Hindawi
Complexity
Volume 2023, Article ID 9847216, 1 page
https://doi.org/10.1155/2023/9847216

WILEY | Hindawi

Retraction

Retracted: The Measurement of Chinese Sentence Semantic Complexity

Complexity

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Zhu, J. Song, W. Peng, D. Guo, and J. Sun, "The Measurement of Chinese Sentence Semantic Complexity," *Complexity*, vol. 2020, Article ID 8871263, 10 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 8871263, 10 pages
https://doi.org/10.1155/2020/8871263

Research Article

The Measurement of Chinese Sentence Semantic Complexity

Shuqin Zhu,[1,2] **Jihua Song**,[2] **Weiming Peng**,[2] **Dongdong Guo**,[2] **and Jingbo Sun**[2]

[1]*Teacher's College of Beijing Union University, Beijing 100011, China*
[2]*School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China*

Correspondence should be addressed to Shuqin Zhu; sftzhushuqin@buu.edu.cn, Jihua Song; songjh@bnu.edu.cn, and Weiming Peng; pengweiming@bnu.edu.cn

Received 13 August 2020; Revised 19 September 2020; Accepted 6 October 2020; Published 20 November 2020

Academic Editor: Zhihan Lv

The complexity of language is usually reflected in the complexity of sentences. At present, the research of sentence complexity mainly focuses on the analysis of syntactic complexity. In this paper, from the perspective of Leech's theory of sentence semantic structure, the predication structure is taken as the semantic unit to explore the sentence semantic complexity. The predication structures are extracted based on the result of sentence-based syntactic analysis, and then the linear expression sequence of a sentence is converted into a semantic hierarchy based on predicate semantic frameworks; the universality of predicate semantic frameworks is obtained by using the spectral clustering algorithm; and the sentence semantic complexity depends on the universality of predicate semantic frameworks at various layers. The experimental results show that the measurement method of sentence sematic complexity based on predicate semantic frameworks is more effective by comparing with the method that only considers the semantic categories of words in the sentence.

1. Introduction

Language complexity refers to a property or quality of a phenomenon or entity in terms of (1) the number and the nature of the discrete components that the entity consists of and (2) the number and the nature of the relationships between the constituent components [1]. The complexity of language is embodied in vocabulary, pronunciation, grammar, and other subsystems. Among them, each plane subsystem (syntax, semantics, and pragmatics) within the grammar subsystem also has complexity [2]. This paper will focus on the semantic complexity, especially the measurement of sentence semantic complexity.

According to Leech's theory of sentence semantic structure, the predication structure is the main semantic unit of a sentence [3]. A predication structure can be divided into arguments and the predicate connecting arguments. Among them, the predicate is the main component of the predication structure, which determines the number and nature of arguments. Moreover, there are subordinate predication structures and degraded predication structures, and the difference between them lies in their different layers and positions in sentences [4]. Yushu Hu pointed out the sentence semantics should not be sought from the lexical semantics in the sentence, but from the form or structure of the sentence. "Only by structural analysis can we summarize the common semantics from the same structures, and only by structural analysis can we find different semantics in different structures" [5].

According to the existing theory and analysis method of sentence semantic structure, this paper starts from the sentence structure and converts the linear expression sequence of the sentence into semantic hierarchy based on the results of sentence-based syntactic analysis. That is, the predication structure is used as the analysis unit. The predication structures of a sentence that need to be expressed preferentially are selected as the important parts, and the unimportant predication structures are selected as the additional components. The predication structures are arranged in layers according to the direct or indirect relationship between the various sentence components. Secondly, combined with the definition of words in HowNet [6], the arguments of the predication structures are further abstracted and generalized to obtain predicate semantic

frameworks (PSFs). In this way, the linear expression sequence of a sentence is converted into a semantic hierarchy, and the sentence semantic complexity is converted into the complexity of PSFs which are measured by the universality of PSFs. Spectral clustering is used to cluster the PSFs of a predicate, and the universality of PSFs in a large class is relatively higher. Finally, the sentence semantic universality depends on the universality of PSFs at each layer, and different weights are given to PSFs at different layers. The sentence semantic universality reflects the sentence semantic complexity. The sentences with high semantic universality are frequently used and the learning order is in the front. Sentences with low semantic universality make it difficult for learners to learn and understand [7, 8]. That is, the higher the sentence semantic universality is, the lower the sentence semantic complexity is.

The main innovations of this paper are as follows: one is to propose a measurement method of the universality of PSFs based on the predication structure, so as to obtain the universality of different PSFs of a predicate; the second is to propose an assessment method of sentence semantic universality based on PSFs, and the sentence semantic complexity is reflected by the sentence semantic universality.

2. Related works

At present, sentence complexity is mainly analyzed from structure and syntax. In [9], it is considered that two kinds of commonly used operations to complicate the content of clauses are parallel compound structure and nesting clause structure. Among them, the parallel compound structure takes the total number of commas and parallel conjunctions appearing in clauses as the quantitative estimation basis for difficulty, and the nesting clause structure takes the number of core verbs appearing in clauses as the estimation basis for difficulty. The mean of the difficulty estimation value of all clauses is taken as the difficulty estimation value of the sentence. In [10], a linear comprehensive evaluation model is used to calculate the complexity of Chinese structure. The indicators used in the model include the total number of clauses, the number of embedded or subordinate clauses in clauses, and the ratio of the word number to the clause number.

In addition, in the field of second language teaching, syntactic complexity is mainly used to measure the syntactic usage of learners' language output, which is an important indicator of learners' language level and language development trajectory. L2SCA is a syntactic complexity analysis tool for English second language, which covers 14 indicators including 5 dimensions of syntactic length, dependency, collocation, phrase complexity, and sentence overall complexity [11, 12]. Paper [13] also selects 14 measurement indicators from three categories and five subcategories for the syntactic complexity of Chinese as a second language, namely, the number of characters, words, syntactic components, phrases, clauses, consortiums, partial relations, complement structures, conjunctions, disjunctions, disposals, and passive, existential, and relative clauses in a basic unit. Papers [14–19] also study sentence complexity, and researchers try to use various quantitative indicators to quantify sentence complexity.

