

Multimodal Discourse Analysis of the Interpersonal Meaning and Interactive Meaning in Documentary *China Bouquet*

Renyi Xia

School of Jinan University, Guangzhou 510000, China

Abstract: With the development of media technology and the rise of visual culture, discourse communication increasingly relies on the coordinated expression of multiple semiotic resources. As a result, multimodal discourse analysis has emerged as a significant and expanding field in linguistics. This study investigates the documentary *China Bouquet* within a multimodal discourse analysis framework, focusing specifically on how interpersonal relationships and interactive meanings are constructed through multimodal resources. Drawing on Systemic Functional Grammar and visual grammar theory, the study analyzes how visual and linguistic modes collaborate to express attitudes, emotions, and stances, thereby achieving interactive meaning. Representative segments from the documentary are selected for analysis to uncover the realization of cultural dissemination and audience engagement.

Keywords: Multimodal Discourse Analysis, Interpersonal Meaning, Interactive Meaning.

1. Introduction

With the rapid development of digital media technologies, information dissemination is no longer limited to written language, and multimodal communication has become a dominant form of everyday interaction. This shift has expanded linguistics from the study of language itself to the exploration of various semiotic modes. Multimodality is understood as composite discourse featuring not only text but also visual elements such as images and charts; equivalently, it may refer to any text in which meaning is encoded through more than one semiotic resource[1]. Based on Halliday's Systemic Functional Linguistics, Kress and van Leeuwen developed Visual Grammar by extending the three metafunctions of language to visual modes such as images and layout, establishing multimodal discourse analysis as an important research field.

Among various genres of multimodal texts, documentaries stand out as particularly potent vehicles for national narrative construction, cultural dissemination, and audience engagement. By integrating multiple modes, including language, images, and sounds, documentaries construct meanings and influence viewers' perceptions. However, existing studies have mainly focused on news media, advertisements, and social media, while the interpersonal and interactive dimensions of documentaries remain insufficiently explored from a multimodal perspective.

This study focuses on the short documentary *China Bouquet*, which presents the inheritance and transformation of traditional Chinese civilization from a contemporary perspective and reflects the values and spiritual aspirations of Chinese people. Employing a mixed-methods approach, this study conducts a multimodal discourse analysis of *China Bouquet* based on Systemic Functional Grammar and Visual Grammar, aiming to investigate the construction of interpersonal and interactive meanings in the documentary.

2. Literature Review

Roland Barthes was one of the earliest scholars to explore

the relationship between images and language in meaning-making. In *Rhetoric of the Image* (1977), he examined how visual and verbal elements interact in communication, providing a foundation for later multimodal discourse studies. However, multimodal discourse analysis is primarily grounded in Halliday's Systemic Functional Linguistics (SFL), which views language as a social semiotic system with three metafunctions: ideational, interpersonal, and textual[2,3].

Building on SFL, O'Toole (1994) extended Halliday's metafunctions to visual analysis and argued that images could also function as discourse. He proposed three corresponding visual functions: representational, modal, and compositional. Based on Halliday's theories, Kress and van Leeuwen (1996) developed Visual Grammar, systematically exploring how visual modes construct meaning through representational, interactive, and compositional functions. Their studies further expanded multimodal research to areas such as visual imagery, color, media, and layout design[4]. With the development of multimodal discourse analysis internationally, Chinese scholars have also made significant contributions to this field. Li Zhanzi introduced Kress and van Leeuwen's Visual Grammar into China and discussed its implications for multimodal discourse analysis and English language teaching[5]. Hu Zhuanglin further explored multimodal semiotics and clarified the relationship between multimodality and multimedia[6]. Zhang Delu proposed an integrated framework for analyzing multimodal discourse from multiple levels, including culture, context, discourse meaning, grammar, and medium[7,8]. Since the 2010s, multimodal studies in China have expanded into fields such as education, advertising, film, and media.

