ISSN 2055-0286
Advances in Social Sciences Research Journal; Vol. 13 No. 04 (2026) (217-231)
https://doi.org/10.14738/assrj.1304.20233

An Investigation into Text-Image Coherence in Vocabulary Teaching: A Case Study of the "Text-Image Interactive Units" in the New Grade 7 English Textbook (PEP)

Huiling Yu*, Yujie Su, Yangyang Cao, Xiaofan Qiu, Hongda Teng

1. School of Foreign Studies, Wenzhou University, Wenzhou, Zhejiang, China

Abstract: This study explores text-image coherence in the "Text-Image Interactive Units" of the 2024 new Grade 7 English textbook (PEP) using content analysis and case study, based on a four-dimensional framework (temporality, semantics, functionality, multimodality). Results show that text-image coherence in vocabulary teaching presents cross-volume features of word-image synchronization, direct matching and vocabulary recognition assistance, with obvious adaptability differences among parts of speech. Abstract vocabulary has prominent multi-dimensional coherence defects, and insufficient functional support is the core problem. Differentiated design by part of speech and cross-volume gradient adjustment are suggested for optimization. The innovation of this study is to break static text-image analysis, adopt a dynamic perspective of vocabulary teaching, combine multimodal symbol analysis with defect diagnosis, and make up for the limitations of previous static research.

Keywords: Text-image coherence, Vocabulary teaching, Text-Image Interactive Units, Grade 7 English textbook (PEP), Multimodality.

INTRODUCTION

The Compulsory Education English Curriculum Standards (2022 Edition)[15] clearly emphasizes the core position of "multimodal discourse teaching" and "viewing" skill, and textbook illustrations, as key visual resources, need to be used in conjunction with textual symbols to support the entire process of vocabulary acquisition, including "recognition, comprehension, and application." The PEP junior high school English textbook is one of the most widely used and influential mainstream textbooks in China. Its 2024 edition innovatively strengthens the design of "Text-Image Interactive Units" in the vocabulary teaching section, making it the core carrier of vocabulary acquisition. It not only conforms to the new curriculum standards, but also directly affects the teaching practice of millions of teachers and students, granting it high research representativeness and application value.

However, from the actual teaching situation, such units often encounter some problems when they are implemented in the classroom, such as cognitive breakdown caused by temporal misalignment, semantic ambiguity that hinders vocabulary internalization, and a single-function design that fails to support practical vocabulary application. These problems directly weaken the effectiveness of vocabulary teaching. Currently, research mostly focuses on the static mapping relationship between images and texts, such as the system of equal and unequal relations proposed by Martinec & Salway [12]. Nevertheless, there is little systematic exploration on the corresponding mechanisms of temporal, semantic, and functional interactions in vocabulary teaching. There is also a lack of relevant diagnosis for the cohesion bottleneck formed within words of different parts of speech.

Thus, such units are clearly unable to adapt to the constantly changing curriculum standards and the flexibility of course arrangements, indicating that there is a lot of room for adaptive development hidden within them.

This study addresses the current situation and gaps in previous research, focusing on the "Text-Image Interactive Units" in the new Grade 7 PEP textbook. It attempts to clarify the categories and distribution of illustrations and corresponding words in terms of collocation, and analyzes the differences in the deficiencies in the connection between various parts of speech, in order to find a connection optimization path that is suitable for the teaching process of "recognition, comprehension, application". Theoretically, the breakthrough point of this study lies in the first dissection of the process of vocabulary teaching's "recognition, comprehension, application" as a dynamic development process, and a detailed breakdown of the connection dimensions between illustrations and corresponding words in the "Text-Image Interactive Units" of textbooks, gradually unraveling the mechanisms underlying the graphic-text mapping of these units. Practically, the research results can provide quantitative standards and specific design guidelines for textbook writers to design the connection between illustrations and corresponding words, helping to further optimize and improve the new version of textbooks. Simultaneously, it can offer junior high school English teachers remedial strategies addressing coherence deficiencies categorized by part of speech and module, which is expected to gradually solve the difficulties in classroom teaching and effectively improve the implementation effect and promotion applicability of vocabulary teaching.

LITERATURE REVIEW

Textbook Illustrations represent a powerful mode of non-verbal communication that is highly effective and intuitive, constituting an important part of the pedagogic discourse system. In this study, the relationship between illustrations and target vocabularies in the recently published PEP junior high school English textbook will be explored. This chapter will provide a systematic literature review from two dimensions, which are multimodal learning theory and symbol synergy research, and the current research on illustrations in English textbooks. The former serves as the theoretical basis and frame of this coherence study, whereas the latter is focused on the practice in the use of illustrations in teaching English and their coherence problems and optimizing trends.

