

Readability Assessment for Chinese L2 Sentences: An Extended Knowledge base and Comprehensive Evaluation Model-based Method

Shuqin ZHU*, Jihua SONG*, Weiming PENG, Jingbo SUN

Abstract: The paper assesses sentence readability based on the standards in the field of Chinese L2 teaching. In view of the inapplicability of the field standards in text readability assessment, the study focuses on two aspects. On the one hand, the graded lexicon of the HSK syllabus is extended to obtain a large-scale graded lexical knowledge base. On the other hand, sentence-based features in the existing teaching grammatical knowledge base are supplemented to achieve the automatic recognition of grammatical points and obtain quantitative grammatical indicators regarding sentence readability. Besides, based on the extended knowledge bases, comprehensive evaluation models are created to calculate the lexical and grammatical difficulties of sentences, as well as the sentence readability. The results of experiments show that the sentence readability is well differentiated in all levels of texts. Furthermore, the correlation between sentence readability and text readability is significantly improved in comparison with existing methods.

Keywords: Chinese L2 teaching; comprehensive evaluation model; extended knowledge base; sentence readability

1 INTRODUCTION

Reading is the process of obtaining information from visual materials, i.e., using text to acquire information, understand the world, develop thinking, and obtain esthetic experience. As one of the four basic learning skills (listening, speaking, reading, and writing), reading plays an important role in Chinese L2 teaching [1]. Given the prominent position of reading in language learning, the provision of suitable reading materials for learners has become one of the key issues in Chinese L2 teaching. According to Vygotsky's zone of proximal development [2] and Krashen's comprehensible input [3], the level of reading materials provided for L2 learners should be slightly higher than the current level of the learners. However, for L2 teaching staff, it takes considerable energy and time to select reading materials with appropriate difficulty from a large number of teaching resources. Hence, an automatic assessment of text readability is urgently required.

The text representation methods and machine learning methods in natural language processing technology can complete the general assessment of text readability, but they cannot explain the details of text readability and lack interpretability which is necessary for L2 teaching. L2 teaching is different from mother-tongue teaching. L2 learners are affected by their mother tongue in terms of sentence comprehension and other aspects. Therefore, the readability features of Chinese L2 texts should consider the characteristics of L2 acquisition [4]. The relevant teaching and examination syllabuses in the Chinese L2 teaching follow the rule of L2 acquisition, which are the reference standards for L2 teaching and the important bases for the readability assessment of Chinese L2 texts. However, owing to the difficulty of automatically identifying grammatical points [5] and the limited number of words in the graded lexicon [6], these syllabuses are not applicable when applied directly to the task of readability assessments.

Presently, studies on the readability of Chinese L2 texts mainly focus on the quantitative analysis of textbooks or the readability analysis of texts in chapter units from a

macroscopic perspective, and few studies analyze the sentence readability [7, 8]. Meanwhile, sentences are an important part of the reading aspect of L2 teaching, and sentence readability is an important part of text readability [9]. The measurement of sentence readability can assist teachers to select difficult sentences to focus on during the teaching process or adapt and downgrade difficult sentences to suit the level of L2 learners. The assessment of sentence readability enables the text readability to be analyzed from a more microscopic and finer perspective.

In this study, we assess the readability of Chinese L2 sentences and extend the graded lexicon of the Chinese Proficiency Test (HSK) syllabus, which is a standard in the field of Chinese L2 teaching, to obtain a more practical graded lexical knowledge base for text readability assessments. Besides, we improve the existing grammatical knowledge base of Chinese L2 teaching [10] for the automatic recognition of grammatical points in sentences, and extract and quantify the relevant indicators of sentence grammatical knowledge. Text readability assessment is a function that maps the text to a value. The input variable of the function is a set of readability features, and the output variable is usually a readability level or score, which indicates a comprehensive evaluation. In other words, to establish an indicator system for the evaluated object, certain methods or models are used to analyze the collected data, and an overall quantitative judgment for the evaluated object is made [11]. Based on the extended knowledge bases, comprehensive evaluation models are built in this study to calculate the lexical and grammatical difficulties of sentences, as well as sentence readability.

The following parts of the paper are arranged as follows: In Section 2, related works and existing problems on sentence readability assessment and Chinese L2 sentence readability assessment are introduced. In Section 3, the extended methods of domain knowledge bases and the assessment method of sentence readability are introduced. The experimental results are introduced and analyzed in Section 4. Finally, the conclusion and limitations of this study are discussed in Section 5.

2 RELATED WORKS

Presently, most researches on text readability are based on chapters. The most representative approach is to evaluate chapter texts using readability formulas, e.g. Flesch-Kincaid Grade Level [12, 13], Gunning Fog Formula [14, 15], SMOG Formula [16] and other related formulas [17, 18]. With recent advances in technology, researchers have applied machine learning methods to text readability assessment. For example, in [19], the authors discussed the application of decision tree, bagging decision tree, linear regression, SVM, and NB methods in text readability assessment.

Sentence is the basic language unit in language learning, and sentence readability assessment is a fine-grained text readability assessment. Current research on sentence readability mainly focuses on the analysis of sentence complexity regarding structure or syntax. Meanwhile, the measurement indicators of syntactic complexity have advanced from coarse-grained indicators based on length [20] to indicators based on clause complexity [21, 22] to indicators based on phrase complexity [23, 24].

In the cognitive study of L2 learning, it is generally accepted that language knowledge, such as lexical knowledge or grammatical knowledge, is a significant factor that affects reading comprehension [25]. Meanwhile, lexical knowledge is considered to be the best factor for reading measurement, as the correlation between lexical knowledge and reading comprehension is 0.5 - 0.85 [26]. The success of reading comprehension significantly depends on the reader's mastery of the words in the text. Thus, lexical knowledge is closely related to reading comprehension. In a study on L2 sentence readability assessment [27], the authors discussed four categories of features that affect the readability of Swedish sentences, namely, traditional features, syntactic features, lexical morphological features, and semantic features. They found that lexical morphological features (the percentage of B1 words in a sentence) had the greatest impact on the readability of L2 sentences. In addition to lexical knowledge, grammatical knowledge is also very important for text comprehension because it is crucial to the construction of discourse coherence and integration [25, 28, 29]. According to the Common Core State Standards, grammatical knowledge can aid reading comprehension and interpretation. When lexical information exists but lacks the necessary grammatical clues, reading comprehension becomes impossible. Because of the importance of lexical and grammatical knowledge in reading comprehension, it is feasible to predict sentence readability based on lexical and grammatical difficulties.