Research on documentaries has been conducted from various perspectives, including rhetoric, genre studies, discourse analysis, pragmatics, and increasingly, multimodal analysis. In China, documentary studies have mainly focused on subtitle translation, imagology, narrative analysis, and multimodal discourse analysis. Although existing studies have provided valuable insights into documentary discourse

from various perspectives, they have primarily focused on translation, narrative strategies, and national image construction. The interpersonal and interactive meanings constructed through the interaction of verbal and visual modes in documentaries remain relatively underexplored.

3. Theoretical Framework

3.1. Interpersonal Meaning in Systemic Functional Grammar

SFG holds that language has three metafunctions: the ideational, the interpersonal, and the textual. Within this theoretical framework, the interpersonal metafunction refers to the interactive relationship between the speaker and the listener (or the writer and the reader) in a discourse context, as well as the speaker's attitude toward what is being said or written. Interpersonal meaning is primarily realized through mood and modality.

Mood refers to how the grammatical structure of a clause embodies speech roles and exchange functions, while modality concerns the speaker's expression of the probability, necessity, or willingness regarding a proposition or proposal, reflecting attitudes and judgments. The mood system plays a central role in constructing interpersonal meaning. According to Halliday (2004), communication involves two fundamental speech roles: giving and demanding. When combined with two types of exchange, these roles generate four speech functions: offer, statement, command, and question. These functions are typically realized through three major clause types: declarative, interrogative, and imperative clauses. Statements are most naturally expressed by declarative clauses; questions by interrogative clauses; and commands by imperative clauses. The Mood element consists of the Subject and Finite components of a clause [9].

In SFG, modality refers to the intermediate space between positive and negative polarity and reflects the speaker's evaluation of information or attitude toward an action. It indicates degrees of probability, usuality, obligation, or inclination. Halliday categorizes modality into two major types: modalisation and modulation. Modalisation concerns the degree of probability and usuality, while modulation relates to obligation and inclination. Modality can be realized through modal operators and mood adjuncts. Since modality involves different degrees of commitment, Halliday (1994) further divides it into three values: high, median, and low.

3.2. Interactive Meaning in Visual Grammar

In 1996, Kress and van Leeuwen published *Reading Images: The Grammar of Visual Design*, in which they extended Halliday's systemic functional theory to visual communication and proposed the theory of Visual Grammar. They argued that images, like language, are meaningful systems through which meaning is constructed. Based on Halliday's three metafunctions, they identified three visual metafunctions: representational meaning, interactive meaning, and compositional meaning, providing a theoretical foundation for multimodal discourse analysis.

Visual Grammar emphasizes how images construct relationships between represented participants and viewers. Interactive meaning, as one of the three metafunctions, concerns the interaction among image producers, represented participants, and viewers [9]. It is mainly realized through four elements: contact, social distance, perspective, and modality. Contact refers to whether participants directly engage with

viewers through gaze, distinguishing between demand and offer images. Social distance reflects the degree of closeness between participants and viewers through different framing choices. Perspective indicates the viewer's position and attitude toward the represented participants, with different angles conveying different relationships of involvement, equality, or power. Modality refers to the degree of realism and credibility represented in images, which can be constructed through visual resources such as color, background, detail, depth, and illumination. Through these visual choices, images establish specific relationships with viewers and influence how meanings are interpreted. Therefore, Visual Grammar offers an effective analytical framework for examining how visual elements convey meanings and construct interactions in multimodal discourse.

4. Results

4.1. Analysis of Interpersonal Meaning in Text Modality

This study aims to analyze the interpersonal meaning in the documentary *China Bouquet*, with a particular focus on mood and modality in the subtitles.

4.1.1. Interpersonal Meaning in Mood System

The mood structure of a clause typically consists of two core components: the Subject and the Finite. The sequencing of these elements determines the mood type of a clause and the interpersonal function. According to SFG, there are three basic types of mood: declarative, interrogative and imperative. The analysis adopts the clause as the basic unit. Through manual annotation, there are a total of 923 clause in the subtitles.