Multimodal Learning Theory and Symbol Synergy Research

The multimodal learning theory is a strong base for the analysis and evaluation of textbook images and building up an analytical framework. Conventional single-sensory learning is not capable of addressing the needs of meaning transmission effectively. Meaning generation and knowledge transfer do not depend on linguistic signs alone, but on the synergy among different modes of language, images, sounds, and movements [6]. Moreover, the connection between text and images influences learning. According to Mayer [14], relevant text and images help promote meaningful learning while unrelated images disrupt the process. These studies form the base of the analytical framework.

The following theories, sorted out in a logically coherent manner, are based on cognitive theory, semiotics, operational rules, and localized adaptations. These include:

Sweller's [18] Cognitive Load Theory, which reveals that the use of "picture-first, word-second" is more appropriate for vocabulary learning, providing the cognitive basis for temporal coherence. Multimodal Semiotic Synergy theory of Kress & van Leeuwen [7] proposes that the cognitive load can be minimized by making connections between different content through the synergistic effects of color, action, and scene, which is a basic starting point for the semiotic study of coherence. Mayer's [13] Multimedia Learning theory optimizes text-image coherence pathways through means such as "spatial contiguity" and "multiple representation," adding to the framework and norms of multimodal theory. Further, the Dual Coding Theory of Paivio [16] provides deep semantic coherence effectiveness by enhancing vocabulary efficiency with the help of bidirectional interaction between verbal and non-verbal memory mechanisms. Finally, domestic scholar Song Zhenshao [17], after combining both Chinese and international experiences, put forward rules such as "spatial contiguity" and "temporal matching" for illustrations in textbooks, translating cognitive principles and semiotic synergy into textbook analysis and providing empirical grounding for coherence evaluation.

At the specific application level of multimodal semiotic synergy, international core theories echo and expand with local research: Kress & van Leeuwen [8] proposed a grammar of visual design, in *Reading Images,* dividing visual signs into three categories: representational, interactive, and compositional meanings, clarifying the synergistic logic between multimodal signs and text-image coherence, and providing core basis for sign classification. van Leeuwen [20] further strengthened the principle of "semantic association", emphasizing that symbols such as color, action, and identification should serve the core meaning transmission and avoid redundant interference; Domestic scholar Gu Yueguo [4] classified multimodal symbols in textbooks into visual, auditory, and other types, providing an analytical framework suitable for local textbooks; Cheng Xiaotang [1] focuses on the image resources of English textbooks, reiterating the principle of "semantic association" of multimodal symbols, and pointing out the direction for multimodal connection in subject specific fields. These studies collectively constitute a theoretical framework for multimodal symbol segmentation and collaborative applications, providing direct support for the diagnosis and optimization of interface defects.

The theory of multimodal learning and the collaborative study of symbols progress and complement each other layer by layer: the former builds a core analytical framework for the connection between graphics and text from dimensions such as cognitive principles and memory mechanisms, while the latter clarifies the classification criteria and application principles of "semantic association" for multimodal symbols. Together, they provide a solid theoretical basis and practical guidance for the construction of the four-dimensional analysis framework for temporal, semantic, functional, and multimodal aspects in this article, as well as the diagnosis of segmentation and connection defects in multimodal symbols.

Current State of Research on English Textbook Illustrations

The research on illustrations in English textbooks has formed a relatively complete research system, focusing on the teaching value, application path, existing defects, and optimization direction of illustrations. It is deeply in line with the goals of "multimodal discourse teaching" and "core literacy cultivation" in the English curriculum standards, providing a

foundation and reference for the special research of the vocabulary module "Text-Image Interactive Units".

This body of research mainly focuses on the teaching scenarios and application strategies of illustrations, and organizes them according to the logic of general functions, vocabulary specialization, target textbooks, and cross-scenario reference, which can provide reference for the practical analysis of "Text-Image Interactive Units." Levin [9] categorized illustration functions into decorative, reinforcing, explanatory, and organizational types. This classification framework has been widely applied in designing coherence for the "recognition, comprehension, application" stages of vocabulary teaching, serving as a core reference for the division of functional coherence dimensions. Dou Mengting [3] specifically studied unit-theme-related vocabulary teaching, exploring how illustrations help establish form-meaning connections for vocabulary, also emphasizing the role of illustrations in basic vocabulary coherence, though without addressing dynamic coherence features. Dong Yawen [2], using the new Grade 7 PEP textbook as her research subject, analyzed the pathways for cultivating "viewing" skills through illustrations. Although not focusing on vocabulary modules, her study provides a contextually relevant reference for illustration application scenarios in the target textbook. Wu Shishu [21] explored application strategies for illustrations in listening and speaking teaching, proposing the core approach of "illustration context interpretation". Although it is not directly related to vocabulary teaching, it can provide reference for the design of connecting situations in the vocabulary application stage.

