CIT. Journal of Computing and Information Technology, Vol. 31, No. 1, March 2023, 39–55
doi: 10.20532/cit.2023.1005719

N-gram Language Model for Chinese Function-word-centered Patterns

Jie Song[1], Yixiao Liu[2] and Yunhua Qu[2]

[1]School of Foreign Languages, Zhejiang University of Finance & Economics, Hangzhou, China
[2]School of International Studies, Zhejiang University, Hangzhou, China

N-gram language modelling, a proven and effective method in NLP, is widely used to calculate the probability of a sentence in natural language. Language pattern is a linguistic level between word/character and sentence, which exists in pattern grammar. In this research, the approach of language model and language pattern are combined for the first time, and language patterns are studied by use of the N-gram model. Chinese function-word-centered patterns are extracted from the LCMC corpus, and aligned into pattern chains. The language model is trained from these chains to investigate the properties and distribution of Chinese function words, the interaction of content words and function words, and the interaction between patterns. The results indicate that there are approximately 10,000 function-word-centered patterns in the texts, which are distributed exponentially. This research summarizes the most common function-word-centered patterns and content-word-centered patterns, and discusses the interactions of patterns based on corpus data. The bigram language model of these patterns reflects the restrictions of function words. In addition, the research adopts an innovative method to visualize the interactions between patterns. This research fills the research gap between word/character and sentence, and reveals basic Chinese pattern categories and the interactions between patterns, which makes a significant contribution to Chinese linguistic research, and improves the efficiency of NLP.

ACM CCS (2012) Classification: Applied Computing → Arts and humanities → Language translation

Keywords: N-gram language model, language pattern, function word

1. Introduction

There are two major concepts in this research: the N-gram language model and language patterns. Corpus-based language modeling is one of the basic methods in Natural Language Processing (NLP), which refers to different probabilistic and statistical methods used to determine the chances of a given series of words occurring in a text. Language models have been shown to achieve remarkable performance across a variety of natural language tasks, such as machine translation [1], word segmentation [2], text-classification [3], *etc.* Furthermore, the models can deal with specific languages in both spoken and written form.

Language patterns is a concept from pattern grammar [4] and is an approach to English grammar that generalizes from the patterning of individual words as observed through concordance lines from a large corpus of general English [5, 6]. Moreover, language patterns are a way to describe the behavior of lexical items [7]. A pattern can be regarded as a cluster of words/characters on the level between words and sentences, and these clusters are bounded and connected by meanings, *e.g.*, "verb-noun-noun" "verb-it", *etc.* Compared with phrase structure grammar and dependency grammar, pattern grammar conducts a linear analysis of sentence construction, rather than hierarchic [8]. This proves to be an ideal theoretical framework for exploring Chinese linguistic features, because word ordering is the foundation of Chinese grammar structure [9–11] and, furthermore, function words constitute an essential grammatical means in Chinese. Analyzing Chinese function-word-centered patterns N-gram model has its advantages as N-gram model is an ideal model in representing a linear structure and, thus, it can be adopted to observe how patterns are formed and how patterns interact with each other.

It is believed that the development of large language models is mainly a feat of engineering and thus far has been largely disconnected from the field of linguistics [12]. Therefore, this study presents a novel application of statistical language modeling to analyze Chinese function word patterns. It provides data-driven insights into prevalent Chinese patterns and quantitative pattern interaction probabilities. At the same time, it explores the links between language models and linguistic studies.

1.1. N-gram Language Model

Natural Language Processing (NLP) is one of the major fields in Artificial Intelligence (AI), and it contains many topics, including machine translation (MT), automatic summarizing/ abstracting, information retrieval, document categorization/classification, information extraction, automatic proofreading, and speech recognition [13]. In addition, NLP is also concerned with numerous issues related to social science, language teaching, *etc.* A language model assigns a probability to a piece of unseen text [14], based on some training data. It is a crucial concept of NLP, especially in the study of speech recognition [15], machine translation [16, 17], Chinese automatic word segmentation [18, 19], and syntactic analysis [20].

A language model that deals with an N-gram is called the N-gram language model. N-gram model predicts the occurrence of a word based on the occurrence of its N-1 previous words, and it's trained on a corpus of text. The N-gram language model is adopted extensively in NLP practice, including automatic handwriting recognition [21], machine translation [22] speech recognition [23], *etc.* N-gram language modeling has been adopted in a few linguistic and cultural studies. For example, Zeng and Greenfield [24] investigates changing cultural values in China from 1970 to 2008 based on Google N-gram, Jiang and Wen [25] explored the automatic grading of E-C translation based on N-gram model, and Qin and Kong [26] made comparison of English and Chinese based on N-gram. However, few research works apply language models to study language patterns in Chinese so far. In this research, N-gram language modeling is used to study the behavior of pattern grammar for the first time. Studying pattern grammar with N-gram model has its advantages as it can be adopted to observe how patterns are formed and how patterns interact with each other.

1.2. Pattern Grammar

A proportion of our everyday language is formulaic, making it is predictable in form and idiomatic and is typically stored in a fixed or semi-fixed form. Pattern grammar [7] is an approach to English grammar that generalizes from the patterning of individual words as observed through concordance lines from a large corpus of general English [5, 6]. It is a description of the behavior of a lexical item. By its definition, a pattern is a series of words that are bounded together at the level of meaning. A string of symbols identifies the pattern, with the node word in capitals. For example, the pattern *V n to-inf* specifies that the verb (*V*) is followed by, and governs, a noun phrase (*n*) and then a to-infinitive clause (*to-inf*), *e.g.*, told us to go home [27].

One of the pioneering works about pattern grammar is Hornby's *A Guide to Patterns and Usage in English*, published in 1954. In this book, Hornby attempts to provide practical guidance to language learners on usage rather than focus on analysis [28].

"Analysis is helpful, but the learner is, or should be, more concerned with sentence-building. For this, he needs to know the patterns of English sentences and to be told which words enter into which patterns." [28]

As quoted above, Hornby points out that patterns are the building blocks of sentences, and he emphasizes that the language learner should "be told which words enter into which patterns". To do that, he listed the most common verbs used in each pattern, *e.g.*, the common verbs used in the pattern of "*V of N*". It can be seen that the grammar pattern coding uses abbreviated symbols to stand for word classes or clause types. For example, it expresses verbs, nouns, and adjectives by '*v*', '*n*', and '*adj*', that clauses by '*that*', and to-infinitive clauses by '*to-inf*'. Other kinds of patterns also exist that include specific words rather than classes, such as "*N from n*", which normally includes prepositions, such as "*at*", "*for*", "*with*", "*from*", *etc.*

At the time of Hornby's seminal work, corpus linguistics was not established, such that the examples were limited both in quantity and category. In modern linguistics, this work can be accomplished by a corpus which provides much more raw material from the real natural language, "*and using a large corpus to study pattern grammar will lead to observations about language that it has not been possible to make before.*" [4]. The new methods of using corpus in the study of language patterns have led to new observations based on the language data, which are the source of novel theories and deeper understanding.

When observing pattern and phraseology in a corpus, it can be seen that lexical facts and grammatical facts appear simultaneously, and the choices of patterns are closely related to contexts, which is markedly different from the traditional view that the distinction between lexis and grammar is blurred in the actual use of language. This observation urges us to reconsider the definitions of lexis, grammar, sense, and pattern, to identify and describe patterns that exist in our language, how they are formed, and how they interact with each other. In this way, we can obtain a clearer picture of the relationship of meaning and form, as well the relationship between content words and function words. This is also a major aim of this research.

