A Contrastive Phraseological Study of Chinese and English "Appearance" Verb Patterns

Maocheng Liang* and Siwen Guo

Beihang University, Beijing, China
Email: frankliang0086@163.com (M.C.L.); ninagreen818@163.com (S.W.G.)
*Corresponding author
Manuscript received October 4, 2024; accepted December 25, 2024; published April 25, 2025.

***Abstract*—The cross-linguistic study of appearance verbs provides valuable insights into language patterns and translation practice. By using the model of the extended unit of meaning, the research aims to explore typical usage patterns of English and Chinese appearance verbs. Embedding models were used to identify potential translation equivalents of "出现" (chu1 xian4) from a parallel corpus, and to assist the analysis of semantic preference and attitudinal polarity. Results show that "出现" is translated into "appear" with the highest percentage of 48.11% among all counterparts identified. Taking this typical pair of translation equivalents as an example, we then used two monolingual corpora to describe their extended units of meaning, respectively. The findings reveal that "出现" and "appear" exhibit similar colligation patterns, however, they diverge in terms of semantic preferences and evaluative connotations. This research contributes to enhancing the understanding of Chinese and English appearance verbs, offering valuable insights into contrastive phraseology and the application of pre-trained language models in linguistics.**

***Keywords*—appear, appearance verbs, contrastive phraseology, pre-trained language models**

I. Introduction

Verbs of appearance describe "the appearance of an entity on the scene" [1]. Chinese verbs of appearance are typically "出现 (chu1 xian4)", "呈现 (cheng2 xian4)", "显现 (xian3 xian4)", etc. In English, typical verbs of appearance include "appear", "arise", "emerge", etc. These preliminary translation equivalents are to some extent interchangeable in translation practice while previous research has shown that perfect equivalents between languages are few [2]. Therefore, the current research aims to compare the patterns of Chinese and English appearance verbs from the theoretical view of corpus-based contrastive phraseology. Two appearance verbs investigated in this paper are "出现" and "appear" both of which are representative and highly frequent in corpora.

Contrastive phraseological perspective underscores that the meaning of a lexical item is established by the union of collocation, colligation, semantic preference, and semantic prosody, which is called the model of the Extended Unit of Meaning (EUM) proposed by Sinclair [3, 4]. English and Chinese appearance verbs are shown to be distinct in semantic preference and semantic prosody including verbs or phrases of "spring up", "crop up", and "涌现 (yong3 xian4)" [5]. However, it should be noted that the analysis procedure of the extended unit of meaning is, to some degree, limited in manual operation such as the identification of semantic sets or attitudinal meaning. This research attempts to employ pre-trained language models to assist the analysis of semantic preference and attitudinal polarity to include more language data and to improve the efficiency.

Based on previous introduction to English and Chinese appearance verbs and the model of the extended unit of meaning, two research questions are proposed as follows:

(1) What are typical patterning features of "出现" and "appear", respectively?
(2) To what extent are typical patterning features of "出现" and "appear" equivalent or divergent?

By answering these two questions, we might further clarify the intricated relations between Chinese and English appearance verbs through the case of "出现" and "appear". This research might shed light on the explanation of differences between highly frequent translation equivalents and offer insights for cross-linguistic contrastive phraseological study, particularly through the application of pre-trained language models.

II. Research Design

A. Corpora

In this research, two types of corpora were used, one Chinese-English parallel corpus and two monolingual corpora. The parallel corpus was used to calculate the translation corresponding rates of "出现" and "appear". Monolingual corpora were employed to explore the extended units of meaning of "出现" and "appear".

The parallel corpus was devised by Center for Chinese Linguistics Peking University (CCL). In the part of Chinese-English parallel data, CCL corpus consists of 5,434,346 Chinese-English sentence pairs with 200 million tokens [6]. Monolingual corpora for Chinese and English were the CCL Chinese part and the British National Corpus (BNC) [7].