In terms of literature review on coherence defects and optimization direction, this area focuses on existing problems and improvement strategies for textbook illustrations. This area organizes the literature in terms of part-of-speech defects, structural defects, general matching defects, and illustration adaptation across textbook versions, which could serve as a foundation for the identification of defects in terms of coherence. Li Yanfen [10] directly pointed out that the core flaw of abstract vocabulary illustrations is semantic ambiguity, believing that illustrations in a single scene are difficult to accurately convey abstract meanings, but did not propose optimization solutions from the perspective of coherence. Luo Fang [11], through statistical analysis, identified a structural imbalance in textbook illustrations, characterized by "an overabundance of concrete types and a deficiency of abstract types," directly pointing to functional coherence defects, i.e., abstract vocabulary lacks suitable multimodal sign support. Hou Ying [5] mentioned in her analysis on the PEP senior high school textbooks that there was a problem with inadequate "text-image matching degree". Although the three-dimensional connection of junior high school vocabulary modules was not analyzed, it provided a core indicator of "matching" for defect diagnosis; Tan Ling [19] compared the graphic and textual relationships between the old and new People's Education Press textbooks and found that the graphic and textual relationships in the new textbooks are more closely related. Although there was no in-depth analysis of the differences in the connection defects of different parts of speech vocabulary, it provides a reference for the adaptability analysis of the new textbooks.

Summary

Existing research has laid a foundation for studying the coherence of illustrations and corresponding vocabulary across multiple dimensions—theory, practice, defects, and

multimodality. However, three core gaps remain. First, much research focuses on static theoretical construction or the application of a single function. It does not holistically analyze the dynamic process of vocabulary teaching ("recognition, comprehension, application") nor decompose coherence into the four dimensions of temporality, semantics, functionality, and multimodality, making it difficult to reflect the dynamic characteristics of coherence. Second, while some studies address defects and applications of textbook illustrations, none conduct specialized research on the "Text-Image Interactive Units" within vocabulary modules. There is a particular lack of targeted diagnosis for coherence defects across different parts of speech (e.g., concrete nouns, abstract adjectives, verbs, prepositions). Consequently, research findings are poorly suited to the vocabulary teaching needs of the new textbook. Third, a systematic "coherence evaluation framework" has not yet been constructed, and the synergistic logic between multimodal signs and the three-dimensional coherence aspects is unclear. This results in optimization suggestions lacking operability, unable to directly guide textbook compilation and actual teaching practices.

This study addresses these research gaps by taking the "Text-Image Interactive Units" in the new Grade 7 PEP textbook as its research object. It constructs a four-dimensional evaluation framework encompassing temporality, semantics, functionality, and multimodal analysis, focusing specifically on diagnosing coherence defects and optimizing methods for vocabulary across different parts of speech, thereby filling the gaps in existing research.

RESEARCH METHODS AND DESIGN

Research Object

This study focuses on all "Text-Image Interactive Units" in the first and second volumes of the new Grade 7 PEP English textbook. The first volume includes 3 Starter Units and 7 regular units (Units 1-7), while the second volume includes 8 regular units (Units 1-8). The study focuses on all vocabulary teaching units adopting the unified format of "English word + picture." Selection criteria were: vocabulary acquisition as the core objective, a direct association between illustration and corresponding word, and clear labeling of the teaching stage (e.g., Section A 1a). Based on the vocabulary lists in the books, 14 valid analysis units were identified (8 in the first volume, 6 in the second), encompassing 110 core vocabulary items. Classified by part of speech, these include: 64 concrete nouns; 6 abstract nouns; 15 verb phrases; 8 prepositions; 4 time-related phrases; 4 weather-related adjectives; and 9 collective nouns.

Research Methods

This study employs a combination of content analysis and case study methods. For the content analysis, a coding table was designed based on the four-dimensional framework : "temporality, semantics, functionality, multimodality". Before formal coding, two researchers underwent pre-coding training to reach a consensus on the coding standards. After formal coding, Cohen's Kappa coefficient was used to test coding consistency, yielding a Kappa value of 0.89 (≥0.75), meeting reliability requirements. For the case study method, typical coherence defect cases across different parts of speech and textbook volumes were selected for in-depth analysis. Combined with multimodal sign theory, this analysis

examined coherence issues under the unified format and cross-volume common characteristics, providing empirical support for optimization pathways.

Research Tool

The core research tool is a cross-volume, cross-part-of-speech Coherence Coding Table. Designed based on four primary dimensions—temporality, semantics, functionality, and multimodal sign supplementation—it incorporates detailed explanations for secondary dimensions, refined based on the vocabulary characteristics and textbook compilation realities of the Text-Image Interactive Units in the new Grade 7 PEP textbook (both volumes). Typical vocabulary examples from the textbook are integrated to enhance the coding table's practicality and adaptability, ensuring standardization and consistency in the subsequent content analysis. The specific coding table is shown in Table 1. Here, temporality, semantics, and functionality are the core coherence dimensions, serving as the direct basis for determining coherence defects. Multimodal sign supplementation is an auxiliary dimension; it enhances the effectiveness of the core dimensions through the synergy of visual elements. Insufficient synergy in this auxiliary dimension can indirectly induce defects in the core dimensions but is not independently classified as a defect.