Grammatical analysis of English by using the theory of pattern has played an important role in language teaching. In Britain, pattern-based dictionaries and grammar books have been published, such as *The Collins Cobuild English Language Dictionary* [29], *The Collins Cobuild English Grammar* [30], *The Collins Cobuild English Dictionary* [31], *etc.* These works focus on describing common and typical English patterns in real language contexts and enhance English learners' ability to recognize and use English vocabulary and structure. Some studies applied pattern grammar to English research. For instance, Huang *et al.* [32] applied pattern grammar to a grammar checking system of language learners. Chen and Liang [33] built an error checking system of English written verbs for Chinese students. Xiong [34] studied the English patterns with the word "it" in different registers. Yu [35] developed a program for automatic recognition and extraction of English verb patterns.

However, compared with the researche of pattern grammar in English, the study and application of pattern grammar in China is still in its initial stage. Wang [36] introduced pattern grammar, while Chen and Liang [37] reviewed the characteristics, and application value of pattern grammar. The advantage of pattern grammar in analyzing Chinese language lies in the following two points: word order is the foundation of Chinese grammar structure [9–11]; and function words constitute essential grammatical means in Chinese [10, 38]. Interactive relations between functional words and content words in the Chinese sentence are not fully depicted and described by phrase structure grammar and dependency grammar. To fill in this gap, this study combines the N-gram model and pattern grammar to investigate what patterns exist in Chinese sentences, as well as the distribution of functional patterns and interactions between the patterns, which contributes to Chinese pattern grammar study, as well as NLP practice in general.

1.3. Research Questions

In previous research, language modeling usually focusses on specific words or characters only, none of them focus on the overall distribution and interaction of Chinese patterns. Similarly, previous studies on pattern grammar were only aimed at teaching and dictionary compiling, and thus far no research has attempted to describe patterns with language modeling. To fill in the gap, in this research, a tentative step is made to combine the N-gram language model and language patterns. The general principles and formalism of the idea are discussed theoretically, and a specific application is made by combining the bigram ($N = 2$) language model and Chinese function-word-centered patterns. Since there are too many pattern types to study thoroughly in a single study, the focus of this research is on Chinese patterns with function words. The specific research questions are as follows:

1. How do function words combine with content words or other function words to form patterns in Chinese language?
2. What patterns exist in Chinese sentences, and how do they distribute in texts?
3. How do patterns interact with each other in Chinese language?

2. Methodology

To train a language model for Chinese patterns, an appropriate Chinese corpus is needed. In this research, the ICTPOS-tagged LCMC corpus is adopted as the database and its details are presented in section 2.1. The language model is trained, and a more mathematically formal description is developed in section 2.2, in which the language model is described by a tensor operator. The index form of tensor is more appropriate for computers to process. Statistical computation methods used in the study are illustrated in section 2.3.

2.1. Corpus

The language model in this project is built on the LCMC corpus, *The Lancaster Corpus of Mandarin Chinese*, which has been constructed as part of a research project undertaken by the Linguistics Department at the Lancaster University. The corpus is designed as a Chinese match of the Freiburg-LOB (FLOB) Corpus of British English. It provides a valuable resource for contrastive studies between English and Chinese, as well as a sound basis for monolingual investigations of Chinese [39].

The LCMC corpus is constructed by using written Mandarin Chinese texts published in mainland China to ensure some degree of textual homogeneity. It contains approximately 1,000,000 words. The text categories are listed in Table 1 [39]. The plain written texts of the LCMC corpus have been transcribed, with tables, figures, formulae, and special symbols omitted and replaced with a gap, which is marked by the word "omission". Long citations from translated texts or texts produced outside of the sampling period were also omitted so that the effect of translations is excluded and the quality of the targeted language is guaranteed.

Alphabetic languages, such as English, in which words are separated with spaces, can thus be easily counted. In contrast, Chinese contains a running number of words. As a consequence, it is impossible to count word occurrence numbers within raw texts. As the proofreading of raw electronic texts is both time-consuming and expensive, it was economical to proofread an excessively large sample, but use only approximately 2,000 words. Based on a pilot study of the ratio of words to characters, the ratio of 1:1.6 is adopted, which means that a 3,200-character running text corresponds to a 2,000-word sample.

Table 1. List of text categories.

A	Press: reportage
B	Press: editorials
C	Press: reviews
D	Religion
E	Skills, trades, and hobbies
F	Popular lore
G	Biographies and essays
H	Miscellaneous: reports and official documents
J	Science: academic prose
K	General fiction
L	Mystery and detective fiction
M	Science fiction
N	Adventure and martial arts fiction
P	Romantic fiction
R	Humor

The POS tag of words in LCMC are tagged by the ICTPOS-tagging algorithm, and the notations of all kinds of POS tag are listed in Appendix. The texts after tagging are in the following form:

$$\cdots \text{word}_i/\text{POS}_i \ \text{word}_{i+1}/\text{POS}_{i+1} \ \text{word}_{i+2}/\text{POS}_{i+2} \cdots$$

2.2. Pattern Segmentation

After the preparation of the corpus, the next step is to extract patterns from texts. Therefore, an explicit definition of Chinese function-word-centered patterns is given, and the pattern segmentation algorithm is presented.

A pattern, as reviewed in the previous section, is a collection of words/characters, which serves both grammatical and semantic functions. The

segmentation of patterns in texts depends on the definition of pattern. In this section, a function-word-centered segmentation of pattern is proposed in order to study the properties of function words in Chinese.

In this segmentation scheme, all of the POS tag of LCMC are classified into three categories:

1. function word
2. content word
3. punctuation.

These three categories are denoted by i, r, and w, respectively. In this project, function words consist of preposition (p), conjunction (c), auxiliary (u), and directional locality (f), *i.e.*, $i \in \{p, c, u, f\}$. Starting from a leading word (LW, beginning of a sentence or following the end of the last pattern), it is either r or i (w is skipped).

If $LW = r$ (which means that the leading word is a content word), consider the next-to-leading word (*NLW*). If $NLW = r$ or w, end the pattern, and *LW* is considered to be a single [r] pattern. If $NLW = i$, then consider the next-to-next-to-leading word (*NNLW* or N^2LW). If $N^2LW = i$ or w, end the pattern, and $LW + NLW$ is considered to be a [$r\ i$] pattern. If $N^2LW = r$, $LW + NLW + N^2LW$ is considered to be a [$r\ i\ r$] pattern. If $LW = i$, end the pattern with r or w. Specifically, if:

$$\underbrace{LW + NLW + ... + N^nLW}_{n+1} = \underbrace{i + i + ... + i}_{n} + w$$

then $LW + NLW + ... + N^{n-1}LW$ forms a pure i pattern with n elements, *i.e.*, [$i \cdot \cdot \cdot i$]. If the following expression holds

$$LW + NLW + ... + N^nLW = i + i + ... + i + r$$

then the whole chain forms a [$i \cdot \cdot \cdot i\ r$] pattern with $n + 1$ elements. All five kinds of patterns are summarized as follows:

1. [r] 2. [$r\ i$] 3. [$r\ i\ r$]
4. [$i \cdot \cdot \cdot i$] 5. [$i \cdot \cdot \cdot i\ r$].