The study of Chinese sentence readability is still in its infancy. In [30], the author determined the sentence difficulty based on the number of lexical semantic categories in a sentence and determined that sentences with more semantic categories are relatively difficult to understand. In another study [31], the author deduced the main factors that affect sentence comprehension from a questionnaire and constructed two sentence readability formulas for primary school students and international students using linear regression. Meanwhile, based on the field standards of Chinese L2, Wanghao deduced quantitative indicators at the character and word aspects and used the CRITIC method to establish a relationship between sentence readability, character difficulty, and

vocabulary difficulty [32]. Dong et al. extracted sentence difficulty features from three aspects, namely, character, word, and syntax, and used a machine learning classification algorithm to predict the sentence absolute and relative difficulties [33].

In the above studies, the application of the field standards in Chinese L2 teaching is not in-depth. Particularly, the application of lexical and grammatical knowledge in the field standards to the study of sentence readability is still inadequate, and the grading syllabus of Chinese L2 teaching is not applicable for assessing text readability. Additionally, the readability assessment of Chinese L2 texts currently uses readability formulas and traditional machine learning algorithms. Therefore, the study of sentence readability assessment needs to be improved and supplemented in several aspects.

3 METHOD

3.1 Extension of the Graded Lexical Knowledge Base

HSK is a standardized international Chinese proficiency test for non-native Chinese speakers (including foreigners, overseas Chinese, and Chinese ethnic minority candidates). HSK syllabus has a complete rating system, with six levels from low to high. There are 5000 words in the graded lexicon of the HSK syllabus. Besides, there are some exemplary words in the syllabus. Although these words are not listed in the graded lexicon, they are graded with additional explanations. By sorting all the words, a basic word list comprising 5650 words was obtained. When analyzing sentences or texts based on this list, there will be a large number of pseudo outline words, which will affect the assessment of text readability or sentence complexity. When learners master the words, they can learn other words by analogy. For example, when learners master "电影院 (cinema)" in the list, they can learn "电影 (movie)" by analogy. In fact, 650 additional words in the HSK syllabus were derived from the extension of the 5000 basic words. Meanwhile, there are two methods for extending the words. One method is to split them, for example, "表格 (table)" belongs to level 4, and "表" is also level 4 after splitting. The other method is to combine the words, for example, "游客 (tourist)" is obtained by combining "旅游 (tourism)" and "客人 (guest)" where "客人 (guest)" belongs to level 3 and "旅游 (tourism)" belongs to level 2. The higher level is considered as the level of the combined word; hence, "游客 (tourist)" belongs to level 3.

Based on the extension method of the HSK syllabus, we adopted split and combination analogies to extend the word list, and two basic principles for extending the word list are given below:

(1) Follow the word formation, for example, "火车站 (railway station)" is a combination word comprising "火车 (train)" and "站 (station)". The structural relationship between them is centering, i.e., the former word describes or restricts the latter word. When a one-layer split is performed, only "火车 (train)" and "站 (station)" can be obtained by following the word formation, and no wrong splits, such as "火 (fire)" and "车站 (station)", is obtained. Some examples of split methods are listed in Tab. 1.

Table 1 Follow the word formation

Word	Word formation	Structural relationship	Split method
董事长 (chairman)	<n mod = "n27n"><n sen = "001">董事</n><n sen = "108">长</n></n>	定中 (Centering)	董事/长
飞禽走兽 (birds and animals)	<n mod = "n2...n2"><n sen = "001">飞禽</n><n sen = "001">走兽</n></n>	并列(Juxtaposition)	飞禽/走兽
没关系 (no problem)	<v mod = "v n2"><v sen = "001">没</v><n sen = "003">关系</n></v>	动宾 (Verb-object)	没/关系
面条儿 (noodles)	<n mod = "n7n-u"><n sen = "101">面</n><n sen = "002">条</n><u sen = "101">儿</u></n>	其它 (Other)	面条/儿
归根到底 (in the final analysis)	<c mod = "v2←v2"><v sen = "001">归根</v><v sen = "001">到底</v></c>	述补 (Verb-complement)	归根/到底

(2) Follow the semantic consistency. In this process, the meaning of morphemes does not change when they form a word, and the meaning of the word is the combination of morpheme meanings. However, the word cannot be split when the way of word meaning combination is a general reference, loss of reference, and metonymy, among others, (as listed in Tab. 2). For example, the word formation of "平白 (gratuitously)" is "

平...白" (<a>平白) and "平 (flat)" and "白 (white)" are word-forming morphemes. The semantic of "平白 (gratuitously)" is not a combination of morpheme meanings but a metaphorical extension. Another example is "人7事" which refers to something special. The semantic changes when the word is split into "人 (person)" and "事 (thing)".

Table 2 Follow the semantic consistency

Word	Word formation	The way of word meaning combination
窗户 (windows)	<n mod = "n...n"><n sen = "001">窗</n><n sen = "001">户</n></n>	失指 (loss of reference)
保驾 (escort)	<v mod = "v n"><v sen = "001">保</v><n sen = "004">驾</n></v>	泛指 (general reference)
爱人 (lover)	<n mod = "v7n"><v sen = "001">爱</v><n sen = "001">人</n></n>	特指 (special reference)
平白 (gratuitously)	<a>平白	喻指 (metonymy)
人事 (human affairs)	<n mod = "n7n"><n sen = "001">人</n><n sen = "001">事</n></n>	代指 (Anaphora)

The extension process of the graded lexicon is shown in Fig. 1.