Linguistic data were retrieved from these corpora. In the parallel corpus, 20,000 sentence pairs were randomly retrieved to describe the translation equivalents of "出现". In the Chinese corpus, 200 concordances were collected to describe the extended unit of meaning of "出现". In BNC, 100 concordances were collected to analyze the extended unit of meaning of "appear".

B. Research Instruments

The model of the extended unit of meaning consists of four elements. First, collocation refers to the syntagmatic relation between lexical items and surrounding words. Usually, collocates within the span of five tokens are taken into consideration. In this research, significant collocates are identified by log-likelihood values calculated based on the frequencies of collocates in a corpus and their co-occurrences

with node words [8]. Second, colligation shows the relation between lexical items and grammatical categories. It is usually identified by the scrutinization of concordances and significant collocates traditionally in the Key Words in Context (KWIC) format. Third, semantic preference refers to the common semantic sets of significant collocates. In this research, significant collocates are vectorized and visualized to assist the categorization of semantic preference. Fourth, semantic prosody aims to show the function of a whole extended unit and presents the communicative purpose of the unit. Previous research has also noted that semantic prosody encompasses the evaluative meaning expressed by speakers [9, 10]. However, the functional perspective on semantic prosody offers a more comprehensive understanding. In this study, we begin by automatically identifying the evaluative part of semantic prosody, followed by presenting a functional explanation. Specifically, we will use pre-trained language models to automatically identify the evaluative polarity of concordances. To elaborate, these four steps are sequenced based on their abstract levels: lexical, syntactic, semantic, and finally pragmatic. The current research will examine the extended units of meaning following this four-step procedure.

The research tools involved are pre-trained language models and python toolkits. Pre-trained language models trained on extensive corpora or fine-tuned according to specific tasks such as sentiment analysis, have revolutionized the field of computational linguistics. Inspired by the distributional hypothesis [11], these models learn and capture complex linguistic patterns, making them invaluable tools for linguistic research [12–15]. This research employed BERT-based models for vectorized representation of lexical items and the automatic identification of evaluative meaning [16, 17]. In addition, the python toolkits of spaCy and HanLP (https://github.com/hankcs/HanLP) were used to parse concordances of "出现" and "appear" [18].

C. Research Procedures

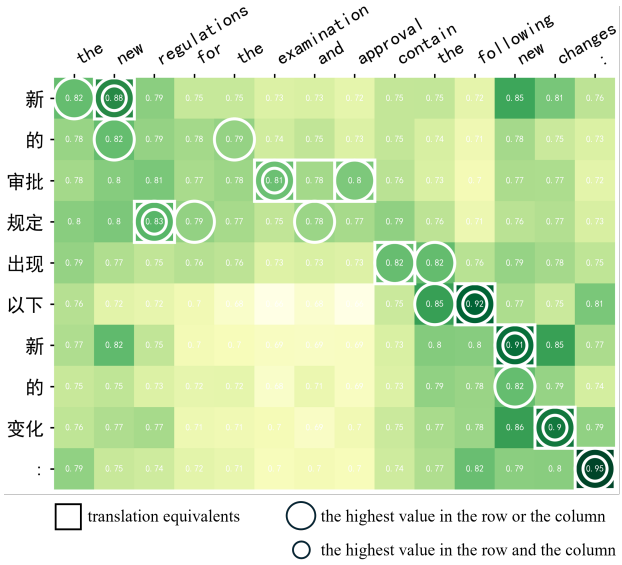


Fig. 1. Example of automatic alignment of a Chinese-English sentence pair.

Firstly, a pre-trained multi-lingual model (bert-base-multilingual-cased) was used to identify the potential English translation equivalents of "出现" as illustrated in Fig. 1. The figure represents a matrix of cosine similarity values between vector representations of each pair of Chinese and English tokens. Higher values indicate greater semantic similarity between the token pairs. In the figure, squares denote potential translation equivalents, which may include single words or multi-word correspondences. Large circles indicate the highest value in either the row or the column, while small circles denote the highest value in both the row and the column. After the identification and description of the matching frequency, "出现" and "appear" were examined by the EUM model. Due to the polysemy of "appear," the distilbert-base-uncased model was employed to categorize concordances of "appear" into distinct sense groups and to filter those concordances that pertain specifically to the meaning of "appearance".