In these pattern types, function words play a crucial role. This segmentation scheme focuses on the connection between content words and the cluster of function words. This is why a succession of content words is considered as a discrete pattern while a succession of function words is regarded as a single pattern. Note that the second pattern is consistent with the fourth and the fifth pattern. For example:

$$r + i + i + i + i + r + w \quad \Longrightarrow [r\ i] + [i\ i\ i\ r] + w$$

Here is an example from the LCMC corpus to illustrate how segmentation works:

他/*rr* 在/*p* 惊慌/*an* 中/*f* 残忍/*a* 地/*ude2* 打/*v* 昏/*v* 了/*ule* 女孩儿/*n*, /*wd* 自己/*rr* 跌跌撞撞/*z* 地/*ude2* 逃/*v* 出/*vf* 了/*ule* 小屋/*n* ... /*ws*

Extract the sentence structure, and use the segmentation rules:

$$r+i+i+i+r+i+r+r+i+r+w+r+r+i+r+r+i+r+w$$
$$\Longrightarrow [ri]+[ri]+[rir]+[rir]+w+[r]+[rir]+[rir]+w$$

The pattern segmentation rules defined in the previous section can be described by the following algorithms. For patterns with $LW = r$, see Figure 1; and for patterns with $LW = i$, see Figure 2.

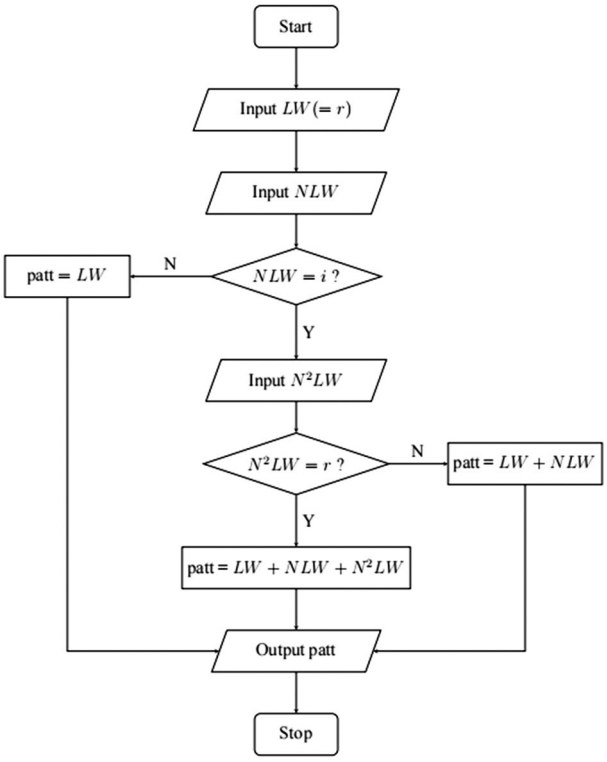

Figure 1. Pattern segmentation algorithm for $LW = r$.

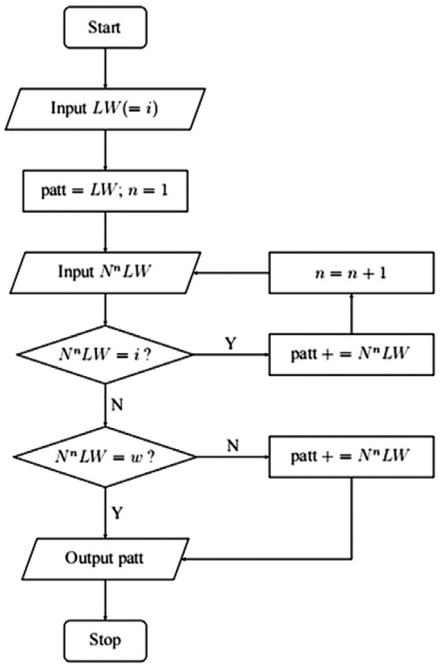

Figure 2. Pattern segmentation algorithm for $LW = i$.

Patterns starting with r, as shown in Figure 1, can be easily identified by nested "if-else" judgements. Once a [*r i*] pattern is confirmed, add a "#" mark to the NLW ; and for a [*r i r*] pattern, tag both *NLW* and N^2LW. Patterns starting with *i*, as shown in Figure 2, are produced from a loop. Once a new word is added to the pattern, tag that word with "#". Then, the next pattern always starts from a non-punctuation word without "#". In this way, every word is guaranteed to appear in a sole pattern, and crossing of patterns is avoided.

The general principles of building an N-gram language model have already been stated in the literature review. However, those expressions are inappropriate to run on a computer. Therefore, this section is devoted to establishing a different formalism by applying new mathematical language to reformulate the previous concepts and statements.

As stated in section 1.1, a language model computes the probability of given sentences. In this project, the subjects are replaced by pattern chains, which can be illustrated as:

$$LM\colon S \to P(S) \quad S = \{p_1, ..., p_l\} \equiv p^l{}_1 \tag{1}$$

If LM is an N-gram language model, the computing process is as follows:

$$P(S) = LM\left(p_1{}^N\right) \cdot LM\left(p_2{}^{N+1}\right) \cdot ... \cdot LM\left(p^l{}_{1-N+1}\right). \tag{2}$$

For each operation, *LM* acts on *N* patterns in the pattern set, and outputs a number. This suggests that an N-gram language model should be regarded as an *N* rank tensor-like operator *T*, such that:

$$T\colon \mathbb{P} \times \mathbb{P} \times ... \times \mathbb{P} \to [0, 1] \tag{3}$$

where $\mathbb{P}$ represents the set of patterns with *d* elements, $\|\mathbb{P}\| = d$, namely:

$$\mathbb{P} = \{p_1, ..., p_d\} \tag{4}$$

It is worth noting that *T* is actually not a real tensor in mathematical terms, because $\mathbb{P}$ is not a linear space (there is no operation of adding and multiplying). However, since we will not consider the transformation of coordinates in the current situations, calling *T* a tensor will not cause any ambiguity. Then, we can employ the index notation of tensor to run on a computer. An *N* rank tensor contains *N* indexes: T_x1, ..., *xN*. Each index runs over the whole $\mathbb{P}$, *i.e.*, from 1 to *d*. Once all of the indexes are determined, *T* can find one of its components, which is defined as a conditional probability:

$$T_{x_1,...,x_N} = P\left(p_{x_N} \mid p_{x_1}, ..., p_{x_{N-1}}\right) \tag{5}$$

Taking a bigram as an example, *T* has two indexes, which means that we can fill its components in a *d*d* matrix:

$$T = \begin{pmatrix} T_{11} & \cdots & T_{1d} \\ \vdots & \ddots & \vdots \\ T_{d1} & \cdots & T_{dd} \end{pmatrix}, \tag{6}$$

where

$$T_{ij} = P\left(p_j \mid p_i\right) = \frac{C\left(p_i, p_j\right)}{\sum_j C\left(p_i, p_j\right)} \tag{7}$$

For the trigram language model, the *ijk*-component is defined as follows:

$$T_{ijk} = P\left(p_k \mid p_i, p_j\right) = \frac{C\left(p_i, p_j, p_k\right)}{\sum_k C\left(p_i, p_j, p_k\right)} \tag{8}$$

To train an N-gram language model means exactly to determine all d^N elements.

Smoothing technique is vital in N-gram modeling [40], and common smoothing methods include Jelinek-Mercer smoothing, Katz smoothing, Witten-Bell smoothing, absolute discounting [41], Kneser-Ney smoothing [42] and modified Kneser-Ney smoothing [40]. Comprehensive contrastive study shows that the modified Kneser-Ney smoothing exerts ideal performance when compared to other soothing techniques [40]. Therefore, the Kneser-Ney smoothing is adopted as the smoothing technique in this study.