Most of the existing researches on syntactic complexity focus on the analysis of sentence structure and formal features. Biber believes that simply considering sentence complexity from the perspective of structure does not really reflect its essence [20]. Ortega also believes that the semantic, function, and communicative value of sentence complexity should be analyzed and studied [21]. In addition, according to Bulté and Alex Housen, the complexity of language learning cognition consists of at least three parts: proposition complexity, discourse interaction complexity, and language complexity [1]. Among them, proposition refers to the semantics expressed in the text, not just the statement itself. The semantic structure of a proposition can be expressed as a "predication structure." Proposition complexity is a relatively new concept, which has received far less attention than language complexity [22, 23].

Therefore, this paper attempts to analyze the sentence semantic complexity based on the basic proposition. In Section 3, the extraction of predication structures, the acquisition of PSFs, and the calculation method of the universality of PFSs are introduced. In Section 4, the calculation method of sentence semantic universality is introduced. The experimental results are introduced and analyzed in Section 5. Finally, the conclusion and limitations of this study are discussed in Section 6.

3. Universality of PFSs

The calculation method of the universality of PSFs is shown in Figure 1. Based on the results of sentence-based syntactic analysis, the predication structures are extracted layer by layer, and the PSFs are obtained by combining the definition of words in HowNet. All the PSFs of a predicate are clustered to get the universality of the PSFs. In addition, it is necessary to calculate the similarity of PSFs through lexical similarity and sememe similarity in order to cluster PSFs.

3.1. Extraction of Predication Structures. The extraction of predication structures is based on the result of syntactic analysis in the sentence-based treebank [24, 25]. The analysis and annotation of sentences in the sentence-based treebank are in the form of visual diagram, as shown in Figure 2. The horizontal line is the benchmark to observe the sentence layer. The subject, predicate, object, attribute, adverbial, complement, and other sentence components attached to the same horizontal line belong to the same layer. The subject, predicate, and object are located above the line, which are the "main components" of the sentence pattern; the attribute, adverbial, and complement are located below the line, which are the "additional components" of the sentence pattern; for the complex additional components, the syntactic analysis goes deep layer by layer. The annotation results are stored in XML form. The diagram and XML can be transformed in both directions.

Based on the results of sentence-based syntactic analysis, the long horizontal line with predicate component is taken

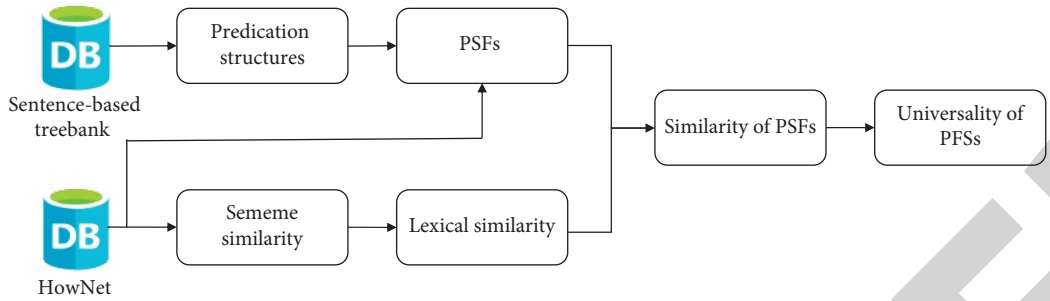

Figure 1: The calculation method of the universality of PSFs.