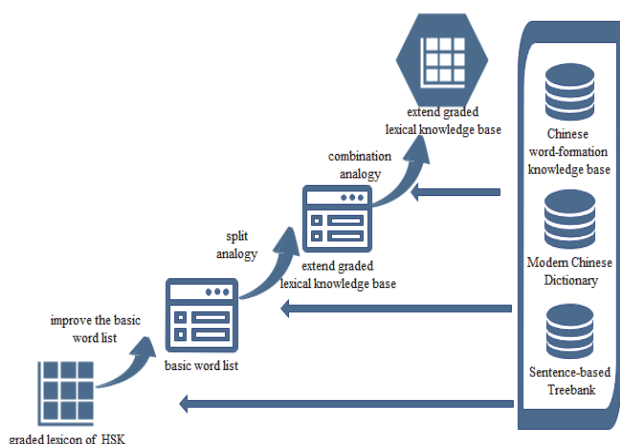


Figure 1 The extension process of the graded lexical knowledge base

By improving the basic word list, split analogy, and combination analogy, an extended graded lexical knowledge base is obtained. Each step requires a sentence-based tree bank [34], Chinese word-formation knowledge base [35], and Modern Chinese Dictionary. The sentence-based tree bank and word-formation knowledge base are created by the Language and Character Resource Research

Center of Beijing Normal University. The sentence-based tree bank includes Chinese mother-tongue teaching materials, international Chinese teaching materials, and literary literature. In this study, the corpus of international Chinese teaching materials in the tree bank was used. The sentence-based tree bank uses the Modern Chinese Dictionary to tag the meaning items and POS of each word. The frequency of different POS and meaning items of each word were obtained by the statistical analysis of the tagged corpus. Some POS and meaning items of words appeared more frequently in the corpus, while others appeared less frequently. Generally, the high-frequency POS and meaning terms of words need to be mastered by learners. Meanwhile, Chinese word-formation knowledge base mainly comprises POS, meaning items, structural relation, word formation, and word meaning combinations.

3.1.1 Improvement of the Basic Word List

The basic word list of the HSK syllabus includes the POS, level, and analogy basis of each word. To improve the basic word list, values were assigned to the attributes of each word in the list, including meaning item, word formation, and the way of word meaning combination, for the subsequent analogies. The main process of improving the basic word list is as follows:

(1) Both the word formation and the way of word meaning combination are obtained from the Chinese word-formation knowledge base. First, based on the sentence-based treebank, the word frequency is supplemented for each word in the word-formation knowledge base. Meanwhile, the frequency of the meaning item of each word rather than the overall frequency of the word is counted based on the sentence-based tree bank. Owing to the particularity of the field, some meaning items do not appear in the international Chinese teaching corpus.

(2) Based on the Chinese word-formation knowledge base after supplementary information, the information of the basic word list is improved. If a word in the Chinese word-formation knowledge base has only one POS, the corresponding information of the record with the highest frequency of meaning items is considered as the attribute of the word in the basic word list. However, if the word has more than one POS in the Chinese word-formation knowledge base, the corresponding information of the

most frequent meaning item with the same POS is added to the basic word list.

Some examples of the basic word list after obtaining the attributes are summarized in Tab. 3, where meaning code refers to the meaning code of words in the Modern Chinese Dictionary. For example, "爱好 (hobby)" is a noun (n) with a meaning code of 002, which means "strong interest in something". In the word formation, "爱好 (hobby)" is composed of two juxtaposed verbs (<n mod = "...v">), in which the meaning code of the first morpheme "爱 (love)" is 002 and that of the second morpheme "好 (like)" is 101. The transferred meaning of 0 indicates that the morpheme meaning has not changed after the word formation. "爱好 (hobby)" corresponds to two records in the Chinese word-formation knowledge base: noun and verb. In the basic word list, "爱好 (hobby)" is a noun. Therefore we added the information corresponding to the noun in the Chinese word-formation knowledge base to "爱好 (hobby)".

Table 3 Examples of the basic word list after obtaining the attributes

Word	POS	Level	Meaning code	Word formation	Transferred meaning
爱好 (hobby)	n	3	002	<n mod = "...v"><v sen = "002">爱</v><v sen = "101">好</v></n>	0
爱护 (care)	v	5	001	<v mod = "...v"><v sen = "001">爱</v><v sen = "001">护</v></v>	0
爱心 (love)	n	5	001	<n mod = "v7n"><v sen = "001">爱</v><n sen = "002">心</n></n>	0
暧昧 (ambiguous)	a	6	002	暖昧	4
安静 (quiet)	a	3	001	安静	0

3.1.2 Split Analogy

Split analogy is used to split words in the basic word list to obtain their morphemes and their structural relationships based on the word formation. If the meanings of the morphemes do not change when they form words, the words are split according to the structural relationship such that the level of the morphemes is the same as that of the original word. In this way, more morphemes with levels can be obtained. Three methods are used for split analogy:

level-by-level split, iterative split, and only split one level. The specific process is shown in Fig. 2.

(1) Level-by-level split traverses the words of level 1 and determines whether they can be split according to the word formation and the transferred meaning. If the words can be split, only one layer is split for each word. After completion, it traverses the words of level 2 and splits them. The same process is repeated for the other levels in sequence.

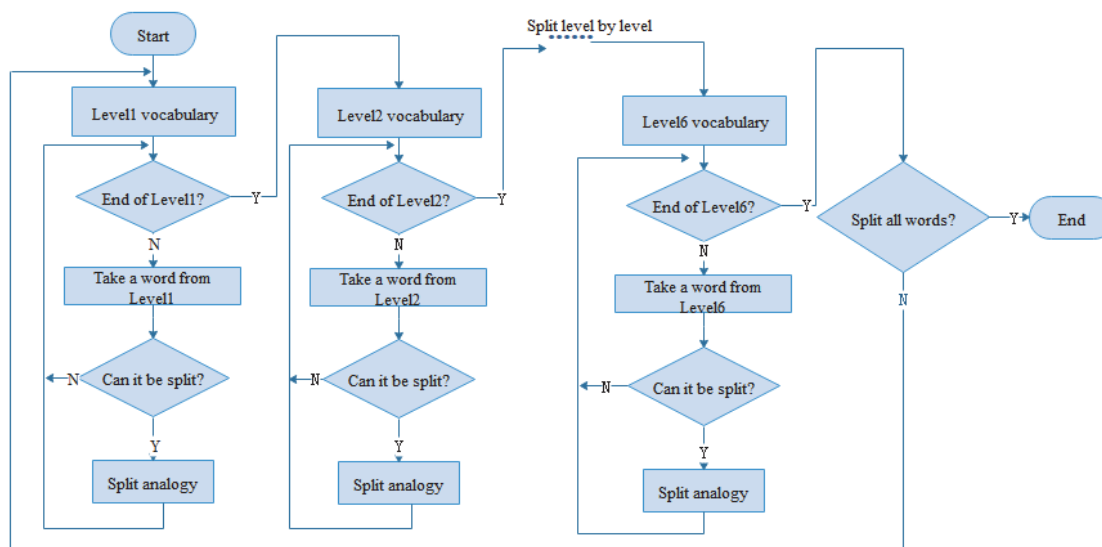


Figure 2 Process of split analogy

(2) After the level-by-level split is completed, the split is restarted from the first level of words and iteratively loops until there are no words that can be split.