2.3. Model Computations

After the ICTPOS-tagged LCMC corpus has been converted to a collection of patterns, or pattern chains, the frequency of different types of patterns as well as their co-occurrence are then processed with a self-drafted Python script and the scientific computing software Mathematica. Then, the data are fitted by a bounded exponential function. All patterns are analyzed and interpreted from a linguistic perspective. The bigram language model of function-word-centered patterns is presented with the "Rainbow" function of the Mathematica software, and bubble charts are adopted to visualize three types of pattern interactions, including [*r*]+[*r*], [∃*i*]+[*r*], [*r*]+[∃*i*] and [∃*i*]+[∃*i*]. ([∃*i*] represents a pattern which contains at least one function word)

3. Results and Discussion

In section 3.1, data regarding the internal structure and mechanism of the Chinese pattern set are presented. Distributions of Chinese pattern set and their types are fitted using a non-linear function model. In section 3.2, the author defined different types of function-word-centered patterns, and presented the explicit distribution of patterns, which reflects the internal structure of the pattern set. In section 3.3, the research demonstrates the distribution of the most frequent 267 function-word-centered patterns. In section 3.4, the bigram language model for function-word-centered patterns is presented. Bubble charts are adopted to visualize the interactions between patterns.

3.1. Pattern Set

The ICTPOS-tagged LCMC corpus is converted to a collection of 659,114 patterns, or 43,279 pattern chains (15.23 patterns per chain), which are generated by a pattern set consisting of 9,160 different patterns. Figure 3 reflects the dynamic expansion of the pattern set.

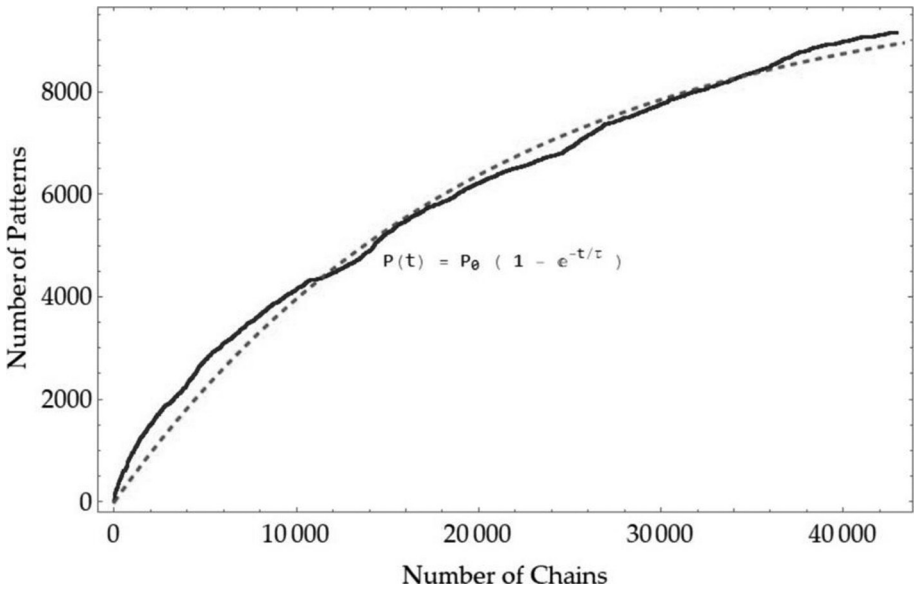

Figure 3. Growth of the pattern set.

Note: the solid line shows the growth of the pattern set, and the dashed line is the nonlinear fit of the data.

Specifically, every time a chain is processed, the instantaneous size of the pattern set is recorded. The data can be fitted by a bounded exponential function:

$$P(t) = P_0\left(1 - e^{-t/\tau}\right) \quad (9)$$

where t is the number of chains, and P_0 and τ are parameters. The best fitting performed using Mathematica finds that $P_0 = 10,139$, $\tau = 20,171$, with correlation $R = 0.998$. These two parameters reflect important properties of the pattern set.

Mathematically, P_0 is the value of $P(t)$ when $t \to \infty$. This limit allows to enlarge the corpus to infinity and exhaust all language phenomena in Chinese. Therefore, this limit value informs us that the size of the pattern set is bounded from above by P_0. In other words, a finite number (approximately 10,000) of patterns can cover all texts in Chinese. Obviously, however, the corpus cannot be enlarged to infinity in practice. However, the asymptotic behavior of the exponential term in Eq. (9) guarantees that if the corpus is sufficiently large, it can be regarded as infinity. This so-called "sufficiently large" threshold is determined by the value of τ, which this study defines as the "corpus unit".

Table 2 shows that when $t = 5\tau \approx 100,000$ chains, the pattern number occupies 99.33% of P_0, which approximately equals P_0. Note that for LCMC, $t = 43,279 \approx 2.15\tau$. From the data, it is found that a corpus with approximately 5τ chains can cover almost all patterns in the set. By sorting the pattern set in numerical order, it can be seen that the fitting function decays exponentially.

Eq. (9) can also be derived theoretically. Assume that at some text length t, the size of the pattern set is $P(t)$. Then, add some small length of text Δt, and observe how many new patterns are added to the set, *i.e.*, $\Delta P = P(t + \Delta t) - P(t)$. A natural guess is that this increment ΔP is proportional to Δt, as well as the number of unknown patterns:

$$P(t + \Delta t) - P(t) = \frac{P_0 - P(t)}{\tau}\Delta t \quad (10)$$

where P_0 is the final size of the pattern set, and therefore $P_0 - P(t)$ is the number of unknown patterns, and Δt is a coefficient of proportionality. Then, take the infinitesimal limit, $\Delta t \to dt$, and Eq. (10) becomes a differential equation:

$$\frac{dP(t)}{dt} = \frac{P_0 - P(t)}{\tau} \quad (11)$$

With the initial condition that $P(0) = 0$, the solution to Eq. (11) is:

$$P(t) = P_0\left(1 - e^{-t/\tau}\right) \quad (12)$$

which is exactly what can be found from the data. Determination of the values of P_0 and τ depends on the distribution of patterns and is not calculated in this research. However, at least we believe that the exponential form is a satisfactory description for the growth of the pattern set in written Chinese.

3.2. Pattern Types

In this section, we defined different types of function-word-centered patterns. Here, we present the explicit distribution of Chinese patterns over 11 basic pattern types, which reflects the internal structure of the pattern set (see Table 3).

As listed in Table 3, all patterns are distributed over 11 pattern types. By simple calculation, it is determined that there are 747,822 content words and 164,911 function words (4.5:1) in LCMC. The result shows that most patterns with function words exist in the form [*r i r*]. The number of this type outweighs the sum of other types with function words. It is also found

Table 2. Coverage rate of pattern set in five corpus units.

t	τ	2τ	3τ	4τ	5τ
$1 - e^{-t/\tau}$	63.21%	86.47%	95.02%	98.17%	99.33%

Table 3. Pattern distribution over different pattern types.

Pattern Type	Number	Ratio
[*r i r*]	93,452	59.75%
[*i r*]	44,685	28.57%
[*r i*]	14,595	9.33%
[*i i r*]	3,300	2.11%
[*i i i r*]	193	<1%
[*i i*]	149	<1%
[*i i i i r*]	15	<0.01%
[*i i i*]	7	<0.01%
[*i i i i*]	3	<0.01%
[*i i i i i r*]	3	<0.01%
[*i i i i i i i i i r*]	1	<0.01%

that, except for the [*i r*] pattern, other pattern types which begin with *i* contribute very little to the total number. Therefore, these patterns (<1%) are generally considered trivial in our study, while non-trivial parts exist among the former four types in Table 3.