Table 1. Mood types of the subtitles

	Declarative	Imperative	Interrogative	Total
Frequency	903	8	10	921
Percent	98%	0.9%	1.1%	100%

The statistical data shows that declarative statement account for the largest proportion among all clause at 98%, while imperative command and interrogative question each represent a smaller share.

Declarative primarily conveys information and facts. The use of declarative mood establishes an asymmetrical relationship between the speaker as an information provider and the audience as information receivers.

Example 1: After weathereing, grinding, and other processes the sediment becomes the purple clay for porters to use. In the hands of a pottery master, the purple clay comes to life.

These statements are used to present the transformation process of purple clay in a chronological and objective manner. The frequent use of declarative clauses in presenting traditional crafts constructs the narrator as an authoritative cultural interpreter. Rather than merely transmitting information, this objective narrative style enables the documentary to present traditional culture as reliable knowledge and strengthens viewers' recognition of cultural values.

The imperative mood is typically used to express commands, requests, suggestions, or instructions. Compared with declarative clauses, imperatives tend to create a more interactive relationship by directly engaging the audience.

Example 2: Add a few dry leaves, pour boiled water into the cup...

In this extract, imperative clauses provide direct instructions related to tea preparation, positioning the narrator as a guide and the audience as participants in the communicative process. Rather than imposing commands, these imperatives create an engaging and intimate interaction, reducing the distance between the narrator and viewers.

Interrogative mood contributes to interpersonal meaning by initiating interaction and engaging the audience. It can be realized through polar questions and wh-questions, which respectively seek confirmation and specific information.

Example 3: What is the symbol of Chinese food?

Example 4: Can we still detect traces of Chan today?

In *China Bouquet*, these interrogatives transform the

audience from passive receivers into active participants in meaning-making. The wh-question introduces a cultural inquiry and creates expectations for further explanation, while the polar question invites viewers to reflect on the continuity of traditional culture. Through these questions, the narrator reduces interpersonal distance and strengthens audience engagement with the cultural themes of the documentary.

4.1.2. Interpersonal Meaning in Modality System

This study mainly focuses on the role of modal verbs and modal adjuncts in constructing interpersonal meaning. Through statistics, a total of 76 modal verbs were found in the subtitles. All modal verbs are categorized and displayed in the table below by different types and values.

Table 2. Modal verbs with different types and values

	Modalisation				Modulation				Total	
	Probability		Usuality		Obligation		Inclination			
High					must	7			7	8.6%
Medium					should	5	will	10	25	30.9%
							would	10		
Low	may	0					can	31	49	60.5%
	might	3								
	could	15								
Total	18	22.2%	0	0%	12	14.8%	51	63%	81	100%

From the table, it can be seen that modal verbs of modulation account for 77.8%, which is higher than those of modalisation. In terms of modal value, low-value modal verbs occur most frequently, with 49 occurrences (60.5%), followed by medium-value modals (30.9%) and high-value modals (8.6%). High-value modals such as *must* mainly express obligation, while medium-value modals such as *should* indicate suggestion or moderate obligation. Low-value modals, including *may*, *might*, and *could*, are mainly used to express possibility.

Example 5: The medical combinations *must* be measured in precise doses.

Example 6: They follow the laws of nature and think human and nature *should* live in harmony.

Example 7: The motion *might* look simply and monotonous, but every strike required just the right amount of force.

Modal verbs serve as important linguistic resources for constructing interpersonal meaning by expressing degrees of certainty, obligation, and evaluation. In Example 5, *must* conveys a high degree of obligation and establishes the narrator's authoritative stance toward traditional knowledge. In Example 6, *should* expresses moderate obligation and presents cultural values in a persuasive but less forceful manner. In contrast, *might* in Example 7 reduces the certainty of the statement and creates a more objective narrative style. Through different modal choices, *China Bouquet* balances the authority of cultural narration with audience engagement, enabling viewers to develop a closer connection with the cultural meanings represented in the documentary.