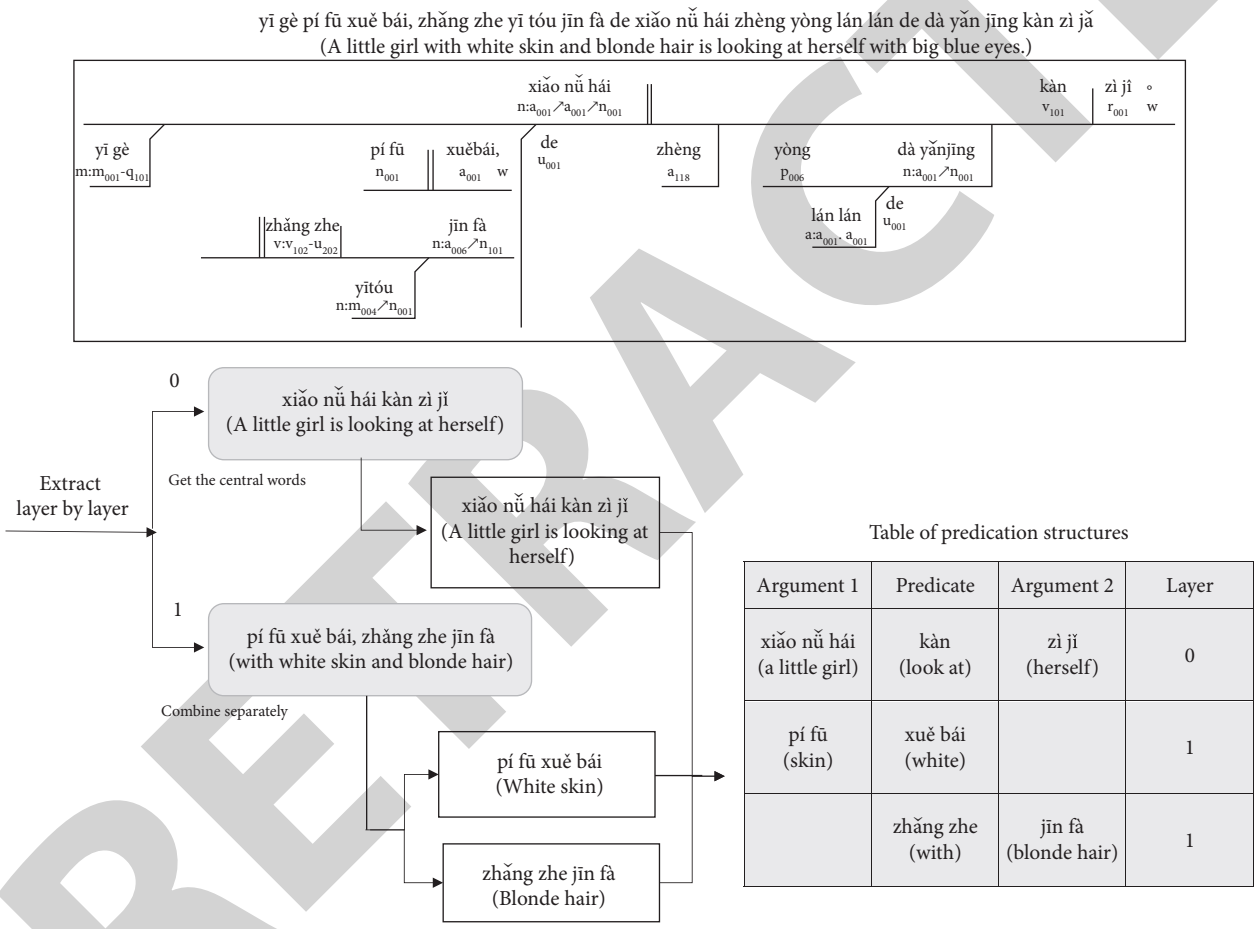

Table of predication structures

Argument 1	Predicate	Argument 2	Layer
xiǎo nǚ hái (a little girl)	kàn (look at)	zì jǐ (herself)	0
pí fū (skin)	xuě bái (white)		1
	zhǎng zhe (with)	jīn fà (blonde hair)	1

Figure 2: The extraction of predication structures.

as the baseline to extract the central word sequence directly related to the predicate. After the central word sequence of each layer is obtained, the predication structures are obtained by splitting and combining multiple predicates, and the process is shown in Figure 2.

It is possible that there are juxtaposed components in the subject or object. At this time, each component needs to be combined with core predicate separately. For example, in the sentence " yán sè , yàng zi dōu bǐ gāng cái kàn de qí páo hǎo (The color and style are better than those of the cheongsam I saw just now)," the subject includes juxtaposition, namely, "yán sè (color)" and "yàng zi (style)." The predication structures of layer 0 are "yán sè hǎo (The color is good) " and "yàng zi hǎo(The style is good)."

The sentences with multiple predicates need to be split. Table 1 lists the split methods of the compound predicates, joint predicates, linked predicates, and pivotal sentence.

Considering the complexity of Chinese language, sentence components not only are acted by words, but also may contain a new predication structure, which is directly identified by the "VP." For example, in the predication structure at layer 0 of the sentence "lì shǐ yǐ jīng zhèng míng tā zhǔ zhāng huáng quán shì cuò de(The history has proved that he is wrong in claiming imperial power)," " zhèng

Table 1: The split methods of multipredicates.

Type	Sentence	Split results
Compound predicates	wǐ shì fú nǐ le(I follow you!)	wǐ||shì wǐ||fú |nǐ
Joint predicates	mǔ qīn qiú shén bài fó(Mother prays for God and Buddha)	mǔ qīn ||qiú |shén mǔ qīn || bài|fó
Linked predicates	tā zhuā zhù wǐ de shǐu bù fàng (He held my hand)	tā || zhuā zhù | shǐu tā ||[bù] fàng
Pivotal sentence	wǐ qǐng nǐ chī zhōng cān (I invite you to eat Chinese food)	wǐ || qǐng | nǐ nǐ || chī | zhōng cān

míng(proof)" is the predicate and " lì shǐ(history)" and "VP" are arguments.

3.2. Acquisition of PSFs. Based on HowNet, the predicate structures are transformed into the PSFs. HowNet is a common sense knowledge base, which takes the concepts represented by Chinese and English words as the description object, and reveals the relationship between concepts and their attributes. HowNet defines a word as follows:

①wǐ:{human|rén:PersonPro = {1stPerson|wǐ }}

②wǐ:{specific|tè dìng:PersonPro = {1stPerson|wǐ }}

The first sememe in the definition of a word is the basic sememe, which points out the most basic meaning of the concept, such as "wǐ" referring to "human" or "specific." The colon is followed by a detailed explanation of the basic sememe.

Combined with the semantic definition of words in HowNet [6], the PSFs can be obtained by abstracting and generalizing the arguments of predication structures, as shown in Table 2. Each word only takes the first basic sememe of each definition. Since it is impossible to know the exact semantics of each argument, if a word has multiple definitions in HowNet, all definitions in HowNet will be listed here for use in subsequent steps. If the word is not defined in HowNet, the word is used directly.

3.3. Sememe Similarity. Sememe similarity is the basis of calculating lexical similarity. Sememe similarity can be obtained by calculating sememe distance [26]. The most classical calculation method is as follows:

$$\mathrm{sim}(s_1, s_2) = \frac{\alpha}{\mathrm{dis}(s_1, s_2) + \alpha}. \tag{1}$$

$\mathrm{dis}(s_1, s_2)$ is the distance between s_1 and s_2 in the sememe tree. If s_1 and s_2 are in the same tree, the distance is the sum of the path lengths from s_1 and s_2 to their minimum common sememe. If s_1 and s_2 are not in the same tree, the distance will take a maximum of 20; α is an adjustable parameter.