By observing the pattern types in the trivial part, it can be seen that the number falls drastically with the number of *i*. By appropriate functional fitting, the numbers of [$i \cdots i$] and [$i \cdots i\ r$] decay exponentially, and the number of [i^n] is in the same order of magnitude as the number of [$i^{n+1}\ r$]. The best fitting functions are as follows:

$$[i \cdots i] \rightarrow f_i(x) = 4585\mathrm{e}^{-3.4256(x-1)} \quad (13)$$

$$[i \cdots i\ r] \rightarrow f_r(x) = 3300\mathrm{e}^{-2.8366(x-1)} \quad (14)$$

The fitting lines are plotted in Figure 4.

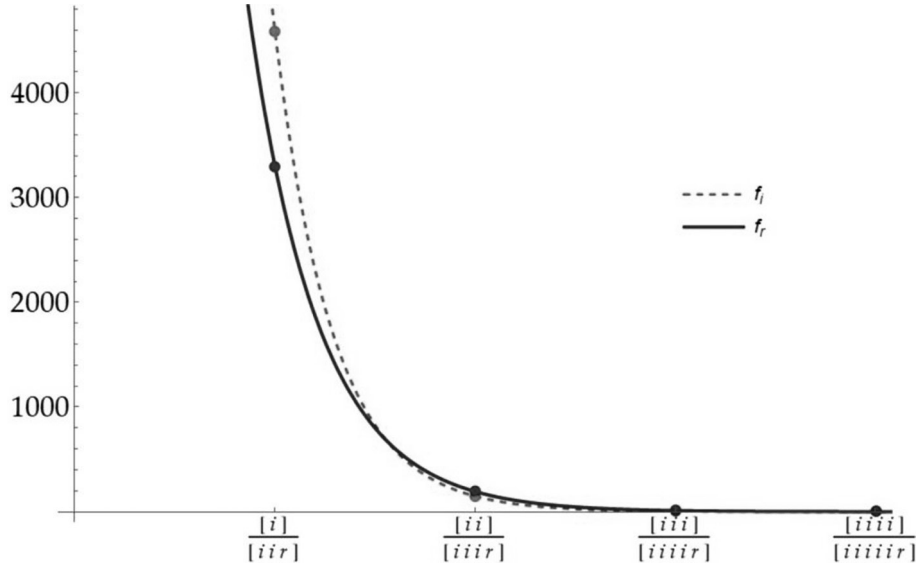

Figure 4. Exponential fit of [i^n] and [$i^{n+1}\ r$].

Note: dashed line for [i^n], and solid line for [$i^{n+1}\ r$].

3.3. Pattern Distribution

In this section, pattern distributions are presented and computed. First, pattern types such as "[*r i r*]" are restored to patterns by replacing "i" with specific function words and replacing "r" with its original POS tag in corpus. Then, 11 pattern types are split into 9,160 specific patterns. Based on those patterns' frequency in the LCMC, the top-36 patterns are listed in Table 4. The numbers of the most frequent 25 patterns are above 1,000, and the numbers of the most frequent 267 patterns are above 100. In this research, only these 267 most frequent patterns are studied, while the rest of the patterns are considered trivial.

Table 4 indicates that the single [*r*] pattern [*v*] comprises the most patterns (161,450) in the pattern set. From Table 4, it can be concluded that the frequency of [v] is larger than the frequency of [n] in written Chinese, which indicates that Chinese is a dynamic language, which is consistent with conclusions from comparative studies of English and Chinese. For example, from the table, it is obvious that patterns with functional word "的", such as "n+的+n", "v+的+n", "n+的+v" rank as the most frequent used patterns in written Chinese, which is consistent with the viewpoints presented in previous studies [43]. Thus, frequent use of functional word "的" is believed to be a typical feature of written Chinese [44], and it is an informational function [45].

Figure 5 presents the distribution of 267 non-trivial Chinese function-word-centered patterns. After data visualization, a similar exponential decay tendency can be seen in Figure 5. Indeed, the graph is well fitted by exp(–0.4t) with R = 0.97.

3.4. Pattern Interactions

In this section, interactions of Chinese pattern units (see Table 5) are discussed with a bigram ($N = 2$) language model. Higher values of N are not studied in this research because the distribution of pattern sets is extremely nonuniform (exponential form, as shown in the last section). Therefore, data sparsity is large at the "tail" part.

Four types of interactions of pattern units (see Table 5) are discussed in this section. These four types come from different combinations of [*r*] and [∃*i*] (which contain function words). Here [∃*i*] represents a pattern which contains at least one function word. Specifically:

1. [*r*]+[*r*]; 2. [*r*] + [∃*i*];
3. [∃*i*] + [*r*]; 4. [∃*i*] + [∃*i*].

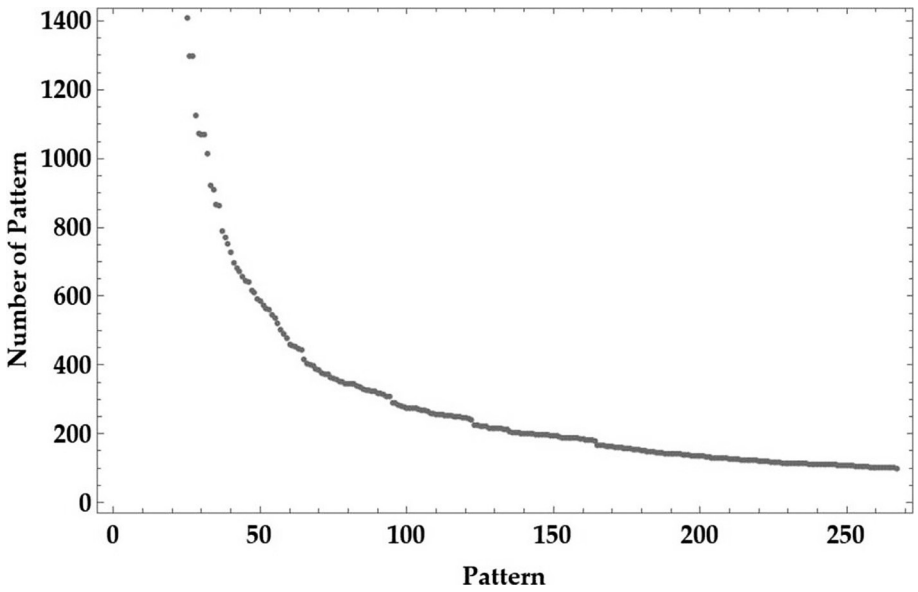

Figure 5. Pattern distribution of the most frequent 267 patterns.

Table 4. Top-36 patterns.