(3) Some words are complex in the word formation and can be nested. For example, "出租车 (taxi)" is the centering structure, and "出租 (rent)" is a word whose word formation is $\langle v \text{ mod} = "v \rightarrow v" \rangle \langle v \text{ sen} = "001" \rangle$ 出 $\langle /v \rangle \langle v \text{ sen} = "002" \rangle$ 租 $\langle /v \rangle \langle /v \rangle$. "Only split one layer" means that "出租车 (taxi)" will only be split into "出租 (rent)" and "车 (car)", and not into three separate morphemes "出(out)", "租(rent)" and "车 (car)". In other words, the morpheme "出租 (rent)" will not be split. However, during the iterative split, "出租 (rent)" will be split into two morphemes "出 (out)" and "租 (rent)". The method of splitting only one layer is used because internal morphemes may be analogized from other words. This method conforms to the top-down and gradually refined cognitive style.

After a word is split, the pairs of words and meaning codes are traversed. If a pair of words and meaning code already exists in the lexical knowledge base, it will not be added; otherwise, it will be added.

3.1.3 Combination Analogy

Combinational analogy theoretically performs various permutations and combinations based on the extended lexical knowledge base (derived from split analogy) and analyzes the rationality and level of words. However, the permutations and combinations have a large order of magnitude, resulting in a relatively high probability of not forming a word. Moreover, some words lack word formation; hence, it is impossible to determine whether they can be combined.

Therefore, combination analogy should first be used to obtain a large-scale vocabulary with word formations and transferred meanings and then analyze the level of words in the large-scale vocabulary. The large-scale vocabulary used in this study was extracted from the XML data of the sentence-based tree bank and Chinese word-formation knowledge base. It comprises 118400 vocabularies, including word, meaning code, POS, word formation, and transferred meaning.

Combination analogy was used to analyze the 118400 words. If the morphemes do not have transferred meanings when they form words, the word was split according to the word formation to obtain the morpheme pairs and the meaning codes. Afterwards, each pair of morphemes and meaning code was checked whether they exist in the extended lexical knowledge base obtained by the split analogy. If each pair of morphemes and meaning code exist, it indicates that each morpheme has a level information. The word was then added to the extended lexical knowledge base, and the level of the word is the maximum value of each morpheme level.

3.2 Improvement of the Teaching Grammatical Knowledge Base

The Language and Character Resource Center of Beijing Normal University created a grammatical knowledge base for Chinese L2 teaching based on the HSK syllabus, the Grading Syllabus, and the Practical Chinese Grammar for Foreigners. The grammatical points of the knowledge base are divided into nine categories, namely: morpheme, content word, function word, phrase, sentence component, single sentence, sentence pattern, tense and complex sentence. According to the HSK syllabus, the levels of grammatical points are marked manually.

Table 4 Sentence-based features

Category	Type	Grammatical points	Sentence-based features	Remarks
Word	Verb	V 了 - V	//v[@mod = "v·u·m·v" and u = "了" and m = "-"]	Describe with POS, morphology, and words
Phrase	Coordinate phrase	联合短语 (coordinate phrase)	//cc[@fun = "COO"] //*[contains(@mod, "...") and substring(@mod,4,1) = "..."]	Describe with syntax Describe with morphology
Sentence component	Complement	简单趋向补语与宾语的相对位置 (The relative position of simple directional complement and object)	//prd[@scp = "VO" and v[(@mod = "v←v-u" or @mod = "v2←v-u" or @mod = "v←v" or @mod = "v2←v") and (*[2] = "来" or *[2] = "去" or *[2] = "上" or *[2] = "下" or *[2] = "进" or *[2] = "出" or *[2] = "回" or *[2] = "过" or *[2] = "开" or *[2] = "起")]] //*[prd[@scp = "VOC"]/v[not(@mod) or @mod = "v-u"] and cmp[v = "来" or v = "去"]]	Description with syntax, POS, and morphology
Single sentence	Sentence class	用“几、多少”提问 (ask questions with “how many”)	//r[(= "多少" and @sen = "001") or (= "几" and @sen = "201")]	Describe with POS, words, and word meaning
Sentence pattern	“把” sentence	后接结果补语 (动词/形容词) (Followed by result complement (verb/adjective))	//*[adv/pp[p = "把"] and prd/v[*[2] = v or *[2] = a]]	Description with syntax, POS, and morphology
Compound sentence	Progressive complex sentence	不但……, 而且……(not only……,but also……)	/ju[xj/cc/c[= "不但"] and xj/cc/c[= "而且"]]	Describe directly with POS and morphology

Different grammatical points correspond to different text features. In [10], the authors added text features to grammatical points and used a regular expression pattern matching method to find the grammatical points contained in the text. However, grammatical points, such as word formation, POS, phrase type, and sentence composition involve deep syntactic structure information and cannot be recognized by regular matching from the character surface. Therefore, as a supplement to the above work, we added sentence-based features to the grammatical points based on XPath (XPath finds specific information by traversing the elements and attributes of an XML document) and attempted to automatically identify the grammatical points contained in the sentence from the corpus of the sentence-based tree bank. Some examples of sentence-based features represented by XPath are given in Tab. 4. Each grammatical point is described by word, POS, morphology, word meaning, syntax, or other relevant contents based on specific situations.

3.3 Measurement of Sentence Readability

In this paper, readability assessment is regarded as a comprehensive evaluation problem, and the entropy weight models were created to measure lexical difficulty, grammatical difficulty and sentence readability.