Table 1: Cross-Volume, Cross-Part-of-Speech Coherence Coding Table

Primary Dimension	Secondary Dimension	Coding Description
Temporality	Word-Image Synchrony	Vocabulary text and corresponding picture are presented simultaneously on the page in a left/right or top/bottom layout. This is the basic textbook layout format, suitable for most concrete nouns, e.g., *ruler*, *fox*, *watermelon*.
	Picture before Word	The complete visual content of the picture is presented first on the page, and the corresponding vocabulary text is then labeled next to/below it. This establishes a visual representation before assigning the textual symbol. Suitable for prepositions, verb phrases, collective nouns, etc., e.g., *in front of*.
Semantics	Direct Matching	The picture content precisely restores the entity, attribute, or concrete object referred to by the vocabulary. Form and meaning correspond completely with no informational deviation, e.g., concrete nouns/color nouns like *red*, *baseball glove*, *desk*.
	Scene Association	The picture does not directly present the vocabulary itself but indirectly conveys its meaning by showcasing its application scenario, background environment, or associated context, e.g., *visited a science museum*, *pet dog*.
	Vague Association	The semantic connection between the picture and the vocabulary is weak. It only reflects a superficial association and cannot accurately convey the core meaning or complete attributes, easily leading to semantic misunderstanding, e.g., *stormy*, *dry*.
Functionality	Recognition Assistance	The picture plays only a basic role in aiding recognition, helping students establish a simple association between the vocabulary text and its visual image, without additional semantic explanation or usage hints. It satisfies only the need for vocabulary identification, e.g., abstract nouns for subjects like *biology*, *maths*, *IT*.
	Comprehension Reinforcement	Building on recognition, the picture further helps students clarify the core semantics of the vocabulary, differentiate between near-

		synonyms, or understand the word's collocation usage, e.g., *between* (clarifies positional relationship), *baseball* and b*aseball glove* (differentiates collocations), *across from* (clarifies directional semantics).
	Application Support	The picture constructs a specific usage context or life scenario for the vocabulary, helping students understand its practical application context. It can directly support pragmatic expression like sentence making and dialogue, e.g., *Chinese chess*, *play baseball*.
Multimodal Sign Supplementation	Scene Elements	The picture contains scene-based visual content related to the vocabulary, such as characters, natural environments, life scenarios, or architectural backgrounds, used to enrich the vocabulary's contextual information, e.g., *a red house* (house and surrounding environment), *art club* (studio scene).
	Action Elements	The picture contains action-based visual content related to the vocabulary, such as dynamic behaviors, body movements, or operational details. Suitable for verb phrases and action-related vocabulary, e.g., *cooked food* (cooking action), *playing volleyball* (action of playing).
	Color Elements	The picture uses specific color combinations or color highlighting to imply the vocabulary's attributes, characteristics, or connotations. Suitable for color nouns and concrete nouns with color characteristics, e.g., *black and white cows*, *red*, *yellow*.
	Sign Elements	The picture contains sign-based visual content related to the vocabulary, such as text labels, directional arrows, icons, or numerical markers, used to precisely convey vocabulary semantics, e.g., clock numbers for time-related phrases, positional arrows for prepositions, prohibition/indication symbols.

RESULTS AND ANALYSIS

Types and Distribution of Coherence Between Illustrations and Corresponding Vocabulary

The coherence of the Text-Image Interactive Units exhibits a pronounced feature of cross-volume concentration. Regarding temporal coherence, as shown in Table 2, "Word-Image Synchrony" accounts for 70%. In the first volume, it is mainly concentrated in the basic vocabulary matching tasks; in the second volume, it covers nouns like animals, food, and sports equipment. "Picture before Word" accounts for 30%, concentrated in the first volume for prepositions, concrete nouns, and collective nouns.

Table 2: Distribution of Temporality Coherence Types

Temporality Coherence Type	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
Word-Image Synchrony	77	70.00%
Picture before Word	33	30.00%
Total	**110**	**100.00%**

Semantic coherence is reflected in Table 3. "Direct Matching" occupies the largest share of 35.45% and applies to concrete nouns. "Direct Matching + Scene Association" takes up 27.27%, referring to the dual matching of entities and scenes. "Scene Association" accounts for 26.36% and focuses on the non-entity abstract vocabulary. "Vague Association"

occupies 10.91% and includes vocabulary with incomplete semantic transmission. The vagueness is generally found in collective nouns in both volumes.