In the above calculation method, the weight of all paths is set to 1, but in HowNet, the difference between the top classes is large; the difference between the bottom classes is small. In view of this situation, [27] not only considers the depth of sememe tree, but also considers the regional density

Table 2: PSFs.

Argument 1	Predicate	Argument 2	Layer
{human|rén}	kàn(look at)	{human|rén} {inanimate| wú shēng wù} {self| jǐ}	0
{part|bù jiàn}	xuě bái(white)		1
	zhǎng zhe(with)	{AppearanceValue| wài guān zhí} {part|bù jiàn}	1

of sememe tree. The calculation method of sememe similarity is as follows:

$$\mathrm{sim}(s_1, s_2) = \frac{\alpha}{d + \alpha},$$

$$d = \delta \cdot \frac{\mathrm{dis}(s_1, s_2)}{\mathrm{con}(s_1) + \mathrm{con}(s_2)},$$

$$\mathrm{con}(s_1) = \gamma \mathrm{deep}(s_1) + \eta \mathrm{desity}(s_1) \, (\gamma < \eta \mathrm{and} \gamma + \eta = 1),$$

$$\mathrm{desity}(s_1) = \frac{nc(s_1)}{\beta}, \tag{2}$$

where $\mathrm{dis}(s_1, s_2)$ is the distance between s_1 and s_2 in the sememe tree. $\mathrm{deep}(s_1)$ is the depth of s_1 in the sememe tree, that is, the path length from the root node to the sememe s_1. $nc(s_1)$ is the sibling node number of s_1. The parameters are set as follows: $\alpha = 1.6, \beta = 50, \gamma = 0.3, \eta = 0.7, \delta = 3$.

3.4. Similarity of PSFs. There may be n parts (arguments) in a PSF. For two different semantic frameworks of a predicate ($F1$ and F_2), if n is different, the possibility of similarity is small, and the similarity of the two PSFs is taken as 0. If n is the same, each framework has $\mathrm{ar}1, \mathrm{ar}2, \ldots, \mathrm{ar}n$ parts (arguments), and $\mathrm{sim}(F_1, F_2)$ is determined by the similarity of each part.

$$\begin{aligned} \mathrm{sim}(F_1, F_2) = {} & \alpha_{\mathrm{ar}1} * \mathrm{Sim}_{\mathrm{ar}1}\left(W_{F_1 1}, W_{F_2 1}\right) + \alpha_{\mathrm{ar}2} \\ & * \mathrm{Sim}_{\mathrm{ar}2}\left(W_{F_1 2}, W_{F_2 2}\right) + \cdots + \alpha_{\mathrm{ar}n} \\ & * \mathrm{Sim}_{\mathrm{ar}n}\left(W_{F_1 n}, W_{F_2 n}\right). \end{aligned} \tag{3}$$

$\alpha_{ar1}, \alpha_{ar2}, \ldots, \alpha_{arn}$ are the adjustable parameters, namely, the weight of each part, and $\alpha_{ar1} + \alpha_{ar2}, \ldots, +\alpha_{arn} = 1$. If W_{F_1k} has m definitions in HowNet: $S_{11}, S_{12}, \ldots, S_{1m}$ and W_{F_2k} has l definitions in HowNet: $S_{21}, S_{22}, \ldots, S_{2l}$, $\text{sim}(W_{F_1k}, W_{F_2k})$ is the maximum value of similarity between definitions:

$$\text{sim}\left(W_{F_1k}, W_{F_2k}\right) = \max_{i=1\ldots m, j=1\ldots l}\left(\text{sim}\left(S_{1i}, S_{2j}\right)\right). \quad (4)$$

For each part of a PSF, the first basic sememe of each definition is obtained from HowNet, so the similarity between definitions is the similarity between sememes.

The subject is the person or thing to be described in a sentence. It is the statement object of the predicate. The predicate and the object are generally combined to describe the subject. In view of the closer relationship between the predicate and the object, the parameters are set as follows:

predicate + object + object(VOO) structure: $\alpha_{ar1} = 0.5$, $\alpha_{ar2} = 0.5$

subject + predicate + object(SVO) structure: $\alpha_{ar1} = 0.2$, $\alpha_{ar2} = 0.8$

subject + predicate + object + object(SVOO) structure: $\alpha_{ar1} = 0.2, \alpha_{ar2} = 0.4, \alpha_{ar3} = 0.4$

3.5. Clustering of PSFs. The similarity matrix of PSFs is obtained by calculating the similarity between the semantic frameworks of each predicate. The method of spectral clustering is used to cluster the semantic frameworks of each predicate, and PSFs in large classes have a high universality.

Spectral clustering is a kind of clustering method based on graph theory [28–30]. All data vertices $V = \{v_1, v_2, \ldots, v_n\}$ form undirected weighted graph G (V,E). Vertices can be connected by edges, and the weight w_{ij} on each edge represents the relationship between v_i and v_j. Because G is an undirected graph, the weight on the edges is independent of the direction of the two points, $w_{ij} = w_{ji}$. The matrix composed of the weights between any two points is the adjacency matrix W of a graph. For any point v_i in a graph, its degree d_i is defined as the sum of the weights of all the edges connected with it, that is, $d_i = \sum_{j=1}^{n} w_{ij}$. The degree matrix can be expressed as D. D is a diagonal matrix whose value is the degree of each vertex.

Each semantic framework of each predicate can be regarded as a vertex in graph G. The relationship between the semantic frameworks of each predicate is represented by the adjacency matrix W, that is, the PSFs similarity matrix of a predicate. Clustering is to cut the graph G into k subgraphs, so that the sum of edge weights between different subgraphs is as low as possible, while the sum of edge weights within subgraphs is as high as possible, as shown in Figure 3. The number of vertices contained in each subgraph is the universality of this kind of PSF u_i.