3.3.1 Lexical Difficulty

(1) The matrix WL was obtained by counting the number of words at each level in each sentence.

$$WL = (wl_{ij})_{m \times 7} \quad (1)$$

where wl_{i1} to wl_{i7} represent the number of first-level, second-level, third-level, fourth-level, fifth-level, sixth-level, and outlined words, respectively.

(2) The indicators in the matrix were standardized. The more the number of first-level words, the less the difficulty of the text, which belongs to a negative index. The standardization formula of wl_{i1} is given by Eq. (2), and the standardization formula of other indices is given by Eq. (3).

$$wl_{i1} = \frac{\max\{wl_1\} - wl_{i1}}{\max\{wl_1\} - \min\{wl_1\}} \quad (2)$$

$$wl_{ij} = \frac{wl_{ij} - \min\{wl_j\}}{\max\{wl_j\} - \min\{wl_j\}} \quad (3)$$

(3) The contribution of the i -th sentence to the j -th indicator was calculated using:

$$p_{ij} = \frac{wl_{ij}}{\sum_{i=1}^m wl_{ij}} \quad (4)$$

(4) The total contribution of all the sentences to the j -th indicator, i.e., the information entropy of the j -th indicator, was calculated using:

$$e_j = -K \sum_{i=1}^m p_{ij} \ln p_{ij} \quad (5)$$

where $K = 1 / \ln m$.

(5) The redundancy of the information entropy was calculated using:

$$d_j = 1 - e_j \quad (6)$$

(6) The weight of each indicator in the lexical difficulty was calculated using:

$$w_j = d_j / \sum_{j=1}^7 d_j \quad (7)$$

(7) The lexical difficulty of the i -th sentence was represented using:

$$wc_i = \sum_{j=1}^7 w_j \cdot wl_{ij} \quad (8)$$

3.3.2 Grammatical Difficulty

(1) The matrix GL was obtained by counting the number of grammatical points at each level in each sentence.

$$GL = (gl_{ij})_{m \times 7} \quad (9)$$

where gl_{i1} to gl_{i7} represent the number of first-level, second-level, third-level, fourth-level, fifth-level, sixth-level, and outlined grammatical points, respectively.

The other steps were the same as those of the lexical difficulty. Thus, the grammatical difficulty of the i -th sentence is given as:

$$gc_i = \sum_{j=1}^7 w_j \cdot gl_{ij} \quad (10)$$

3.3.3 Sentence Readability

The matrix $SWGL$ was obtained by calculating the lexical and grammatical difficulties of each sentence.

$$SWGL = (wc_i, gc_i)_{m \times 2} \quad (11)$$

where wc_i and gc_i represent the lexical and grammatical difficulties of the i -th sentence, respectively.

The indicators in the matrix were standardized using:

$$wc_i = \frac{wc_i - \min\{wc\}}{\max\{wc\} - \min\{wc\}} \quad (12)$$

$$gc_i = \frac{gc_i - \min\{gc\}}{\max\{gc\} - \min\{gc\}}$$

After standardization, the weight of each indicator ($w_i (i = 1, 2)$) in the sentence difficulty was calculated based

on the entropy weight method, and the sentence readability (sc_i) was given as:

$$sc_i = w_1 \cdot wc_i + w_2 \cdot gc_i \quad (13)$$

The entropy weight represents the differentiation degree of an indicator to the evaluation object and does not reflect the importance of the indicator. Hence, unreasonable weights may appear. The comprehensive weight of the indicators was obtained by combining the importance and entropy weights, and it is given by:

$$\beta_i = \frac{\alpha_i w_i}{\sum_{i=1}^2 \alpha_i w_i} \quad (i=1, 2) \quad (14)$$

where $\alpha_i (i = 1, 2)$ represents the importance weight of the indicators, which will be determined through experiments.

Finally, the sentence readability (sc_i) was assessed using the comprehensive weight, and it is given by:

$$sc_i = \beta_1 \cdot wc_i + \beta_2 \cdot gc_i \quad (15)$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental data

The experimental data were from Boya Chinese textbooks in the sentence-based tree bank. The textbooks have nine volumes which are divided into three levels: primary, intermediate, and advanced. The sentences in each level of the text are listed in Tab. 5.

Table 5 Experimental data

Level	Chapter	Sentence	Average number of characters in a sentence	Average number of words in a sentence
Primary	66	1075	12,24	7,46
Intermediate	67	2549	24,77	14,52
Advanced	48	2942	27,00	15,04

4.2 Extended Graded Lexical Knowledgebase

After the split and combination analogies were performed, the final extended lexical knowledge base comprised 43892 words. The number of words in all the levels is listed in Tab. 6.

Table 6 Graded lexical knowledge base

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Total
598	982	2400	6282	13144	20486	43892

Based on the graded lexicon of the HSK syllabus and the extended lexical knowledge base, we performed

a statistical analysis of words in the sentences of different levels of the texts, and the results are summarized in Tab. 7. The proportion of outline words in the sentences at all levels of the texts based on the extended lexical knowledge base is significantly reduced, which makes the sentence readability more accurately reflected by lexicon. Additionally, the proportion of outline words in the sentences of the primary text is the smallest, whereas the proportion of outline words in the sentences of the advanced text is the largest. This indicates that the analogy process is reasonable and consistent with cognitive characteristics.