	Pattern	
1	n + 的 + n	8,911
2	的 + n	6,332
3	v + 的 + n	5,816
4	a + 的 + n	4,074
5	n + 的 + v	3,312
6	v + 了 + n	2,427
7	n + 和 + n	2,056
8	r + 的 + n	1,971
9	的 + v	1,966
10	v + 了 + m	1,715
11	v + 的	1,541
12	在 + n	1,527
13	v + 的 + v	868
14	n + 的 + a	864
15	a + 的 + v	789
16	v + 在 + n	770
17	a + 地 + v	754
18	和 + n	727

	Pattern	
19	v + 了 + v	696
20	的+ a	681
21	n + 的	671
22	v + 着 + n	657
23	v + 了 + r	645
24	v + 的 + a	643
25	a + 的	610
26	n + 上	593
27	n + 中	587
28	v + 和 + v	573
29	但 + v	563
30	对 + n	561
31	n + 在 + n	547
32	n + 的 + m	537
33	v + 了 + a	521
34	在 + v	504
35	和 + v	489
36	而 + v	477

Table 5. Top-50 pattern units.

	Pattern unit	
1	v	161,450
2	n	137,475
3	d	55,648
4	r	34,830
5	a	30,917
6	m	29,674
7	q	21,302
8	n + 的 + n	8,911
9	y	6,573
10	t	6,474
11	的 + n	6,332
12	b	5,996
13	v + 的 + n	5,816
14	a + 的 + n	4,074
15	n + 的 + v	3,312
16	x	2,500
17	s	2,457
18	v + 了 + n	2,427
19	n + 和 + n	2,056
20	r + 的 + n	1,971
21	的 + v	1,966
22	v + 了 + m	1,715
23	v + 的	1,541
24	在 + n	1,527
25	k	1,410

	Pattern unit	
26	v + 的 + v	868
27	n + 的 + a	864
28	a + 的 + v	789
29	v + 在 + n	770
30	a + 地 + v	754
31	和 + n	727
32	v + 了 + v	696
33	的+ a	681
34	n + 的	671
35	v + 着 + n	657
36	v + 了 + r	645
37	v + 的 + a	643
38	z	618
39	a + 的	610
40	n + 上	593
41	n + 中	587
42	v + 和 + v	573
43	但 + v	563
44	对 + n	561
45	n + 在 + n	547
46	n + 的 + m	537
47	v + 了 + a	521
48	在 + v	504
49	和 + v	489
50	而 + v	477

3.4.1. Interactions of Content-Word-Centered Patterns

Chinese content patterns in written genres are presented in this section. The bigram language model describes the connections of two nearby pattern units in pattern chains. Since it has already been found that the single [*r*] pattern unit type is the most common pattern unit type in the LCMC corpus, it is natural to suppose that the most frequent pattern connections are in the form [*r*]+[*r*], as presented in Table 6. These pattern connections generally include several content word centered patterns types: v-pattern, n-pattern, d-pattern, a-pattern, r-pattern, m-pattern, and q-pattern in Chinese. Although this research does not focus on the content word pattern, the results are still presented here for the purposes of completeness.

From Table 6, some grammatical and linguistic phenomena in written Chinese can be concluded. For example, "vn" and "vv" take the majority of v-pattern, "nv" and "nn" take the majority of n-pattern, "an" takes the majority of a-pattern, "dv" takes the majority of d-pattern, and "rv" takes the majority of r-pattern. Some of these observations support linguists' intuitive point of view described in previous research, and some complement previous researchers' views with statistical evidence.

3.4.2. Interactions of Function-Word-Centered Patterns

This section examines the interactions of function-word-centered patterns in written Chinese with bigram language model. Specifically, a pattern which contains at least one function word is represented by [∃i]. Equivalently, this notation represents all patterns, except for [r]. Instead of presenting the results as in Table 5, bubble charts (Figures 6–8) are adopted for statistical visualization.

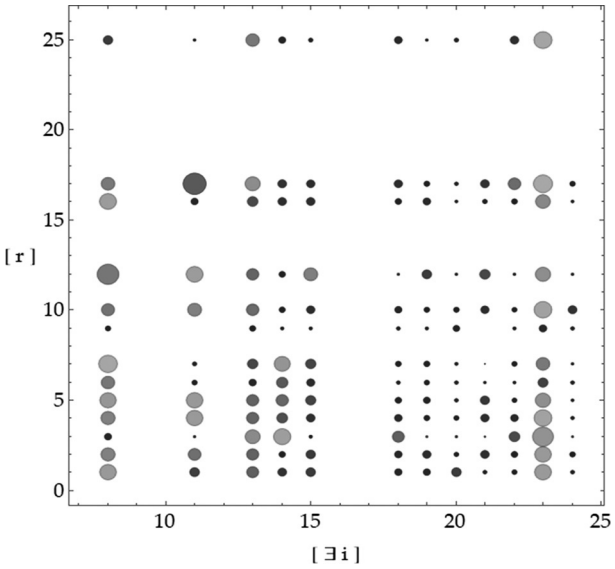

Figure 6. Bubble chart of the language model for [∃*i*]+[*r*].

Table 6. Language model of [*r*]+[*r*].

	v	n	d	r	a	m	q
v	0.1898	0.2378	0.0560	0.0705	0.0468	0.0469	0.0104
n	0.2446	0.1791	0.0919	0.0289	0.0388	0.0249	0.0065
d	0.5634	0.0462	0.1005	0.0177	0.1248	0.0145	0.0055
r	0.2950	0.1819	0.1370	0.0370	0.0463	0.0573	0.0354
a	0.1998	0.2659	0.0532	0.0207	0.0585	0.0230	0.0100
m	0.1204	0.1896	0.0276	0.0139	0.0492	0.0706	0.4101
q	0.1722	0.4060	0.0648	0.0240	0.0850	0.0352	0.0068

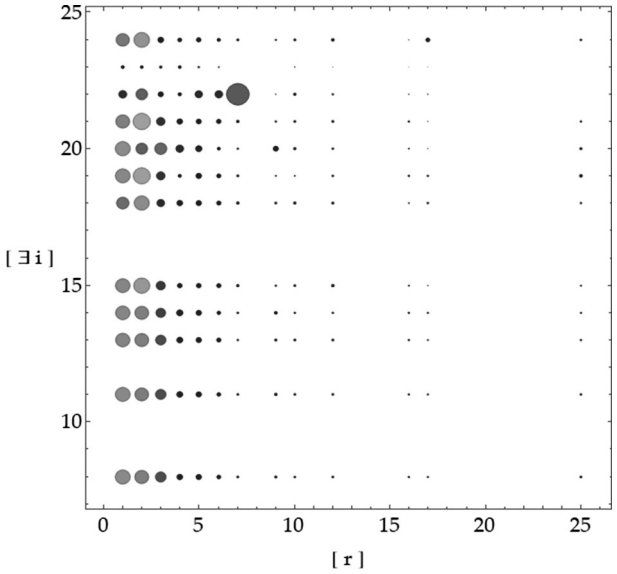

Figure 7. Bubble chart of the language model for [*r*] + [∃*i*].

Both horizontal and vertical axes of these charts represent patterns in the order of Table 5. Every bubble corresponds to a connection of two patterns: $p_y + p_x$, where x and y are horizontal and vertical values, respectively, and p_x and p_y are the x-th and y-th patterns in the pattern set, respectively. The area of each bubble is a quantitative representation of the probability of that connection:

$$A(x, y) = P(p_y + p_x) \tag{15}$$

where $A(x, y)$ is the area of the bubble centered at point (x, y). The color of each bubble is a qualitative representation of the probability of that connection. The color schemes are determined by the default setting of Mathematica's "Rainbow" option. It can be seen that the bubbles with warm color have relatively high probabilities, and vice versa. These three graphs only show the language model of the top-25 pattern units (>1,000).