4. Sentence Semantic Universality

According to Levy, there are two different ways to understand sentences: one is based on memory; the other is based on expectation. Because of the need to complete the timely storage, synthesis, and extraction of input information, it is difficult to understand based on memory [31]. The text that meets reading expectation is relatively easy to understand. For example, the following two sentences have the same number of words, but the premodifiers in the first sentence are juxtaposed, which meet reading expectation and are easy to understand. However, the second sentence is not easy to understand because of its multiple nesting of modifiers [10].

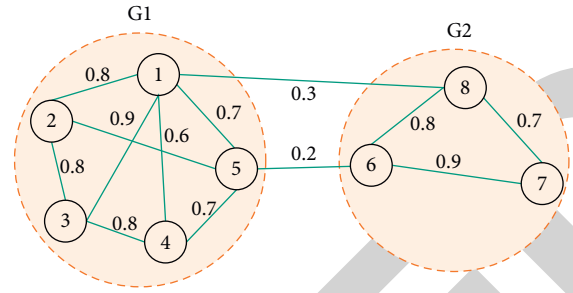

Figure 3: Clustering of PSFs.

(1) zài chù dǎngwěi de dà nào hóng 9 yuè, géxīn 2 qiān jiàn, bǎozhèng bǎi mǐ jǐng, guóqìng bǎ lǐ xiàn de xíngdòng kǐuhào xià (25 words; under the slogan of the party committee's campaign to make a big splash in September, to innovate 2000 pieces, to ensure the 100-meter well and to present gifts on National Day).

(2) duǎnduǎnde guānyú shìjiè shàng de zhǐngzhǐng de lìshǐ de zōngjiào duìyú rénlèi de sǐwáng hòu de shēngmìng suǐ céngjīng qǔ guò de tàidù de xùshù (25 words; a short narrative about the attitudes of various historical religions in the world to human life after death).

Based on the above theory, the sentence semantic complexity can be divided into two parts: the complexity of the main PSFs and the complexity of the additional PSFs. Only by understanding the main PSFs can we grasp the central idea of the sentence. Only by clarifying the additional PSFs can we get a complete understanding of sentence semantics. Different weights are given to PSFs at different layers, and the semantic universality of a sentence (U_{sen}) with n structures is the synthesis of the universality of PSFs (u_i) in every layer.

$$U_{sen} = \sum_{i=0,\ldots,n} \alpha_i u_i, \quad (5)$$

where α_i is an adjustable parameter, that is, the importance of different PSF, which will be determined later by experiments.

5. Experiments and Discussion

5.1. Experimental Data. 244 volumes of international Chinese textbooks in the sentence-based treebank are selected to obtain the universality of PSFs, which includes 4,695 documents and 91,526 sentences (separated by。？!).

Boya Chinese is selected to complete experiments of sentence semantic complexity. *Boya Chinese* contains 9

volumes of textbooks. The difficulty of these textbooks increases in turn, and they can be divided into primary, intermediate, and advanced. The details are shown in Table 3.

5.2. Universality of PFSs. Based on 91,526 sentences, 231,020 predication structures are extracted. 1,138 predicates with a frequency greater than 20 are clustered. The contour coefficient is used to measure the density and dispersion of the classes, so as to automatically select the number of clusters. The calculation method of the contour coefficient is as follows:

$$S = \frac{b - a}{\max(a, b)}. \tag{6}$$

For a predication structure, a is the average distance from other predication structures in the same category, and b is the average distance from the predication structures in the different categories closest to it. The overall contour coefficient is the average value of all the contour coefficients. The larger the contour coefficient is, the better the dispersion between classes is; the smaller the contour coefficient is, the worse the clustering effect is.

After clustering, the percentage of a kind of predication structure can be obtained. As shown in Table 4, in the predication structures of "tí gāo (improve)", the first class of predication structures accounts for 6.7%, and the second class of predication structures accounts for 24.6%. Combined with the occurrence frequency of the predicate, the universality of each predication structure can be obtained. For predicates whose frequencies are less than or equal to 20, the universalities of their predication structures are set to 1.

5.3. Sentence Semantic Universality. This paper analyzes the sentence semantic universality of *Boya Chinese*. At the same time, the setting methods of adjustable parameter in the calculation formula of sentence semantic universality are compared in this experiment.

Method 1: the sentence universality takes the lowest universality of PSFs in the sentence.

$$\alpha_i = \begin{cases} 1, & \text{if } \min(u_i), \\ 0, & \text{other.} \end{cases} \tag{7}$$

Method 2: if there is only one layer of syntactic structure in a sentence, the weights of all the predication structures are the same; otherwise, the weight of predication structures at the backbone layer is 0.8, and the weight of predication structures at the additional layer is 0.2.

First of all, method 1 is used to set adjustable parameters. Table 5 shows the distribution of sentence semantic universality in textbooks at all levels. It can be seen intuitively that, with the increase of text difficulty, the proportion of sentences with low universality gradually increases, from 26.6% to 82.6%, and the proportion of sentences with high universality is declining sharply.

Method 2 is used to calculate the sentence semantic universality, and the distribution of sentence semantic universality in each textbook is shown in Table 6. From the results in the table, the distribution of sentences with semantic universality between 1 and 20 in the textbooks of Book 1 to Book 9 is not rising steadily. The distribution of sentences with semantic universality more than 1000 has not achieved the expected effect, and the distribution law is not obvious in all levels of textbooks.

In order to compare the difference of sentence semantic universality between the two methods on text difficulty, the relative entropy (KL distance) between adjacent level texts is calculated based on sentence semantic universality. KL distances are shown in Table 7. It can be seen that the sentence semantic universality calculated by Method 1 can better distinguish texts at all levels, and the KL distances between textbook texts at adjacent levels are larger, so Method 1 is used to obtain sentence semantic universality. The effect of Method 2 is not as expected. This may be because the split of the sentence is too detailed when obtaining the predication structures, resulting in the frequencies of synthetic predicates being higher, which affects the calculation of sentence semantic universality. For example, the sentence "wǐ néng qù yóu yǐng(I can go swimming)" is divided into "wǐ néng(I can)," "wǐ qù(I go)," and "wǐ yóu yǐng(I swim)." In this case, the frequencies of predicates such as "néng(can)" and "qù(go)" have increased a lot.