Table 3: Distribution of Semantic Coherence Types

Semantic Coherence Type	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
Direct Matching	39	35.45%
Direct Matching + Scene Association	30	27.27%
Scene Association	29	26.36%
Vague Association	12	10.91%
Total	**110**	**100.00%**

"Recognition Assistance" takes up the largest proportion of 37.27%. It is used to teach basic vocabulary knowledge in two volumes. "Recognition Assistance+Comprehension Reinforcement" accounts for 27.27% and serves as the distinction between similar vocabulary and deepens their semantic differences. "Recognition Assistance+Application Support" makes up 25.45%. It helps apply vocabulary in context. "Comprehension Reinforcement" occupies 7.27% and focuses solely on semantic difference reinforcement. The last one "Comprehension Reinforcement+Application Support" takes up only 2.73%.

Table 4: Distribution of Functionality Coherence Types

Functionality Coherence Type	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
Recognition Assistance	41	37.27%
Recognition Assistance + Comprehension Reinforcement	30	27.27%
Recognition Assistance + Application Support	28	25.45%
Comprehension Reinforcement	8	7.27%
Comprehension Reinforcement + Application Support	3	2.73%
Total	**110**	**100.00%**

The use of multimodal signs in teaching vocabulary also differs significantly in contribution ratios among various categories (Table 5). Multimodal signs can be used in more than one type for some vocabulary items. Among all categories of multimodal signs, scene elements contribute the most vocabulary items, totaling 79 vocabulary items, which take up 71.82% of the total number of vocabulary. They form the key signs that help deepen contextual information and semantics. Scene elements account for the greatest amount (71.82%), followed by sign elements (59.09%). Most vocabulary items combine two or more sign types to create synergistic effects.

Table 5: Distribution of Multimodal Sign Types and Contribution Rates

Multimodal Sign Type	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
Scene Elements	79	71.82%
Sign Elements	65	59.09%
Color Elements	39	35.45%
Action Elements	34	30.91%

Differences in Coherence Deficits Across Parts of Speech

The adaptability of the unified format varies significantly across different parts of speech. The coherence deficits exhibit a distribution pattern tied to parts of speech and concentrated in specific dimensions. The judgment criteria are shown in Table 6.

Table 6: Criteria for Determining Text-Image Coherence Deficits

Coherence Deficit Type	Core Determination Dimension	Applicable Parts of Speech	Rule for Determining Deficit
Temporal Deficit	Whether the order of presenting text and image aligns with cognitive rules for vocabulary acquisition.	Prepositions, verb phrases, collective nouns, time-related phrases, weather-related adjectives, abstract nouns. Concrete nouns: uniformly judged as having no temporal deficit.	Adopts Word-Image Synchrony.
Semantic Deficit	Whether the illustration accurately conveys the deeper meaning, logical relationships, or categorical attributes of the vocabulary.	All parts of speech.	Non-concrete nouns: adopts Direct Matching or Vague Association. Concrete nouns: adopts Vague Association.
Functional Deficit	Whether the illustration supports the cross-volume pedagogical gradient needs of "recognition, comprehension, application."	All parts of speech.	First volume: Only achieves Recognition Assistance (i.e., only helps students recognize the word, no comprehension/application support). Second volume: Only achieves Recognition Assistance or Recognition Assistance + Comprehension Reinforcement, failing to meet Application Support requirements.

At the temporality coherence level, 30 vocabulary items have deficits (27.27%), existing only in non-concrete vocabulary such as abstract nouns and verb phrases. The distribution is shown in Table 7.

Table 7: Distribution of Vocabulary Types with Temporal Coherence Deficits

Temporal Deficit	Part of Speech	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
No	Collective Nouns	8	7.27%
	Prepositions	8	7.27%
	Concrete Nouns	64	58.18%
		80	**72.73%**
Yes	Abstract Nouns	6	5.45%
	Verb Phrases	15	13.64%
	Collective Nouns	1	0.91%
	Time-related Phrases	4	3.64%
	Weather-related Adjectives	4	3.64%

		30	**27.27%**
Total		**110**	**100.00%**

At the semantic coherence level, 14 vocabulary items have deficits (12.73%), concentrated in abstract nouns, time-related phrases, and weather-related adjectives. The distribution is shown in Table 8.