Mood adjuncts are important resources for expressing modality, including probability, frequency, and the speaker's evaluation of a proposition or proposal. In Systemic Functional Grammar, mood adjuncts can be divided into two categories: mood adjuncts and comment adjuncts. The analysis shows that comment adjuncts rarely occur in the subtitles of *China Bouquet*; therefore, this section focuses mainly on mood adjuncts. Mood adjuncts can be further classified into three types: modality, temporality, and intensity.

A total of 79 mood adjuncts were identified, and the distribution of different types of mood adjuncts is presented in the following table.

Table 3. Distribution of mood Adjuncts

	Modality	Temporality	Intensity	Total
Frequency	23	18	38	79
Percent	29.1%	22.8%	48.1%	100%

The table shows that intensity adjuncts account for the largest proportion, followed by modality and temporality adjuncts. Modality adjuncts express probability and usuality, temporality adjuncts indicate time reference, and intensity adjuncts emphasize the degree or strength of an evaluation.

Example 8: Before industrial production, silk was a rare material *perhaps* because it was so precious for ordinary garments.

Example 9: 4700 years ago, silkworm rearing, silk reeling and weaving had *already* emerged.

Example 10: Relics from the Bronze Age can be found all over the world, but *only* the ancient Chinese used bronze as ritual vessels.

Mood adjuncts serve as important linguistic resources for constructing interpersonal meaning by expressing the narrator's stance, evaluation, and degree of commitment toward propositions. In Example 12, *perhaps* functions as a modality adjunct expressing possibility. By reducing the certainty of the statement, it creates a more objective and cautious narrative style, allowing viewers to engage with the information rather than receive a fixed interpretation. In Example 13, *already* functions as a temporality adjunct that emphasizes the long history of traditional practices. It connects past cultural achievements with contemporary recognition, strengthening the narrator's authority in presenting Chinese cultural heritage. In Example 14, *only* functions as an intensity adjunct by highlighting the uniqueness of ancient Chinese bronze culture. This exclusive expression reinforces the positive evaluation of traditional

practices and guides viewers to recognize their cultural significance. Through different types of mood adjuncts, *China Bouquet* adjusts the degree of certainty, evaluation, and emphasis in narration, thereby constructing a balanced relationship between the documentary voice and the audience.

4.2. Analysis of Interactive Meaning in Visual Modality

4.2.1. Interactive Meaning in Contact System

Kress and van Leeuwen classify images into two types of gaze acts: “demand act” and “offer act”. A demand act occurs when represented participants establish direct eye contact with viewers, creating a sense of interaction and involvement through the gaze. In contrast, an offer act does not involve direct gaze, positioning viewers as observers who receive information presented by the image.

For quantitative analysis, this study employed the multimodal annotation software ELAN to identify and annotate gaze acts in the 20-episode documentary series. The distribution of demand acts and offer acts is presented as follows.

Table 4. The distribution of contact

	Demand	offer	Total
Frequency	51	103	154
Percent	33.1%	66.9%	100%

According to the data above, the annotation process yielded 154 contact images in total. Among these, images representing “offer” accounted for the majority at 66.9%, while images expressing “demand” make up 33.1%.



Fig. 1 Traditional Tang Dynasty Makeup

Fig. 1, selected from the episode *Traditional Chinese Makeup*, shows the subject gazing directly at the camera, establishing direct eye contact with the audience. The woman in the image is a makeup artist specializing in ancient Chinese cosmetics. Adorned in traditional Tang Dynasty-style makeup, she looks into the lens as if inviting viewers to appreciate and acknowledge the aesthetic and cultural values behind the historical makeup. This contact encourages the audience to reflect on traditional culture, creating a sense of interaction between the participant and the viewers[10].