5.4. Comparative Experiment

5.4.1. Baseline. From the above experiments, it can be seen that when sentence semantic universality is used to represent sentence semantic complexity, sentence semantic complexity has obvious distribution law in all levels of texts (Method 1). The method in this paper closely connects structure and semantic, extracts the predication structures layer by layer based on the results of syntactic analysis, and synthesizes the complexity of the predication structures at all levels of a sentence.

In order to further verify the effectiveness of this method, the following method does not consider sentence structure and only measures the sentence semantic complexity from the diversity of lexical semantics. The calculation method is given as an example below [32].

If there is a dialogue below:

(A) wǐ de bà bà yán jiū de shì shù xué, nǐ de bà bà ne? (My dad studies mathematics, what about your dad?).

(B) wǐ de bà bà shì shū fǎ. (My father studies calligraphy).

So, although the structure of the following two sentences is the same, it is clear that the first sentence is easier to understand than the second sentence, because the semantics of "bà bà(daddy)" and "jūn rén(military)" are the same [32].

(1) wǐ de bà bà shì jūn rén (My father is a soldier).

(2) wǐ de bà bà shì shū fǎ (My father studies calligraphy).

Table 3: Boya Chinese.

Title	Character	Level	Chapter	Sentence	Average of character in a sentence	Average of word number in a sentence
Boya Chinese 1	5876	Primary	66	1075	12.244	7.46
Boya Chinese 2	7259					
Boya Chinese 3	11560	Intermediate	67	2549	24.774	14.519
Boya Chinese 4	17390					
Boya Chinese 5	14475					
Boya Chinese 6	28344					
Boya Chinese 7	26421	Advanced	48	2942	27.002	15.035
Boya Chinese 8	27955					
Boya Chinese 9	34942					

Table 4: Predication structures.

Class	Argument 1	Predicate	Argument 2	%
1	tā(it)	tí gāo(improve)	lì yòng lǜ(utilization rate)	6.7
1		tí gāo(improve)	dān chǎn(per unit yield)	6.7
1		tí gāo(improve)	chǎn liàng(yield)	6.7
1		tí gāo(improve)	jǐng tì(alert)	6.7
1	jié jìng(shortcut)	tí gāo(improve)	shōu shì lǜ(viewing rate)	6.7
2	dī shōu rù jiē céng(low income class)	tí gāo(improve)	shēng huó shuǐ píng(living standard)	24.6
2		tí gāo(improve)	shè huì dì wèi(social position)	24.6
2		tí gāo(improve)	rén kǐu sù zhì(population quality)	24.6
2		tí gāo(improve)	mǎn yì dù(satisfaction)	24.6
2		tí gāo(improve)	néng lì(ability)	24.6
2		tí gāo(improve)	shēng huó zhì liàng(quality of life)	24.6
2		tí gāo(improve)	sù zhì(quality)	24.6

Table 5: Distribution of sentence semantic universality in each textbook (Method 1).

Textbook	1–20	21–100	101–200	201–300	301–400	401–1000	>1000
1	26.60	29.40	8.20	5.60	1.80	7.80	20.60
2	52.00	25.26	7.37	5.26	0.42	2.74	6.95
3	60.89	22.67	5.56	4.00	0.44	2.00	4.44
4	69.01	20.13	2.08	2.40	0.48	2.40	3.51
5	72.77	14.19	2.97	1.37	0.23	2.75	5.72
6	80.02	10.10	1.72	1.29	0.43	1.72	4.73
7	76.75	10.72	2.30	1.70	0.40	1.20	6.91
8	78.68	10.98	2.00	1.27	0.18	1.91	4.99
9	82.60	10.18	1.13	0.14	0.42	1.41	4.10

Table 6: Distribution of sentence semantic universality in each textbook (Method 2).

Textbook	1–20 (%)	21–100 (%)	101–200 (%)	201–300 (%)	301–400 (%)	401–1000 (%)	>1000 (%)
1	10.20	19.20	11.60	8.40	3.00	19.00	28.60
2	13.26	21.47	13.89	9.05	5.26	19.37	17.68
3	12.89	22.00	11.56	7.78	6.44	24.67	14.67
4	12.46	23.48	9.27	7.67	7.19	22.68	17.25
5	14.87	17.62	10.30	5.72	5.03	23.34	23.11
6	22.66	17.83	8.92	5.69	4.73	20.62	19.55
7	23.85	18.84	7.72	6.51	4.61	19.54	18.94
8	19.51	16.88	9.98	5.54	4.26	23.96	19.87
9	24.05	18.39	6.79	4.81	5.23	18.53	22.21

Table 7: KL distances between textbooks of adjacent level.

Text	Method 1	Method 2
Primary & intermediate	0.10832	0.011571
Intermediate & advanced	0.00822	0.005318

The semantics of each word in the sentences obtained from HowNet are as follows (because the semantic classification dictionary in [32] cannot be obtained, we count the number of semantic categories in the sentence based on HowNet):

①wǐ:human| rén
②wǐ:specific| tè dìng
③de:FuncWord| gōng néng cí
④bà bà: human| rén
⑤shì:be| shì
⑥shì:exist| cún zài
⑦shì:expression| cí yǔ
⑧shì:specific| tè dìng
⑨jūn rén: human| rén
⑩shū fǎ: method| fāng fǎ

Only the number of semantic categories is considered, and the occurrence number of semantic categories is not counted. The number of semantic categories in the first sentence (wǐ de bà bà shì jūn rén) is 6 (①②③⑤⑥⑦). The number of semantic categories in the second sentence (wǐ de bà bà shì shū fǎ) is 7 (①②③⑤⑥⑦⑩)). In order to offset the influence of sentence length, the sentence semantic complexity = the number of semantic categories in the sentence/the number of words in the sentence [32]. The semantic complexity of the first sentence = 6/5 = 1.2, and the semantic complexity of the second sentence = 7/5 = 1.4. It can be seen that the second sentence has a higher complexity and is more difficult to understand.