Table 7 The distribution of words in the sentences of different levels of texts

Vocabulary	Level	Level 1 / %	Level 2 / %	Level 3 / %	Level 4 / %	Level 5 / %	Level 6 / %	Outline / %
Graded lexicon of HSK	Primary	56,49	13,52	11,13	7,38	2,30	0,50	7,83
	Intermediate	41,45	12,53	11,41	10,06	6,43	2,52	15,48
	Advanced	35,91	10,10	10,18	11,88	8,62	4,51	18,60
Extended lexical knowledgebase	Primary	58,08	14,13	11,91	8,10	2,89	0,87	3,18
	Intermediate	42,88	13,44	12,57	11,84	8,49	4,39	6,28
	Advanced	36,88	10,79	11,19	14,29	11,33	7,51	7,81

To test the discrimination of the extended lexicon to the sentences of different levels of texts, the distribution difference of words in the sentences of the adjacent level text was calculated based on relative entropy. For the probability distribution of words between texts of two adjacent different levels, the relative entropy can be expressed as:

$$KL(P, Q) = \left(\sum_{i=1}^7 p_i \log \frac{p_i}{q_i} + \sum_{i=1}^7 q_i \log \frac{q_i}{p_i} \right) / 2 \quad (16)$$

where $i = 1, 2, \dots, 6, 7$ represent the corresponding words of levels 1...6 and the outline words, p_i and q_i indicate the distribution of words in level i in different levels of the text, and the base number of the log is e . The results are summarized in Tab. 8. Compared with the graded lexicon

of HSK syllabus, the extended graded lexicon better distinguished primary and intermediate text sentences, as well as intermediate and advanced text sentences, and it can better reflect sentence difficulty from the lexicon. The reason is that when analyzing sentences based on the extended lexical knowledge base, the proportion of pseudo outline words is significantly reduced, and each meaning item of the words is assigned a different difficulty level. For example, when the meaning of "转机 (transfer)" is "好转的机会 (the opportunity to improve)", the level of difficulty is 6. The level of difficulty of interpreting "转机 (transfer)" as "乘飞机中途转乘其他飞机 (transfer to another plane)" is 4, which makes the text difficulty more detailed.

Table 8 Differences of word distribution in the sentences of adjacent level texts

Level	Graded lexicon of HSK syllabus	Extended graded lexical knowledgebase
Primary and intermediate	0,0396919	0,043331
Intermediate and advanced	0,0089794	0,010716

4.3 Lexical Difficulty

Based on the representation of lexical difficulty model, the number of vocabularies in each level in each sentence was counted. After standardization, the contribution of

each indicator was calculated, and the information entropy, redundancy, and weight of each indicator were obtained. The results are summarized in Tab. 9, in which the weight becomes greater as the word level becomes higher.

Table 9 Weights of lexical difficulty indicators

Parameter	Level 1 words	Level 2 words	Level 3 words	Level 4 words	Level 5 words	Level 6 words	Outline words
Information entropy	0,999	0,940	0,937	0,934	0,920	0,906	0,785
Redundancy	0,001	0,060	0,063	0,066	0,080	0,094	0,215
Weight	0,001	0,104	0,109	0,114	0,138	0,162	0,373

The weighted sum of each indicator was used to obtain the lexical difficulty of the sentence. The larger the value, the greater the lexical difficulty of the sentence. Tab. 10 compares the lexical difficulty of sentences in all levels of the text. The average lexical difficulty of the primary, intermediate, and advanced text sentences were 0,469, 1,298, and 1,516, respectively. Thus, the sentence difficulty can be well reflected by the vocabulary. Furthermore, the maximum lexical difficulty and standard deviation of the primary text sentences were 3,079 and 0,474, respectively, which indicates that the range of lexical difficulty in the primary text is limited. Meanwhile, the range of lexical difficulty in the advanced text sentence is the largest, whereas that in intermediate text sentence is in the middle.

Table 10 Analysis of lexical difficulty

Level	Average	Maximum	Standard deviation
Primary	0,469	3,079	0,474
Intermediate	1,298	8,411	0,989
Advanced	1,516	9,289	1,1697

4.4 Grammatical Difficulty

Based on the grammatical difficulty model, the information entropy, redundancy, and weight of each indicator were obtained. The results are summarized in Tab. 11, in which the weight gradually increases as the grammatical level increases, i.e., the higher the grammatical level, the greater the weight.

Table 11 Weights of grammatical difficulty indicators

Parameter	Level 1 grammar	Level 2 grammar	Level 3 grammar	Level 4 grammar	Level 5 grammar	Level 6 grammar
Information entropy	0,998	0,916	0,903	0,848	0,739	0,663
Redundancy	0,002	0,084	0,097	0,152	0,261	0,337
Weight	0,002	0,090	0,104	0,163	0,279	0,361

Furthermore, the grammatical difficulty of sentences in different levels of the texts was compared. The average grammatical difficulties of the primary, intermediate, and advanced text sentences were 0,179, 0,365, and 0,341, respectively. Thus, the sentence difficulty cannot be well reflected by grammar. This is because the difficulties of a large part of grammatical points are determined by vocabularies. For example, "忽然 (suddenly)" is a time adverb in the grammatical knowledge base but it was also included in the graded lexical knowledge base. To avoid repeated calculation, the grammatical knowledge base was manually filtered and 425 grammatical points were screened out. These grammatical points were then used to distinguish the sentence difficulty, and the degree of distinction was poor. Therefore, the combination of grammar and lexicon is required in sentence readability.

4.5 Sentence Readability

Based on the assessment model of the sentence readability, the information entropy, redundancy, and entropy weights of the lexical and grammatical difficulties were obtained, and the results are summarized in Tab. 12. The entropy weight of the lexical difficulty was 0,406, while that of the grammatical difficulty was 0,594. The weight of the grammatical difficulty is relatively large because the entropy weight method is entirely based on

objective data. Besides, the values of the grammatical difficulty indicators considerably vary; hence, the weight obtained by the entropy weight method is relatively large. As the entropy weight method does not consider the importance of indicators, the importance weight of indicators was introduced. After several experiments and comparisons, the importance and comprehensive weights were obtained, as shown in Tab. 12. The comprehensive weight increases the weight of the lexical difficulty and reduces the weight of the grammatical difficulty. This conclusion is consistent with the research results in [29], i.e., lexical knowledge is more strongly associated with reading comprehension than grammatical knowledge. This study also confirms previous research that the prediction ability of the grammatical indicator to the readability of Chinese L2 text is relatively weak [27].

Table 12 Indicators of sentence readability

Parameter	Lexical difficulty	Grammatical difficulty
Entropy	0,963	0,946
Redundancy	0,037	0,054
w_i (entropy weight)	0,406	0,594
α_i (index importance)	0,9	0,1
β_i (comprehensive weight)	0,860	0,140

The sentence readability can be obtained by the weighted sum of each indicator. The average, standard deviation, and confidence interval of the sentence

readability in each level of the text were calculated (assuming that the distribution of sentence readability in each level of the text follows Gaussian distribution, 95% confidence interval is constructed). Meanwhile, to verify the effectiveness of the method proposed in this paper, the method proposed in [30] was used to analyze the experimental corpus in this paper. The results are summarized in Tab.13, in which both methods distinguished sentence readability in different levels of the text. The paper [30] calculated sentence readability by counting the lexical semantic categories in the sentence. The sentence readability became higher as the value became smaller. As the semantic classification dictionary in [30] could not be obtained, we counted the number of semantic categories in the sentence using HowNet.