Table 8: Distribution of Vocabulary Types with Semantic Coherence Deficits

Semantic Deficit	Part of Speech	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
No	Verb Phrases	14	12.73%
	Collective Nouns	9	8.18%
	Prepositions	8	7.27%
	Concrete Nouns	64	58.18%
	Weather-related Adjectives	1	0.91%
		96	**87.27%**
Yes	Abstract Nouns	6	5.45%
	Verb Phrases	1	0.91%
	Time-related Phrases	4	3.64%
	Weather-related Adjectives	3	2.73%
		14	**12.73%**
Total		**110**	**100.00%**

Functional coherence represents the main weakness in text-image coherence, with 68 defective vocabulary items, accounting for 61.82% of the total, covering most parts of speech. Among these, concrete nouns account for the highest proportion of functional deficits, becoming the core issue in functional coherence. The specific distribution is shown in Table 9.

Table 9: Distribution of Vocabulary Types with Functional Coherence Deficits

Functional Deficit	Part of Speech	Number of Vocabulary Items	Percentage of Total Vocabulary (%)
No	Verb Phrases	3	2.73%
	Collective Nouns	9	8.18%
	Prepositions	8	7.27%
	Concrete Nouns	22	20.00%
		42	**38.18%**
Yes	Abstract Nouns	6	5.45%
	Verb Phrases	12	10.91%
	Concrete Nouns	42	38.18%
	Time-related Phrases	4	3.64%
	Weather-related Adjectives	4	3.64%
		68	**61.82%**
Total		**110**	**100.00%**

Analysis of the unique coherence features of each part of speech: Concrete nouns are the largest set (64 items, 58.18% of total). Their defects occur only in the functionality

dimension (42 items, 65.6% of concrete nouns), while they have no temporal or semantic deficits. The prepositions (8 in number, or 7.27% of total) are the only part of speech without any coherence deficits in the three dimensions: temporality, semantics, and functionality, thus making them highly adaptable. Verb phrases, totaling 15 (13.64% of total), have a 100% temporal deficit rate, and 80% of them also have functional deficits. The dual deficit of temporality and functionality is the core issue, e.g., met up with friends uses Word-Image Synchrony and only shows people greeting. Collective nouns (9 in number, or 8.18% of total) demonstrate high adaptability with just one word showing a temporal deficit. Scene illustrations are effective in representing this category and the collective nature of this part of speech. Abstract nouns (6 in number, or 5.45% of total) are the most inadaptably coherent. For example, art, maths are paired only with simple tool images, failing to convey the abstract meaning of the subjects and completely missing teaching needs. Time-related phrases and weather-related adjectives, each totaling 4 (each 3.64% of total), exhibit highly consistent defect characteristics. For instance, a quarter to two lacks time logic markings, and cloudy lacks attribute comparisons with similar words, making it difficult to support deeper understanding and contextualized application.

DISCUSSION

Analysis of the Causes Behind the Distribution of Coherence Types

The cross-volume concentration pattern of "Word-Image Synchrony, Direct Matching, Recognition Assistance" in the Text-Image Interactive Units essentially results from the textbook compilation's effort to balance unity, coherence, and practicality. The current design matches the cognitive theory of Dual Coding proposed by Paivio and Multimedia Learning put forward by Mayer because of its consistency with the image-based cognitive ability of seventh-grade students. In addition, the empirical data analysis of "70.00% Word-Image Synchrony; 35.45% Direct Matching; 37.27% Recognition Assistance" reveals the current design's consistency with standardization, coherence, and practicability in large-scale teaching, vocabulary acquisition, and foundational teaching.

From the viewpoint of multimodal signs, the current distribution is based on "Scene Elements and Sign Elements as primary; Color Elements and Action Elements as secondary." That is, when choosing easily-understood and universal visual signs in terms of the textbook compilation, scene elements and sign elements, which fit the image-based cognitive ability of seventh-grade students, have been used. The elements can establish basic contexts for most vocabulary. However, Color Elements and Action Elements are not universal and applicable to all scenes, which result in semantic ambiguity and insufficiency in the function of verb phrases and abstract nouns due to the absence of specialized multimodal sign synergy.

Core Reasons for Differences in Coherence Deficits Across Parts of Speech

The varying adaptability across parts of speech is closely related to vocabulary attributes, cognitive principles, cross-volume concentrations, and the suitability of multimodal sign synergy. Prepositions show no coherence deficits as their design aligns with cognitive principles for positional relationships. Concrete nouns only have functional deficits. This is because these vocabulary items are easy to visualize. However, the cross-volume design

fails to reinforce application contexts through multimodal sign synergy according to the teaching gradient, resulting in insufficient functional support. Collective nouns have only one item with a temporal deficit; overall, their group scene illustrations suit collective meaning transmission. Verb phrases exhibit dual temporal and functional deficits: they uniformly use Word-Image Synchrony, violating cognitive principles for action vocabulary, and static pictures cannot present complete action processes. In addition, there is no optimal combination of multimodal signs for the applicative learning gradient in Volume Two, with only some of the words conforming to the need for application support. There are deficits in all three areas in the abstract nouns in that there are no concrete objects, appropriate signs, and good visualization for abstract nouns. Time-based expressions and weather-related adjectives have deficiencies in all three aspects as well, and need a Picture before Word design as well as sign elements, neither of which is present in the current design.

