a semi-fixed structure with a functional "slot" between functional words to be filled by content words in actual language usage (e.g. it V to n). Nevertheless, pattern grammar summarizes most of the patterns that are frequently used in practice, avoiding the generation of sentences that are theoretically feasible but not in line with idiomatic usage, thus playing an important role in language application, especially in the field of English teaching and natural language processing.

This research adopts the methods of quantitative linguistics, taking patterns as the research focus, to explore the length-frequency relationship of patterns, verify Zipf law and synergic linguistic model $y = ax^{-b}$ in respect of patterns and to analyze the usage characteristics of patterns in different registers. This study provides bidirectional contributions to pattern grammar and the synergic

relationship between length and frequency. In addition, it offers a refinement method for the application of patterns in different registers.

To fulfill the above purpose, research questions are hereby listed:

1) What is the regular relationship of English pattern length on the frequency of pattern use?
2) What is the best mathematical model to describe the above rules?
3) What kind of regular changes are found with the change of text registers?

2. Methodology

2.1 Extraction of Frequency

The study regards the two properties of patterns as the object: length and frequency of occurrence. The length of the pattern was measured by words. For example, the specific pattern "it suggests that" derived from semi-fixed patterns "it V clause" is classified as a three-word pattern. And the length size will be set between 2~5 considering the basic size distribution of the pattern list; thus, 60 pattern types are to be involved with the one-word pattern "V-ing" excluded from the pattern list presented above. The frequency of the pattern is the data collected from the corpus. As for the data source, BNC Baby distinguishes itself with a wide range of stylistic types and strong representativeness of modern English spoken and written language use. This can meet the needs of this study to reveal the interaction between the length and frequency of English patterns. The tags were excerpted from the BNC tagging grammar when writing the corresponding regular expressions. For instance, when the pattern "it V clause" is searched through the whole corpus, three tag types are used, including "V\w+", which stands for V, "DTQ", "CJT", and "AVQ", which stand for marks of the clause and the tag of "it". And the regular expression of this pattern is combined like this:

<w type="PNP" lemma="it">\w+ </w><w type="V\w+" lemma="\w+">\w+ </w><w type="(DTQ|CJT|AVQ)" lemma="\w+">\w+

2.2 Procedure

The chosen patterns in the pattern list were searched in the corpus through the concordance program Antconc *3.4.4*. And N-grams tool, as well as a regular expression, will be used to extract the linguistic data, especially the pattern frequency. Besides, the frequency value of chosen patterns is to be collected not only from searching through the entire corpus but also from the 4 sub-corpora with different text types.

The frequency of chosen patterns will be computed into mean average using Microsoft Excel 2003. As the length size varies from two to five, four groups in a total of the average frequency statistics were collected. And SPSS v.25 will then be used to conduct both regression analysis and goodness-of-fit test. The regression analysis will be applied to verify the quantitative relationship between pattern length and frequency. Two variables are involved here: the independent variable X represents pattern length, and the dependent variable Y represents average frequency. The two power function models aforementioned will be evaluated by goodness-of-fit values R^2 to distinguish the better model from the other. It is worth mentioning that coefficients like a and b would also be measured and analyzed. As for the power function model, when the corpus samples are equal in length, the value of the model coefficient changes regularly with the variation of the text style: for the frequency of Chinese single-syllable words, the value of the coefficient "b" is relatively stable with the subtle variation found in different text types, while the coefficient "a" is obviously different in that the value of it in spoken style is significantly higher than that in written style (Deng & Feng, 2013).

3. Results

3.1 Length-frequency relationship for English patterns

The original frequency results are counted for each n-word English pattern in Table 1. The second entries in the table, named "type", are the count of different patterns with the same number of words. The third entries are the average frequency of each N-word pattern (N=2~5). An obvious decline in type number is noticed, as well as that of pattern frequency from a 2-word pattern to a 5-word pattern. The average frequency of occurrence decreases from 42181.8 of 2-word patterns to 424 of 5-word patterns for every million words. These results conform to Zipf's "least effort" theory to interpret the regular relationship between word length and word frequency in terms of a macroscopic view.