Fig. 2 Chinese Traditional Shadow Puppetry

Fig. 2 is from the episode *Three Generations of Shadow Puppeteers*. The image presents a master puppeteer performing traditional Chinese shadow puppetry without direct gaze toward the viewer. According to Kress and van Leeuwen's visual grammar, the absence of eye contact represents an offer act, positioning the audience as observers who receive information about the cultural practice presented. By objectively displaying the performance process and artistic features of shadow puppetry, the image encourages viewers to appreciate its cultural significance and reflect on the importance of preserving traditional heritage.

4.2.2. Interactive Meaning in Social Distance System

Social distance reflects the social relationship between the participants in the image and the viewer. This distance is visually represented through varying camera shots, such as close-ups, medium shots, and long shots. The closer the camera distance, the more intimate the social relationship. The distribution of three types of social distance will be shown in Table 5.

Table 5. The distribution of distance

	Close-up shots	Medium shots	Long shots	Total
Frequency	167	245	201	613
Percent	27.2%	40%	32.8%	100%

From the table, it can be seen that a total of 613 images are marked as representing social distance. Among them, medium shots accounts for the largest proportion at 40%, followed by long shots, while close-up shots are the least represented.



Fig. 3 Traditional Tea Leaf Harvesting

Fig. 3 is from the eighth episode *The Mile Taste of Tea*. This episode primarily explores Chinese tea culture, covering the tea-growing environment, tea production, and the art of brewing and tasting tea. Image 5.3 depicts a scene of tea picking, using a close-up shot to focus on the tea worker's hands and the tea leaves. This close-up shot shortens the social distance between the viewer and the tea worker in the image, creating a sense of intimacy. The audience can clearly observe the hand movements and the fine details of the tea leaves, which makes them immerse themselves in the moment of tea harvesting.



Fig. 4 Traditional Tea Processing

Fig. 4 is also from the eighth episode. This is a medium shot showing a tea maker processing tea, capturing the upper body of the tea maker along with the surrounding environment. The shot balances the individual and the setting, reflecting a personal distance. Although it does not convey the same level of intimacy as a close-up, it still helps to reduce the distance between the participant in the image and the viewer. It enables the audience to observe the tea maker's facial expressions and gestures while also perceiving the surrounding context. This framing reduces the psychological distance between the represented participant and the viewer, encouraging audiences to engage with the traditional tea-making practice and appreciate its cultural significance.



Fig. 5 Courtyard Tea Tasting

Fig. 5 depicts several middle-aged people drinking tea and chatting in an open courtyard. The use of a long shot represents a public distance by placing the participants farther away from the viewer. This framing reduces the sense of personal involvement and constructs an observational perspective, positioning the audience as external observers of everyday tea-drinking practices. Besides, the scene reflects the social dimension of Chinese tea culture, where tea acts as a medium for communication, reflection, and harmony. The inclusion of natural elements in the frame further reinforces the traditional Chinese ideal of unity between humans and nature. This long shot allows the audience to truly appreciate the charm of tea culture.

4.2.3. Interactive Meaning in Perspective System

Perspective refers to the choice of camera angle in an image. It reveals the power relations, attitudes, and stances between the participant in the image and the viewer, thereby constructing the interactive relationship[11]. Perspective is mainly divided into two types: horizontal and vertical. The horizontal angle can be categorized into frontal and oblique angle, while the vertical angle includes low, high, and eye-level angle. The following table present the distribution of the visual resources about perspective.

Table 6. The distribution of perspective

	Horizontal		Vertical			Total
	Frontal	Oblique	Low	High	Eye-level	
Frequency	234	212	23	94	122	685
Percent	34.2%	30.9%	3.4%	13.7%	17.8%	100%

685 images of perspective are identified in this documentary. According to the data above, horizontal angle accounts for 65.1% in total, while vertical angle occupies 34.9%. In terms of horizontal, frontal angle appears more frequently than oblique angle. As for vertical, eye-level constitutes the largest portion at 17.8%, followed by high angle. Low angle makes up the smallest share.