The non-trivial part of this matrix is visualized by bubble charts, where big bubbles mean high probability and small bubbles represent low probability. The size of the bubble can thus represent the interaction between patterns. There are two extreme cases: (1) if the bubble is large in size, it indicates that the probability equals 1 or almost equals 1, and the restriction is so strong that the corresponding two patterns are bounded together, such as $A(11, 17) = P([s] + [的\ n]) = 0.0269$ in Figure 6, $A(7, 22) = P([v\ 了\ m] + [q]) = 0.5312$ in Figure 7, and $A(11, 18) = P([v\ 了\ n] + [的\ n]) = 0.0725$ in Figure 8, and (2) if a bubble vanishes, it indicates that the restriction between the corresponding patterns is also strong because their connection is not likely to appear in the natural language, such as (11, 9), (21, 9), (21, 25), and (24, 25) in Figure 6; (7, 23), (9, 23), (25, 23), and (25, 22) in Figure 7; and (19, 23), (21, 23), (22, 23), (22, 22), (24, 22), (14, 20), (15, 20), and (22, 19) in Figure 8. Most bubbles ranging between 0 and 1 can be compared relatively by observing bubbles' size. Through observation of the bubble charts, we can identify and discuss the interactions between patterns and determine the most likely interaction and combination of patterns. Only bubbles within one chart can be compared because different charts have different scales, and the bubbles' sizes only have relative meanings.

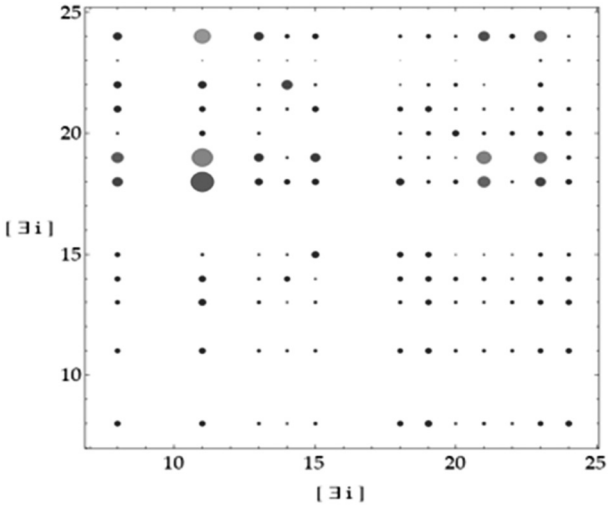

Figure 8. Bubble chart of the language model for [∃*i*] + [∃*i*].

The result also indicates that for [v 的] or [n 的], the words that fill in the blanks will affect the interaction with former patterns. Figure 6 has a relatively small average probability compared to that of the other two graphs. The largest value is only 0.0269 ([s] + [的 n]). Most bubbles

with x = 8 (pairs ending with [n 的 n]) and x = 23 (pairs ending with [v 的]) have larger probabilities. However, for bubbles with x = 13 ([v 的 n]) and x = 15 ([n 的 v]), the probabilities are relatively small.

There is also an evident non-asymmetry between Figures 6 and 7. This means that [*r*] + [∃*i*] models are markedly different from [∃*i*] + [*r*] models. The largest value in Figure 6 is 0.0269; whereas, in Figure 7, most bubbles in the first five columns are greater than 0.03. The comparison also informs us that the probabilities in Figure 6 are more evenly distributed, while the bubbles in Figure 8 become increasingly small when *x* increases. The asymmetry arises from the fact that $C(x_i, x_j) \neq C(x_j, x_i)$. This difference means that [∃*i*] patterns are more likely to appear on the left side of [*r*], rather than on the right side. This finding may bring new insights to the study of written Chinese and might be discussed in future research.

4. Conclusion

In this research, a tentative step is made to combine the N-gram language model and language patterns together for the first time. Specifically, a bigram language model for Chinese function-word-centered patterns is built to study the properties of function words and how function words interact with content words. Besides, in this study, Chinese pattern types are discussed, as well as the distribution of patterns are presented. Major findings reveal the exponential distribution and bounded size of Chinese patterns. The result shows that most patterns with function words exist in the form [r i r]. The number of this type outweighs the sum of other types with function words. It is also found that, except for the [*i r*] pattern, other pattern types which begin with *i* contribute very little to the total number. As for the interaction of patterns, the research found the non-asymmetry between [*r*]+[∃*i*] models and [∃*i*]+[*r*] models, which may be studied in future research activities. At the same time, the research takes into account the most common interactions between patterns units, such as "[s]+[的 n]", "[v 了 m]+[q]", and "[v 了 n]+[的 n]". The results also indicate that for [v 的] or [n 的], the words that fill in the blanks will affect the interaction with former patterns. These findings are consistent with linguistic intuition as well as linguists' observations.

This research fills in the research gap between word/character and sentence, and reveals basic Chinese pattern categories and the interactions between patterns. Thus, it makes a significant contribution to Chinese linguistic research, and improves the efficiency of NLP. We introduced a pattern segmentation rule and the underlying algorithm to extract function-word-centered patterns from the corpus. Some of the research findings support the point of view made in previous research, and some complement previous researchers' views with statistical evidence. Overall, the research confirms that "pattern" can be regarded as a valid linguistic unit in describing and exploring Chinese language. Furthermore, computational techniques open valuable new possibilities for exploring linguistic structures. In future studies, researchers can try to adopt other language modeling methods to explore linguistic rules based on a large corpus.

However, the presented research possesses certain limitations which provide directions for further research. Firstly, the LCMC corpus adopted in this research contains only written Chinese. Further study can also take spoken Chinese into consideration, because written language is different from spoken language in numerous aspects. Secondly, since this research focuses on function words, only function-word-centered patterns are considered. As content word patterns are also vital in Chinese, it is necessary to study the effects of involving content words. Finally, the research adopts only the bigram model for function-word-centered patterns. Higher values of *N* are not studied in this research because the distribution of pattern sets is extremely nonuniform. Further research which is based on a larger corpus may try to apply these methods with higher order N-grams. The study significantly advances understanding of Chinese language features in a data-driven manner.

References

[1] P. K. Nagaraj *et al.*, "Kannada to English Machine Translation Using Deep Neural Network", *Ingénierie des Systèmes d'Information*, vol. 26, no. 1, pp. 123–127, 2021. https://doi.org/10.18280/isi.260113

[2] K. Ravishankar *et al.*, "Floor Segmentation Approach Using FCM and CNN", *Acadlore Transactions on AI and Machine Learning*, vol. 2, no. 1, pp. 33–45, 2023. https://doi.org/10.56578/ataiml020104

[3] S. Zhou *et al.*, "Chinese Text-classification Based on N-gram", *Journal of Chinese Information Processing*, vol. 2001, no. 1, pp. 34–39, 2001.

[4] C. Johnson, "Review of Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English", *Computational Linguistics*, vol. 27, no. 2, pp. 318–320, 2001.

[5] J. Sinclair, "Corpus Concordance Collocation", Oxford: Oxford University Press, 1991.

[6] J. Sinclair, "Trust the Text: Language, Corpus and Discourse", London: Routledge, 2004.

[7] S. Hunston and G. Francis, "Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English", John Benjamins Publishing Company, 2000.

[8] D. Brazil, *The Grammar of Speech*. Oxford: OUP, 1995.

[9] L. Wen and F. Hu, "Some Problems in Chinese Word Order Research", *Studies of the Chinese Language*, vol. 3, pp. 161–165, 1984.

[10] S. Lian, *Comparative Study of English and Chinese*. Beijing: Higher Education Press, 1993.