Comparison and Complement to Existing Research

This study forms an effective dialogue and supplement with existing research: it is consistent with the research conclusion of Li Yanfen [10] on "semantic ambiguity in abstract vocabulary illustrations", but further reveals the distribution of defects in prepositions, verb phrases, and other parts of speech, as well as the core causes of multimodal symbol singularity, and clarifies the cross-volume commonality of defects; It supports the viewpoint of Luo Fang [11] that there are too many physical illustrations in textbooks and insufficient abstract ones, and refines the adaptation differences of cross-part-of-speech vocabulary in specific modules of both volumes; It supports Hou Ying's (2022) research on "insufficient matching between text and image", and constructs a cross-part-of-speech coherence evaluation framework based on the whole of the upper and lower volumes; It aligns with Tan Ling's [19] finding that "the new textbook shows closer text-image associations", but it points out the underlying part-of-speech adaptability defects and insufficient cross-volume optimization behind these associations.

The innovation of this study lies in focusing for the first time on the overall "Text-Image Interactive Units" of the first and second volumes of seventh grade, deconstructing coherence patterns under a unified format, and proposing an optimization pathway of "unified format, differentiated details, cross-volume coherence". This study also includes the lack of multimodal symbol collaboration as the core cause of cohesion defects, improving the existing research perspective that only focuses on static matching of graphics and text, and making the diagnosis of cohesion defects more targeted.

Research Limitations

This study has certain limitations. The sampling process involves only the seventh-grade volumes and excludes the eighth- and ninth-grade volumes. The research methodology mainly employs textual analysis while excluding any empirical evidence from classroom instruction. There is no consideration for auditory signs such as audio in analyzing multiple signs.

CONCLUSION

The core conclusions are as follows: The “Text-Image Interactive Units” in the new seventh-grade PEP English textbooks (Volumes 1 & 2) adopt a unified “English word + picture” format. They exhibit a cross-volume concentration pattern characterized by “temporality mainly as Word-Image Synchrony, semantics mainly as Direct Matching, and functionality mainly as Recognition Assistance.” This highly aligns with foundational vocabulary teaching goals but shows insufficient application support and lacks clear cross-volume optimization.

The adaptability of the unified format varies significantly across parts of speech. Prepositions show the best adaptability; concrete and collective nouns show relatively good adaptability; verb phrases have dual temporal and functional deficits; abstract nouns, time-related phrases, and weather-related adjectives show the poorest adaptability. Defect formation is closely related to vocabulary attributes, cognitive principles, and multimodal sign synergy.

Multimodal signs serve as a core supplementary support for the three-dimensional coherence (temporality, semantics, functionality). Among them, scene elements and sign elements are most effective in enhancing coherence effectiveness. The synergistic interplay of the four-dimensional framework is key to achieving efficient text-image coherence.

Based on the research conclusions, the following practical recommendations are proposed.

In terms of textbook writing, part-of-speech adaptation and cross-volume coherence of temporal connections should be optimized. In the first volume, basic concrete nouns should retain the “word-image synchrony” form, while abstract nouns, verb phrases, time phrases, and weather adjectives should adopt the “graph first word later” temporal sequence uniformly throughout the entire volume. Prepositions should maintain the existing “graph first word later” design, and collective nouns should only be fine tuned for individual temporal defects. The second volume should strengthen the standardization of temporal sequence design, enhance the accuracy of semantic coherence, and supplement prepositions with positional identification elements; Verb phrases should be presented with more details, while abstract nouns, time, and weather adjectives are supplemented with connotation associations, logical contrast markers, and scene elements; The cross-volume concentration of rich functional connections should be enriched, with the first volume focusing on “word recognition assistance+basic understanding” and the second volume strengthening “understanding enhancement+application support”; Establish a quantitative standard for cross-volume coherence and clarify the coherence requirements for different parts of speech and vocabulary in different volumes.

In terms of teaching applications, teachers should make targeted efforts to address the gap in cross-volume and cross-part-of-speech connections, supplement connotation related materials for abstract nouns and carry out scene association activities, supplement action process demonstrations and contextual application exercises with illustrations for verb phrases, and design attribute comparisons and logical sorting tasks for time/weather adjectives; Collective nouns can simply supplement the classification activities of similar items for individual defective vocabulary; Regarding prepositions, use teaching aids to restore the core positional relationships in the image, and explain the usage of prepositions such as “in front of” and “between” across volumes. Optimize the timing presentation strategy, adjust the order of graphic and textual presentation according to part of speech