Fig. 6 Chan Master Practicing Calligraphy

Fig.6 is selected from the nineteenth episode *Traces of Chan*. Chan derives from the Samskrit language, meaning “contemplation in peace”. Chan is a wisdom of east and influenced Chinese culture and spirit. This frontal image shows master Xueshanjingyan, who is fully committed to the practice of Chan in Meili Snow Mountain. The frontal angle used in this image positions the viewer as if they are directly facing the master. According to visual grammar theory, a frontal perspective signifies that the viewer is involved with the represented participant. It creates a sense of inclusion and engagement, inviting the viewer into the scene as an invisible participant. By portraying the master head-on in the sacred setting of Meili Snow Mountain, the image not only

emphasizes spiritual tranquility but also encourages the viewer to engage with the cultural and philosophical depth of Chan.



Fig. 7 Chan Master Practicing in Nature

The oblique angle conveys a sense of distance between the viewer and the participant. Since Chan practice emphasizes inner silence and self-reflection, this angle highlights the pilgrim's focus and inner tranquility. By using an oblique perspective, the image encourages viewers to approach the scene with respect, fostering understanding of the spiritual depth of Chan practice rather than direct involvement.



Fig. 8 Majijishan

Fig. 8 is from the third episode *Statue Gallery on Silk Road* which introduces the culture of Buddhism. The image captures the entirety of Maijishan from a high-angle perspective, allowing viewers to clearly see the mountain's contours, the distribution of grottoes, the natural landforms, and the surrounding environment. It helps the audience understand the spatial layout of Maijishan as a site of Buddhist grotto art and its integration with the natural landscape from a macro perspective.

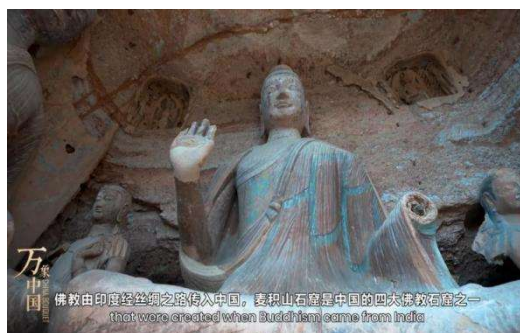


Fig. 9 Buddha Statue

Fig. 9 of the Buddha statue, taken from a low-angle perspective, positions the viewer below the figure and reinforces the statue's sense of solemnity and transcendence. In visual grammar, the vertical angle reflects the power relationship between the participant and the viewer. This upward gaze establishes an interaction based on reverence and awe, highlighting the sacred nature of the Buddha.



Fig. 10 ancient Chinese Ritual Vessel

Fig. 10 is selected from the second episode *Vessels of Ritual*. This image is taken from an eye-level angle and shows the intricate details of ancient Chinese ritual vessels. This perspective establishes an equal and intimate interactive relationship with the viewer. The eye-level perspective places the viewer and the object at the same visual height, avoiding any sense of dominance or subordination. It reduces the distance between the viewer and the historical object, fostering a stronger sense of cultural identification and emotional closeness.

4.2.4. Interactive Meaning in Modality System

Table 7. The distribution of perspective

	High modality	Low modality	Total
Frequency	254	99	353
Percent	72%	28%	100%

Modality refers to the degree of realism, credibility, or naturalness conveyed by an image - that is, how real or objective the image appears to be. The modality of an image can be divided into two degrees: high modality and low

modality. To present the authenticity and objectivity of the content, high modality images are generally used in documentary.

According to the data below, high modality images in this documentary are dominant, accounting for 72%. However, low modality images occupy 28%.