5.4.2. Results. The summary of semantic complexity of sentences in *Boya Chinese* textbooks is shown in Table 8. The sentence complexity metrics obtained by the method in [32] and the method proposed in this paper are different. Using the method in [32], the representation of the sentence semantic complexity is ratio, the minimum is 0.5, the maximum is 12, and the median is 2.42. The representation of the sentence semantic complexity in this paper is frequency, with a median of 7.45.

In order to compare the two methods, the mapping functions of sentence semantic complexity are constructed firstly, and the sentence semantic complexity is divided into 1–6. The larger the value is, the more difficult the sentence is. After statistics and analysis of the distribution of sentence semantic complexity, the constructed mapping functions are shown in Table 9 (it should be noted that, after the analysis of the sentences in texts, it is found that the diversity of lexical semantics is less in the sentences of the more difficult texts, so monotonic decreasing function is also constructed).

The two methods are used to analyze the sentences in the textbooks (*Boya Chinese*) and calculate the average, standard deviation, and confidence interval of the sentence semantic complexity of each level of text (assuming that the distribution of sentence difficulty in each level of text follows Gaussian distribution, a 95% confidence interval is constructed). The results are shown in Table 10. It can be seen that as the difficulty of the text increases, the average of the sentence semantic complexity obtained by the two methods increases, but relatively speaking, the sentence semantic complexity obtained by the method proposed in this paper is better distinguished in all levels of text.

Table 8: Sentence semantic complexity in Boya Chinese.

	Min	1/4 quantile	Median	3/4 quantile	Max
Paper [32]	0.50	1.95	2.42	3.00	12.00
This paper	1.00	2.09	7.45	29.73	5285.50

Table 9: Mapping functions.

Paper [32]	**a**	[0,1.5)	[1.5,2)	[2,2.5)	[2.5,3)	[3,3.5)	[3.5,∞)
	$f_1(a)$	6	5	4	3	2	1
This paper	**a**	**≤ 1**	**≤ 5**	**≤ 10**	**≤ 20**	**≤ 100**	**>100**
	$f_2(a)$	6	5	4	3	2	1

Table 10: Comparison of sentence semantic complexity.

Method	Level of text	Average	Standard deviation	Confidence interval
Paper [32]	Primary	3.10	1.54	[3.01,3.20]
	Intermediate	3.38	1.45	[3.32,3.43]
	Advanced	3.57	1.47	[3.51,3.62]
This paper	Primary	**2.43**	1.43	[2.34,2.52]
	Intermediate	**3.76**	1.63	[3.70,3.83]
	Advanced	**4.10**	1.62	[4.04,4.16]

Table 11: Correlation analysis of sentence semantic complexity and the text level.

Parameters	Paper [32]	This paper
Pearson correlation coefficient	0.11	0.31
T	26.03	43.81
Critical value (99%)	2.33	2.33

Due to the lack of Chinese sentence complexity tagging corpus, Pearson correlation coefficient is used to analyze the correlation between sentence semantic complexity and the text level. The results are shown in Table 11. The correlation coefficient of the method proposed in this paper is 0.31, which is significantly improved compared with the method of [32]. By constructing T to analyze the significance of correlation coefficient, T is not within the critical value ($-2.33 < T < 2.33$), which indicates that there is a significant positive correlation between sentence semantics complexity and the text level at 99% confidence level.

The effect of measurement method based on predicate semantic frameworks is better than that only considering the number of semantic categories in sentences. The reason should be that the measurement method based on PSFs combines structure and semantics and takes predication structure as semantic unit, which not only measures the semantic collocation relationship and quantity between sentence elements from a horizontal perspective, but also examines the hierarchical system and the primary secondary relationship from a vertical perspective. It is a comprehensive analysis of the number and nature of elements in a language system, as well as the number of connections between these different elements.

6. Conclusion

Based on the results of sentence-based syntactic analysis, this paper extracts the predication structures and converts the predication structures into PSFs. The spectral clustering method is used to cluster the semantic frameworks of each predicate to obtain their universality. Then according to the number and importance of PSFs at different layers of the sentence, the sentence semantic universality is obtained. Experiments show that the sentence semantic universality can well reflect the sentence semantic complexity. Furthermore, the method is compared with the method that only considers the semantic categories of words in the sentence. Experimental results show that the proposed method in this paper can effectively measure the sentence semantic complexity.

In this paper, the universality of PSFs is only considered from the collocation universality of subject, object, and predicate, ignoring the relationship between adverbial, complement, and predicate. However, adverbial is the grammatical component that modifies the predicate, and complement is the component that complements and explains the predicate. They are closely related to the predicate. In addition, a predication structure is a reflection of the basic propositional semantic of the sentence. In addition to the basic propositional semantic, the sentence semantics also contain the superpropositional semantics, such as modal semantic, tense and aspect semantic, and degree semantic, which will be considered in the subsequent work.

Data Availability

Sentence-based treebank and text corpus of international Chinese textbooks supporting this study have not been made available because the sentence-based treebank cannot be published until the relevant intellectual property protection application is completed. In addition, these textbooks belong to third party rights; the authors have no right to publish the data source.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61877004 and 62007004) and the Key Project of the National Social Science Foundation of China (Grant no. 18ZDA295).

References

[1] B. Bulté and A. Housen, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, John Benjamins, Amsterdam, The Netherlands, 2012.