and volume in teaching, guide students to observe pictures and infer semantics for abstract nouns, verb phrases, and time/weather adjectives before presenting vocabulary, maintain the existing presentation form of prepositions before words, and only fine tune individual vocabulary for collective nouns. Strengthen the collaborative teaching of multimodal symbols, guide students to focus on the scene, identification, action and other elements in high-frequency vocabulary illustrations such as abstract nouns and verb phrases, and carry out activities such as "looking at pictures to say words, cross-volume vocabulary connection" in combination with speech, action and other modalities. Expand the application scenarios of illustrations and design extension activities based on the illustrations in the first and second volumes. For example, use the noun illustrations in the first volume to create a "backpack sorting list", and combine the illustration design of abstract nouns and verb phrases with subject applications and action description tasks. Deeply integrate illustrations with vocabulary applications to achieve coherent mastery of cross-volume vocabulary.

Future research can further expand the research sample to cover eighth and ninth grade textbooks and analyze the developmental patterns of transitional learning stages; Conduct empirical research on teaching to verify the impact of optimized linkage paths on vocabulary learning outcomes; Expand the analysis dimensions of multimodal symbols, including auditory, motor, and other modalities, and improve the coherence evaluation framework; Compare the cross-volume and cross-part-of-speech graphic and textual design of different versions of textbooks to provide a more comprehensive reference for textbook writing.

REFERENCES

[1]. Cheng, X.T. and Cong, L., The design and use of image resources in English textbook compilation. *Curriculum, Teaching Material and Method*, 2020. 40(8): p. 78-85.

[2]. Dong, Y.W., Research on the strategies of illustration and function utilization in junior high school English textbooks based on improving "viewing" skills. *Progress in Education*, 2025. 15(8): p. 546-552.

[3]. Dou, M.T., The application of textbook illustrations in junior high school English unit theme vocabulary teaching. *Campus English*, 2021. 39: p. 111-112.

[4]. Gu, Y.G., An analysis of multimedia and multimodal learning. *Computer-Assisted Foreign Language Education*, 2007. 102(2): p. 3-12.

[5]. Hou, Y., A study on the relationship between images and texts in the new People's Education Press senior high school English textbooks. *Education Observation*, 2022. 11(35): p. 109-113.

[6]. Jewitt, C., Multimodality, "reading", and "writing" for the 21st century. *Discourse: Studies in the Cultural Politics of Education*, 2005. 26(3): p. 315-331.

[7]. Kress, G. and van Leeuwen, T., *Multimodal Discourse: The Modes and Media of Contemporary Communication*. 2001, London: Arnold.

[8]. Kress, G. and van Leeuwen, T., *Reading Images: The Grammar of Visual Design*. 2nd ed. 2006, London: Routledge.

[9]. Levin, J.R., Pictures as Information Processors: Thematic Variations with Educational Applications, in *Knowledge Acquisition from Text and Pictures*, H. Mandl and J.R. Levin, Editors. 1989, Amsterdam: North-Holland. p. 253-280.

[10]. Li, Y.F., Enlightening wisdom with images: textbook illustrations helping improve the efficiency of junior high school English teaching. *Curriculum, Teaching Material and Method*, 2025. 14: p. 62-64.

[11]. Luo, F., A study on the structural imbalance of textbook illustrations. *Curriculum, Teaching Material and Method*, 2022. 42(9): p. 78-82.

[12]. Martinec, R. and Salway, A., A system for image–text relations in new (and old) media. *Visual Communication*, 2005. 4(3): p. 337-371.

[13]. Mayer, R.E., Multimedia Learning: Are We Asking the Right Questions? *Educational Psychologist*, 1997. 32(1): p. 1-19.

[14]. Mayer, R.E., Multimedia learning. *Psychology of Learning and Motivation*, 2002. 41: p. 85-139.

[15]. Ministry of Education of the People's Republic of China, *Compulsory Education English Curriculum Standards (2022 Edition)*. 2022, Beijing: Beijing Normal University Press.

[16]. Paivio, A., *Mental Representations: A Dual Coding Approach*. 1986, New York: Oxford University Press.

[17]. Song, Z.S., A cognitive psychological study of textbook illustrations. *Journal of Beijing Normal University (Social Science Edition)*, 2005. 42(6): p. 22-26.

[18]. Sweller, J., Instructional Design in Technical Areas. *Australian Journal of Education*, 1999. 43(1): p. 5-23.

[19]. Tan, L., A comparative study of the relationship between images and texts in the old and new People's Education Press junior high school English textbooks. *Progress in Education*, 2025. 15(1): p. 1063-1073.

[20]. van Leeuwen, T., *Discourse and Practice: New Tools for Critical Discourse Analysis*. 2008, Oxford: Oxford University Press.

[21]. Wu, S.S., The interpretation and application strategies of textbook illustrations in integrated listening and speaking teaching of junior high school English. *English Teacher*, 2022. 22(12): p. 89-91.