Table 1. Type and Average frequency of N-word patterns

Length	*pattern type*	*pattern frequency*
2	***25***	***42181.8***
3	***23***	***9838.35***
4	***10***	***771.19***
5	***3***	***424***

3.2 The best fitting function

2 functions, $y = ax^{-b}$ (b > 0) and $y = ax^{b}e^{-cx}$, are the most suitable model to describe length-frequency relations for the Chinese and English languages according to previous studies. Then the initial data were processed to screen out the optimum function based on the test of goodness of fit. The result shows the optimum function should be the power function, which refers to the first model $y = ax^{-b}$ (R^2= 0.960, p < 0.05). After plugging the coefficients into the equation, power function $y = 2066374.747x^{-5.352}$ is chosen to precisely predict the frequency of any English patterns in the 4 million corpora. The fitting curves are shown in the following figure 1, where the observed values of frequency are represented by grey spots, and the theoretical values land on the curves. The dotted line curve stands for the power function model, and the solid line curve for other models.

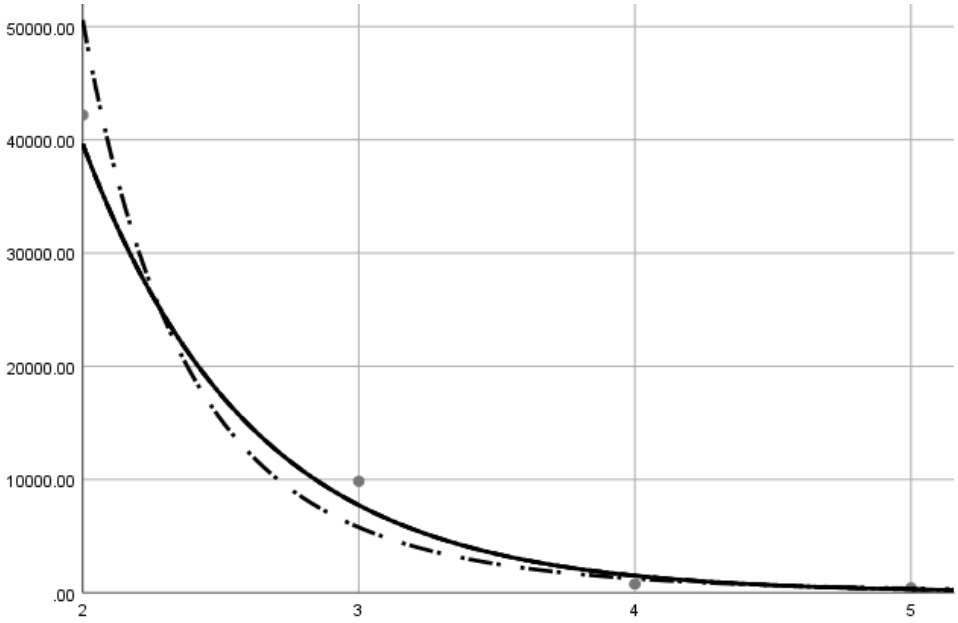

Figure 1. Fitting Curves

3.3 Analysis of different registers

Similarly, the frequency of N-word patterns was counted and summarized in table 2. It could be observed that the frequency of occurrence of the N-word pattern appears to have significant divergence. From the second column to the fifth column, pattern frequency was typed into 4 categories in terms of registers, including academic text, conversation, fiction and news. It could be observed that the data diverge in different registers. The number of 5-word patterns counted in academic text significantly surpasses those found in the other 3 corpora. Moreover, it even surmounts the frequency of 4-word patterns in academic text. Besides, total patterns of different lengths in the conversational text are apparently less than those in other corpora. In fact, texts forming the conversation corpora were collected from spoken English while the other 3 corpora are writing English texts. The results seem to coincide with previous studies that the frequency of multi-word differs in spoken texts and those in academic writing, since in speech, people tend to use fixed sequences with both functional words and content words while most patterns in writing English are formulaic sequences or the semi-fixed patterns with a variable slot.

Table 2. The average frequency of N-word patterns in 4 registers

Length	*Freq of ACAdemic text*	*Freq of conversational text*	*Freq of fiction text*	*freq of news text*
2	*12647.72*	*6528.84*	*10215.96*	*11554.12*
3	*2881.17*	*2064.70*	*2445.43*	*2645.74*
4	*153.8*	*230.8*	*225.3*	*125.7*
5	*217*	*48.33*	*76*	*82.67*

Concerning the coefficients, the value of the coefficient "b" subtly fluctuates among different text types, while the coefficient "a" violates Deng and Feng's finding that for the data of Chinese single-syllable words, the coefficient in spoken style is significantly higher than that in written style. In contrast, the value of "a" computed from the conversational register appears to be the lowest one. Cross-language variations may dominate such differences, or this coefficient cannot suggest the quantitative features among registers. Pieces of evidence from more languages are needed to converge a conclusion.