[2] R. Carston and D. Blakemore, "Introduction to coordination: syntax, semantics and pragmatics," *Lingua*, vol. 115, no. 4, pp. 353–358, 2005.

[3] G. Leech, *Semantics: The Study of Meaning*, Penguin Books, Harmondsworth, UK, 2nd edition, 1981.

[4] L. Si, "Research on sentence semantic generation," *Foreign Languages and Their Teaching*, vol. 184, no. 7, pp. 4–7, 2004.

[5] Y. Hu and X. Fan, "Three planes of grammar research," *Language Teaching & Linguistic Studies*, vol. 2, pp. 4–21, 1993.

[6] Z. Dong and Q. Dong, "HowNet-a hybrid language and knowledge resource," in *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering*, pp. 820–824, Beijing, China, October 2003.

[7] E. Vyvyan, "Lexical concepts, cognitive models and meaning-construction," *Cognitive Linguistics*, vol. 17, no. 4, pp. 491–534, 2006.

[8] V. Evans and J. Zinken, "Figurative Language in a Modern Theory of Meaning Construction: A Lexical Concepts and Cognitive Models Approach," in *Art, Body And Embodiment*, Cambridge Scholars Press, Cambridge, UK, 2007.

[9] T. Mao, "Manual annotation approach to Chinese complex sentences by using bottom-up and top-down," *Journal of Chinese Computer Systems*, vol. 37, no. 4, pp. 716–721, 2016.

[10] H. Qin and L. Kong, "The impact of translational Chinese on original language: a syntactic complexity perspective," *Journal of Foreign Languages*, vol. 41, no. 5, pp. 17–28, 2018.

[11] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[12] X. Lu and H. Ai, "Syntactic complexity in college-level English writing: differences among writers with diverse L1 backgrounds," *Journal of Second Language Writing*, vol. 29, no. SI, pp. 16–27, 2015.

[13] Y. Wang, *A Study on the Measurement of Syntactic Complexity of Chinese as a Second Language*, Beijing Normal University, Beijing, China, 2015.

[14] W. Jiang, "Measurements of development in L2 written production: the case of L2 Chinese," *Applied Linguistics*, vol. 34, no. 1, pp. 1–24, 2013.

[15] J. E. Casal and J. J. Lee, "Syntactic complexity and writing quality in assessed first-year L2 writing," *Journal of Second Language Writing*, vol. 44, pp. 51–62, 2019.

[16] B. R. Ambati, S. Reddy, and M. Steedman, "Assessing Relative Sentence Complexity Using an Incremental CCG Parser," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1051–1057, San Diego, CL, USA, June 2016.

[17] F. Dell'Orletta, S. Montemagni, and G. Venturi, "Assessing document and sentence readability in less resourced languages and across textual genres," *International Journal of Applied Linguistics*, vol. 165, no. 2, pp. 163–193, 2015.

[18] S. Jiang, *Research on Sentence Difficulty Measurement*, Xiamen University, Xiamen, China, 2009.

[19] D. Yu, S. Wu, and C. Guo, "Assessing sentence difficulty in Chinese textbooks based on crowdsourcing," *Journal of Chinese Information Processing*, vol. 34, no. 2, pp. 17–26, 2020.

[20] B. Douglas, G. Bethany, and P. Kornwipa, "Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?" *Tesol Quarterly*, vol. 45, no. 1, pp. 5–35, 2012.

[21] L. Ortega, "Syntactic complexity in L2 writing: progress and expansion," *Journal of Second Language Writing*, vol. 29, pp. 82–94, 2015.

[22] H. Zaki and R. Ellis, "Learning Vocabulary through Interacting with Written Text," in *Learning a Second Language*

through Interaction, John Benjamins, Amsterdam, The Netherlands, 1999.

[23] R. Ellis and G. Barkhuizen, *Analyzing Learner Language*, Oxford University Press, Oxford, UK, 2005.

[24] W. Peng, J. Song, Z. Sui et al., "Formal schema of diagrammatic Chinese syntactic analysis," in *Proceedings of the 16th Chinese Lexical Semantics Workshop*, pp. 701–710, Beijing, China, May 2015.

[25] S. Zhu, Y. Zhang, W. Peng et al., "Construction of the basic sentence-pattern instance database based on the international Chinese textbook treebank," in *Proceedings of 2016 International Conference on Asian Language Processing*, pp. 266–270, Tainan, Taiwan, November 2016.

[26] J. Xu, J. Liu, and Y. Zhang, "Word similarity computing based on hybrid hierarchical structure by HowNet," *Journal of Information Science and Engineering*, vol. 31, no. 6, pp. 2089–2101, 2015.

[27] X. Yuan, "Research on the calculation of semantic similarity of HowNet," *Journal of Liarning University (Natural Science Edition)*, vol. 38, no. 4, pp. 358–361, 2011.

[28] K. Li and Y. Liu, "A spectral clustering algorithm based on self-adaption," in *Proceedings of 6th International Conference On Machine Learning And Cybernetics*, pp. 3965–3968, Hong Kong, China, August 2007.

[29] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Computing and Applications*, vol. 24, no. 7-8, pp. 1477–1486, 2014.

[30] C. Christina, V. Nicholas, and P. Ioannis, "Face clustering in videos based on spectral clustering techniques," in *Proceedings of 2011 First Asian Conference on Pattern Recognition*, pp. 130–134, Beijing, China, November 2011.

[31] L. Roger, F. Evelina, and G. Edward, "The syntactic complexity of Russian relative clauses," *Journal of Memory & Language*, vol. 69, no. 4, pp. 461–496, 2013.

[32] J. Zheng, "Lexical semantics and sentence difficulty measurement," in *Proceedings of Chinese Lexical Semantic Workshop*, pp. 261–265, Xiamen, China, April 2005.