Table 13 Results of sentence readability

Method	Level	Average	Standard deviation	Confidence interval
Paper [30]	Primary	2,87	1,30	[2,79, 2,96]
	Intermediate	2,60	0,90	[2,56, 2,64]
	Advanced	2,50	0,93	[2,47, 2,54]
This paper	Primary	0,391	0,392	[0,368, 0,415]
	Intermediate	1,048	0,780	[1,018, 1,078]
	Advanced	1,201	0,912	[1,168, 1,234]

Additionally, the correlation between sentence readability and text readability was analyzed using:

$$r = \frac{\sum_{i=1}^m (sc_i - \overline{sc})(dc_i - \overline{dc})}{\sqrt{\sum (sc_i - \overline{sc})^2 \sum (dc_i - \overline{dc})^2}} \quad (17)$$

where \overline{sc} represents the average of all the sentence readability and \overline{dc} represents the average of the text readability where the sentence is located. The results are shown in Tab. 14. The correlation coefficient of the sentence readability and text readability obtained by this method is 0,37, which is significantly improved compared with the method in [30]. By constructing the T statistic to analyze the significance of the correlation, the T statistic was not within the T critical value ($-2,33 < T < 2,323$), indicating that at 99% confidence level, there was a significant positive correlation between the sentence readability and the text readability in which the sentence was located.

Table 14 Analysis of the correlation between the sentence readability and text readability

Parameter	Paper [30]	This paper
Correlation coefficient	-0,141	0,37
T statistic	-29,624	49,37
T critical value (99%)	2,33	2,33

Based on the practice of Chinese L2 teaching, we selected the two most important reference bases in the teaching process, namely lexical and grammatical knowledge, and calculated the lexical and grammatical difficulties in the sentences. Afterwards, the sentence readability based on lexical and grammar difficulties was obtained. The experimental results show that the method has good applicability in the field of Chinese L2 teaching.

In [30], the number of semantic categories in a sentence was used to measure the sentence difficulty. For example, although "我的爸爸是军人 (my father is a soldier)" and "我的爸爸是书法 (my father is calligraphy)" have the same length, the number of semantic categories of the first sentence is relatively small and the sentence difficulty is relatively low. Hence, the method in [30] is more suitable for determining the relative difficulty of sentence pairs.

5 CONCLUSION

In this study, the standards in the field of Chinese L2 were extended and improved. First, a method for building a large-scale graded lexical knowledge base was proposed. The method extends the graded lexicon by improving the basic lexicon of the HSK syllabus, split analogy, and combination analogy, and obtains the levels of different meaning items of words. From the experimental results, the extended lexical knowledge base can better distinguish different levels of texts. Additionally, the existing teaching grammatical knowledge base was supplemented with sentence-based feature information to automatically identify grammatical points in the sentence and obtain the relevant quantitative indicators of grammatical knowledge. Based on these knowledge bases, the applicability of the comprehensive evaluation method in sentence readability measurement was explored. The experimental results show that the sentence readability calculated by the comprehensive evaluation method is well distinguished in all levels of texts, and the sentence readability and text readability are significantly positively correlated. The method proposed in this paper is aimed at the practical needs of Chinese as a second language teaching and has a good interpretability. It can not only provide sentence difficulty measurement, but also show specific lexical and grammatical difficulties in sentences.

Due to the lack of sentence readability tagging corpus, this study only analyzes the correlation between sentence readability and text level, and cannot directly evaluate the accuracy of sentence readability. To analyze the accuracy of sentence readability assessment, it is necessary to establish an L2 sentence readability tagging corpus in the future. The extended large-scale graded lexical knowledge base is suitable for text readability assessment. However, its application to Chinese L2 teaching needs to be further verified by experts.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China (Grant No. 61877004 and Grant No. 62007004) and a key project of the National Social Science Foundation of China (Grant No. 18zda295).

6 REFERENCES

- [1] James, H. (2007). Reading, writing, speaking and listening: perspectives in applied linguistics. *Applied Linguistics*, 28(2), 316-320. <https://doi.org/10.1093/applin/amm014>
- [2] Barohny, E. (2019). The zone of proximal development as an overarching concept: a framework for synthesizing vygotsky's theories. *Educational Philosophy and Theory*, 51(1), 18-130. <https://doi.org/10.1080/00131857.2017.1421941>