Table 3. Fitting the power function $y=ax^{-b}$ (b>0) to the data of N-word patterns

$y=ax^{-b}$ (b>0)	Academic Texts	Conversational Texts	Fiction Texts	News Texts
	a=438571.184 b=5.042 R^2 = 0.891	a=408250.602 b=5.419 R^2 = 0.952	a=617782.238 b=5.546 R^2 = 0.970	a=831173.518 b=5.867 R^2 = 0.941

4. Discussion

Differences emerge in the frequency of N-word patterns among registers, especially between conversational spoken text and academic writing text. There are principally two apparent exception values observed in our results. One is the shortage of pattern usage in conversational text, and the other is the frequency of 5-word patterns outnumbering 4-word patterns in the academic register. For the former phenomenon, it is indicated that information density plays an important role in underlying the total number of patterns in different registers. Apparently, the academic text is typical of dense information, while news and fiction registers come next, spoken text the last. The general numbers of patterns that occurred in each register coincidently follow this regular order in Table 2. In addition, language structure appears to be particularly required in registers with dense information, as patterns with slot structure imply the grammar structure of language. Biber and Conrad (2009) pointed out different situational traits and purposes of written and spoken registers. Writing registers allow preparation and revision of the text, while most spoken registers are offhand without time to retrieve the "deep structure" of language usage and organize it.

Besides, most written registers have to communicate purpose to exchange new information, while most spoken registers stress interactive functions to maintain the relationship. Consequently, the language of academic writing is refined and interweaved by grammar structure to achieve information maximization, while spoken language, in general, uses fewer patterns except the fixed bundles consisting of functional words and definite content words. And using structure certainly adds burdens to the cognitive processing of language production. People are intentionally trained to acquire the structure of academic language; in contrast, spoken language in a natural context is inclined to "least effort" without burdens. In fact, Zipf raised the "the least effort" principle and developed it into the assumption of an inverse relationship between the length of words and the frequency of occurrences because he held that human beings have this in nature to input the least effort to achieve the most effect. Results of the present research manifest that this principle has a more intuitive reflection on the conversational register.

Another exception emerging in results helps us to recognize the specific structure used in the academic register. To study the 4-word and 5-word patterns that brought differences among registers, we further retrieved the specific concordance lines of example sentences to probe into more characteristics of patterns used in different registers. The most frequent 4-word and 5-word patterns

contain a series of *it is* clause that is largely adopted in academic writing but seldom used in virtual conversation (e.g., *it V adj clause*, *it be V-ed clause,* and *it V adj to-inf*). The formal subject *it* is a particular placeholder in the English language to balance sentence structure and avoid a long subject. While academic writing requires impact information and objectivity, the sentence length tends to be relatively long, and the usage of the personal subject is necessarily substituted by passive voice and formal subject *it*.

5. Conclusion

We made the first attempt to extend the length-frequency study to patterns. Despite the above findings, there are certain limitations in this research. The first limitation lies in the pattern list. The patterns picked up in the study are relatively adequate for our research; nevertheless, some patterns were omitted in consideration of our focus, which may lead to errors in coefficient estimation to some extent. To unfold a full view of patterns, more studies are required from perspectives besides length-frequency relations. The second limitation is concerned with the method of extraction. Although the regular expression adopted in this study shows more accuracy and efficiency than the N-gram tool in the extraction and identification of patterns, the operation on semi-fixed patterns could still be refined through computational models. Future studies are suggested to conduct optimization of the algorithm or to separate the semi-fixed pattern and flexible patterns to further explore the characteristics of pattern frequency.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Biber, D., & Conrad, S. (2009). Register, genre, and style. Cambridge University Press.

[2] Deng, Y. C., & Feng, Z. W. (2013). A quantitative linguistic study on the relationship between word length and word frequency. *Journal of Foreign Languages, 36*, 29–39.

[3] Zipf, G. K. (1949). Human behavior and the principle of least effort: *An introduction to human ecology*. Cambridge, MA: Addiso-Wesley.