[11] L. Wang, Chinese Grammar Theory. Beijing: Zhonghua Book Company, 2015.

[12] F. Jelinek, "Language Models and Linguistic Theories Beyond Words", *Nature Machine Intelligence*, vol. 5, pp. 677–678, 2023. https://doi.org/10.1038/s42256-023-00703-8

[13] C. Zong, "Statistical Natural Language Processing (2nd ed.)", Beijing: Tsinghua University Press, 2013.

[14] D. Hiemstra, "Language Models", In: L. Liu and M.T. Özsu (eds) Encyclopedia of Database Systems, Springer, Boston, MA, 2009. https://doi.org/10.1007/978-0-387-39940-9_923

[15] W. Y. Zhang *et al.*, "Improving End-toend Single-channel Multi-talker Speech Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1385–1394, 2020. https://doi.org/10.1109/TASLP.2020.2988423

[16] P. F. Brown *et al.*, "A Statistical Approach to Machine Translation", *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[17] J. Yang *et al.*, "Towards Making the Most of Bert in Neural Machine Translation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9378–9385. https://doi.org/10.1609/aaai.v34i05.6479

[18] J. F. Gao *et al.*, "Improved Source-channel Models for Chinese Word Segmentation", in *Proceedings of ACL, 2003*, Sapporo, Japan, 2003, pp. 272–279. https://doi.org/10.3115/1075096.1075131

[19] J. Zhang, "Integrated Chinese Word Segmentation and Part-of-speech Tagging Model Based on CNN and Bidirectional LSTM", Shanghai Jiao Tong University, 2019.

[20] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing", in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430. https://doi.org/10.3115/1075096.1075150

[21] R. Srihari and C. Baltus, "Combining Statistical and Syntactic Methods in Recognizing Handwritten Sentences", AAAI Symposium: Probabilistic Approaches to Natural Language, pp. 121–127, 1992.

[22] P. F. Brown *et al.*, "A Statistical Approach to Machine Translation", *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[23] R. Cattoni *et al.*, "Robust Analysis of Spoken Input Combining Statistical and Knowledge-based Information Sources", in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'01*. Madonna di Campiglio, Italy, pp. 347–350, 2001. https://doi.org/10.1109/ASRU.2001.1034658

[24] R. Zeng and P. M. Greenfield, "Cultural Evolution over the Last 40 Years in China: Using the Google Ngram Viewer to Study Implications of Social and Political Change for Cultural Values", *International Journal of Psychology*, vol. 50, no. 1, pp. 47–55, 2015. https://doi.org/10.1002/ijop.12125

[25] J. Jiang and Q. Wen, "A Comparative Study on the Effects of N-tuples and Translation Units on English-Chinese Automatic Scoring", *Modern Foreign Languages*, vol. 2010, no. 2, p. 8, 2010.

[26] H. Qin and L. Kong. *Corpra and Contrastive Linguistics.* Beijing: Foreign Language Teaching and Research Press, 2019.

[27] S. Hunston and H. Su, "Patterns, Constructions and Local Grammar: A Case Study of 'evaluation", *Applied Linguistics*, vol. 40, no. 4, pp. 567–593, 2019. https://doi.org/10.1093/applin/amx046

[28] A. S. Hornby, "Guide to Patterns and Usage in English", Oxford University Press ELT, 1954.

[29] J. Sinclair, "Collins cobuild English language dictionary", London Glasgow: Collins, 1987.

[30] J. Collins, "Collins cobuild English grammar", London: Harper Collins, 1990.

[31] J. Collins, "The Collins cobuild English dictionary", London: Harper Collins, 1995.

[32] C. C. Huang *et al.*, "EdIt: A Broad-coverage Grammar Checker Using Pattern Grammar", in *Proceedings of the ACL-HLT 2011 System Demonstrations*, 2011.

[33] G. Chen and M. Liang, "The Origin, Features and Applications of Pattern Grammar", *Foreign Language Research*, vol. 1, pp. 17–24, 2017. https://www.doi.org/10.16263/j.cnki.23-1071/h.2017.01.004

[34] S. C. Xiong, "A Corpus-based Approach to the Interaction of English Verb Patterns with 'it' and Registers", Zhejiang University, 2014.

[35] T. Yu, "Automatic Identification and Extraction of English Verb Patterns: A Study Based on the Clustering of Concordance", Beijing: Foreign Language Teaching and Research Press, 2018.

[36] Y. Wang, "On the Borderline Between Grammar and Lexis: A Review of Pattern Grammar", *Contemporary Linguistics*, vol. 10, no. 3, pp. 257–266, 2008.

[37] G. Chen and M. Liang, "Automatic Detection of Verb form Errors in Chinese EFL Learners' Written English-A Study Based on Link Grammar", *Journal of Chinese Information Processing*, vol. 31, no. 6, pp. 196–204, 2017.

[38] J. Lu and Z. Ma, *On Function Words in Modern Chinese* (Revised Edition). Beijing: Language & Culture Press, 2003.

[39] T. McEnery and R. Xiao, "The lancaster corpus of mandarin Chinese (LCMC)", Lancaster: Lancaster University, 2004.

[40] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999. https://doi.org/10.1006/csla.1999.0128

[41] H. Ney *et al.*, "On Structuring Probabilistic Dependences in Stochastic Language Modelling", *Computer Speech & Language*, vol. 8, no. 1, pp. 1–38, 1994. https://doi.org/10.1006/csla.1994.1001

[42] R. Kneser and H. Ney, "Improved Backing-off for m-gram Language Modeling", in *Proc. of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, USA, vol. 1, 1995, pp. 181–184. https://doi.org/10.1109/ICASSP.1995.479394

[43] H. Yang, The Europeanization in Modern Chinese, 2008, Beijing: Commercial Press

[44] Y. Peng, "A study on Chinese lexial bundles in written and spoken genre", Zhejiang University, 2017.

[45] J. Song *et al.*, "A Multi-dimensional Approach to Register Variations in Mandarin Chinese", *Glottometrics*, vol. 51, pp. 39–71, 2021. https://doi.org/10.53482/2021_51_393

Received: September 2023
Revised: October 2023
Accepted: October 2023

Contact addresses:
Jie Song
School of Foreign Languages
Zhejiang University of Finance & Economics
Hangzhou
China
e-mail: 20220102@zufe.edu.cn

Yixiao Liu
School of International Studies
Zhejiang University
Hangzhou
China
e-mail: 1010369037@qq.com

Yunhua Qu*
School of International Studies
Zhejiang University
Hangzhou
China
e-mail: qu163hua@163.com
*Corresponding author

Jie Song received her PhD degree in Applied Linguistics from the Zhejiang University, Hangzhou, in 2022. Since 2022, she has been a lecturer in the School of Foreign Languages, Zhejiang University of Finance & Economics, in Hangzhou, China. Her research interests include corpus linguistics, pattern grammar, and register studies.

Yixiao Liu received a double BSc degree in English and Physics from Zhejiang University, Hangzhou, China, in 2018 and the MSc degree in physics from Zhejiang University, Hangzhou, China, in 2021. His research interests include pattern grammar, and natural language processing.

Yunhua Qu received the MSc degrees in applied linguistics from Zhejiang University, Hangzhou and PhD degree in computational linguistics in from Communication University of China, Beijing. She has been a professor with the School of International Studies, Zhejiang University. She is the author of two books, and more than 50 articles. Her research interests include corpus linguistics, pattern grammar, and register studies.