- [3] Xiaobin, C. & Detmar, M. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4), 418-447. <https://doi.org/10.1080/09588221.2018.1527358>
- [4] Sung, Y. T., Lin, W. C., Dyson, S. B., Chang, K. E., & Chen, Y. C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2), 371-391. <https://doi.org/10.1111/modl.12213>
- [5] Tianbao, S. (2016). Automatic extraction and recommendation of Grammar Points in L2 Chinese. *2016 International Conference on Asian Language Processing (IALP)*, 184-188. <https://doi.org/10.1109/IALP.2016.7875964>
- [6] Shuqin, Z., Jihua, S., Weiming, P., Dongdong G., & Gu W. (2019). Text readability assessment for Chinese second language teaching. *Chinese Lexical Semantics Workshop*, 393-405. https://doi.org/10.1007/978-3-030-38189-9_41
- [7] Nikolay, K., Julia, B., & Fedor, V. (2014). Single-sentence readability prediction in Russian. *Communications in Computer and Information Science*, 436, 91-100. https://doi.org/10.1007/978-3-319-12580-0_9
- [8] Kotani, K., Yoshimi, T., Nanjo, H., & Isahara, H. (2012). Applicability of readability formulae to the measurement of sentence-level readability. *5th International Conference of Education, Research and Innovation*, 6023-6031.
- [9] Kroeger, E. & Sun, M. (2005). Sentence difficulty evaluation for a learner's dictionary. *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 678-682. <https://doi.org/10.1109/NLPKE.2005.1598822>
- [10] Weiming, P., Jihua, S., & Min, Z. (2014). Construction of grammar resource base for international Chinese Teaching. *Distance Education in China*, (8), 90-94.
- [11] Weichao, Y., Kai, X., Jijian, L., Lingling B., & Chao M. (2018). Multiple flood vulnerability assessment approach based on fuzzy comprehensive evaluation method and coordinated development degree model. *Journal of Environmental Management*, 213(1), 440-450. <https://doi.org/10.1016/j.jenvman.2018.02.085>
- [12] Surman, R. & Bath, P. A. (2013). An assessment of the quality of information on stroke and speech and language difficulty web sites. *Journal of Information Science*, 39(1), 113-125. <https://doi.org/10.1177/0165551512469775>
- [13] Williamson, J. M. L. & Martin, A. G. (2010). Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. *International Journal of Clinical Practice*, 64(13), 1824-1831. <https://doi.org/10.1111/j.1742-1241.2010.02408.x>
- [14] Jeanne, G. B. (1981). Computerized readability levels. *IEEE Transactions on Professional Communication*, 24(1), 45-46. <https://doi.org/10.1109/TPC.1981.6447823>
- [15] Zhou, R. & Wang, Z. (2016). The application of readability formulas for the Tibetan teaching textbooks. *3rd International Conference on Education*, 49, 812-817. <https://doi.org/10.2991/icemaess-15.2016.169>
- [16] McLaughlin, G. (1969). Smog grading - a new readability formula. *Journal of Reading*, 12(8), 639-646.
- [17] Burkhead, M. & Ulferts, G. (1977). Sample frequency in application of Dale-Chall readability formula. *Journal of Reading Behavior*, 9(3), 287-290. <https://doi.org/10.1080/10862967709547229>
- [18] Luong, A. V., Nguyen, D., & Dinh, D. (2018). A new formula for Vietnamese text readability assessment. *10th International Conference on Knowledge and Systems Engineering*, 198-202. <https://doi.org/10.1109/KSE.2018.8573379>
- [19] Rohit J. K., Xiaoqiang L., Siddharth P., Martin Franz, Radu F., Raymond J. M., Salim R., & Chris Welty (2010). Learning to predict readability using diverse linguistic features. *23rd International Conference on Computational Linguistics*, 546-554.
- [20] Ute, K., Amir, R., & Neomy, S. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, 21(21), 1-17. <https://doi.org/10.1016/j.asw.2014.01.001>
- [21] Kuiken, F. & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48-60. <https://doi.org/10.1016/j.jslw.2007.08.003>
- [22] Norris, J. M. & Lourdes, O. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Lingus*, 30(4), 555-578. <https://doi.org/10.1093/applin/amp044>
- [23] Lu, X. F. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- [24] Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20(2), 148-161. <https://doi.org/10.1016/j.jslw.2011.02.001>
- [25] Zhang, D. (2013). Vocabulary and grammar knowledge in second language reading comprehension: a structural equation modeling study. *Modern Language Journal*, 96(4), 558-575. <https://doi.org/10.1111/j.1540-4781.2012.01398.x>
- [26] Stæhr, L. S. (2008). Vocabulary size and the skills of reading, listening and writing. *Language Learning Journal*, 36(2), 139-152. <https://doi.org/10.1080/09571730802389975>
- [27] Ildikó, P., Elena, V., & Richard J. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. 174-184. <https://doi.org/10.3115/v1/W14-1821>
- [28] Akbari, Z. (2014). The role of grammar in second language reading comprehension: Iranian ESP context. *Procedia-Social and Behavioral Sciences*, 98(6), 122-126. <https://doi.org/10.1016/j.sbspro.2014.03.397>
- [29] Aryadoust, V. & Baghaei, P. (2016). Does EFL readers' lexical and grammatical knowledge predict their reading ability? Insights from a perceptron artificial neural network study. *Educational Assessment*, 21(2), 135-156. <https://doi.org/10.1080/10627197.2016.1166343>
- [30] Jinquan, Z. (2005). Lexical semantics and sentence difficulty measurement. *Chinese Lexical Semantic Workshop*, 261-265.
- [31] Shaomin, J. (2009). *Research on sentence difficulty measurement*. PhD thesis, Xiamen University.
- [32] Wanghao, G. (2010). *Research on the formula of readability of TCFL text*. Master thesis, Shanghai Jiaotong University.
- [33] Dong, Y., Siyuan, W., Chaoyang, G., & Yuling T. (2020). Assessing sentence difficulty in Chinese textbooks based on crowd sourcing. *Journal of Chinese Information Processing*, 34 (2), 16-26. https://doi.org/10.1007/978-3-319-27194-1_68
- [34] Shuqin, Z., Yinxia, Z., Weiming, P., & Jihua, S. (2016). Construction of the basic sentence-pattern instance database based on the international Chinese textbook Treebank. *2016 International Conference on Asian Language Processing (IALP)*. <https://doi.org/10.1109/IALP.2016.7875983>
- [35] Yinbing, Z., Jihua, S., & Weiming, P. (2019). Construction and analysis of Chinese word-formation knowledge base based on modern Chinese dictionary. *Proceedings of the 20th Chinese Lexical Semantics Workshop*.

Contact information:

Shuqin ZHU

(Corresponding author)

1) Teacher's College, Beijing Union University,

5 Waiguanxiejie St, Beijing 100111, Peoples R China.

2) School of Artificial Intelligence, Beijing Normal University, 5

Waiguanxiejie St, Beijing 100111, Peoples R China.

E-mail: stzhushuqin@bnu.edu.cn

Jihua SONG

(Corresponding author)

School of Artificial Intelligence, Beijing Normal University,

No.19, Xijiekouwai St, Haidian District, Beijing, 100875, China

E-mail: songjh@bnu.edu.cn

Weiming PENG

School of Artificial Intelligence, Beijing Normal University,

No.19, Xijiekouwai St, Haidian District, Beijing, 100875, China

E-mail: pengweiming@bnu.edu.cn

Jingbo SUN

School of Artificial Intelligence, Beijing Normal University,

No.19, Xijiekouwai St, Haidian District, Beijing, 100875, China

E-mail: sunjingbo@mail.bnu.edu.cn