Fig. 11 Chinese Courtyard House

High modality can be illustrated by Image 5.11 which is from the *Courtyard of Harmony*. This image features highly saturated colors - crimson walls, green trees, and blue sky. The vibrant color contrast between the courtyard's green trees and the vermilion walls creates striking visual impact. Sunlight filters through the leaves, casting dappled shadows on the walls, while the natural play of light and shadow enhances the three-dimensional quality of the scene. In terms of interactive meaning, this high modality image enhances realism and persuasive power, effectively evoking audience resonance with the Chinese courtyard house as a cultural symbol while deepening the viewer's sense of immersion.



Fig. 12 Old Courtyard Houses

Fig. 12 shows the low modality. The black-and-white image shows the basic shape of old courtyard houses. This image removes color and simplifies details to just introduce the historical context and convey knowledge and information, rather than to evoke emotional resonance.

5. Conclusion

Based on Systemic Functional Grammar and Visual Grammar, this study conducts a multimodal discourse analysis of the short documentary *China Bouquet*, focusing on the construction of interpersonal and interactive meanings in verbal and visual modes.

The findings show that interpersonal meaning in verbal mode is mainly realized through mood and modality. The frequent use of declarative mood enables the documentary to present information and construct the narrator's authority in introducing traditional Chinese culture. Meanwhile, imperative and interrogative moods, although less frequently

used, contribute to audience engagement and interaction. Modal verbs and modal adjuncts further express different degrees of obligation, possibility, and evaluation, making the narration more persuasive and interactive.

Interactive meaning in visual mode is constructed through contact, social distance, perspective, and modality. The documentary mainly employs offer images, medium and long shots, and objective visual representations to provide information and maintain an observational relationship with viewers. Meanwhile, demand images and closer visual distances are occasionally used to reduce psychological distance and enhance audience involvement. Different visual choices work together to guide viewers' perception and appreciation of traditional cultural practices.

Overall, *China Bouquet* constructs a dialogue between traditional culture and contemporary audiences through the interaction of verbal and visual modes. The documentary not only introduces traditional Chinese cultural practices but also reconstructs their contemporary significance by emphasizing cultural continuity, spiritual values, and national identity. This study reveals how multimodal resources contribute to the representation and transmission of cultural heritage, providing insights into the role of documentaries in shaping cultural understanding and identification.

References

- [1] Zhu, Y. (2007). Theoretical foundations and research methods of multimodal discourse analysis. *Foreign Language Research*, 30(5), 83-85.
- [2] Halliday, M. A. K. (1978). *Language as social semiotic*. Arnold. pp. 127-135.
- [3] Halliday, M. A. K. (1994). *An introduction to functional grammar*. Arnold. pp. 365-378.
- [4] Kress, G., & van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. Routledge. pp. 123-129.
- [5] Li, Z. (2003). A social semiotic analysis of multimodal discourse. *Foreign Language Research*, 20(5), 1-7.
- [6] Hu, Z. (2007). Multimodalization in social semiotic studies. *Language Teaching and Linguistic Studies*, 29(1), 2-5.
- [7] Zhang, D. (2009). An exploration of an integrated theoretical framework for multimodal discourse analysis. *Foreign Languages in China*, 6(1), 24-30.
- [8] Zhang, D. (2018). An integrated framework for multimodal discourse analysis from the perspective of systemic functional theory. *Modern Foreign Languages*, 46(6), 731-740.
- [9] Thompson, G. (2014). *Introducing functional grammar* (3rd ed.). Routledge. pp. 18-26.
- [10] Yuan, X., & Nai, R. (2022). A study on the multimodal discourse meaning construction in international communication of "Cultural China". *Foreign Language Teaching*, 43(5), 23-29.
- [11] Zhou, J. (2022). A multimodal discourse analysis of interactive meaning in the CGTN documentary *Beautiful China* [Master's thesis]. Liaoning University. pp. 23-25.