Second, during the analysis of the extended unit of meaning, spaCy and HanLP were used to perform dependency parsing of English and Chinse sentences to retrieve collocates of "出现" and "appear" according to dependency relations. Then, the colligation was analyzed to describe the grammatical categories frequently co-occurring with appearance verbs. Next, the collocates were vectorized by pre-trained language models and these vector representations were clustered by k-means into potential semantic sets. The description of semantic preference was performed according to the cluster results. Finally, pre-trained language models for sentiment analysis were used to automatically analyze the evaluative meaning of concordances of appearance verbs. The models involved are Erlangshen-Roberta-330M-Sentiment for Chinese and distilbert-base-uncased-finetuned-sst-2-English for English.

Third, the results of the extended units of meaning were compared to explore the similarity and differences between "出现" and "appear".

III. Results and Discussions

In this section, we list preliminary English equivalents of "出现" to illustrate potential options for Chinese-English translation practice. Subsequently, we analyze and compare the extended units of meaning to demonstrate the patterning features of "出现" and "appear".

A. English Translation Equivalents of "出现"

As shown in Table I, the top ten English translation equivalents are presented with their respective percentages. The verb form "appear" ranks first. Other forms include "occur," "there be," come- phrases, and "emerge," among others. This result indicates that context plays a crucial role in translation choices. It also underscores the importance of carefully selecting the English translation of "出现" to fully convey the meaning of the source texts. This table highlights the need for further comparison to examine the equivalence or divergence between the patterns of Chinese and English appearance verbs. The English equivalent "appear" is the most frequent (48.11%) among all translation choices, therefore, it is used as an example to illustrate the workflow and explore the similarities and differences between Chinese and English appearance verb usage.

Table 1. Top ten English equivalents of "出现"

English Equivalents	Percentage	English Equivalents	Percentage
APPEAR	48.11%	PRESENT	4.45%
OCCUR	13.6%	SHOW (up)	4.27%
THERE (BE)	8.83%	SEE	3.72%
COME (up/out/into…)	5.83%	ARISE	3.13%
EMERGE	4.97%	DEVELOP	3.09%

B. The Extended Units of Meaning of "出现"

The extended units of meaning of "出现" will be analyzed from four aspects including collocation, colligation, semantic preference, and semantic prosody.

First, the collocates of "出现" concordances retrieved from the corpus within a span of ±2 tokens are listed in Table II. The decision to use a two-token span is based on preliminary observations of the concordances, which suggest that the most relevant information related to appearance verbs typically falls within this range. Additionally, the use of dependency parsing to retrieve collocates is not restricted by the span value. Therefore, the log-likelihood values are calculated to show the significance of collocates. The greater the values, the more typical the collocates are, which denotes the collocation strength between "出现" and its collocates.

Table 2. Significant collocates of "出现"

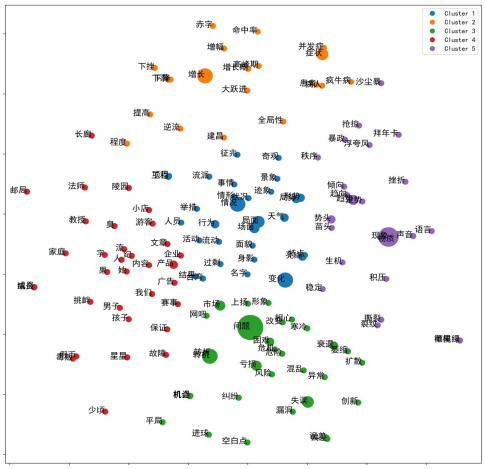
Left 2	**Left 1**	**Right 1**	**Right 2**
形象(10.58)	中 (17.48)	了 (229.89)	类似 (17.09)
甚至 (9.70)	也 (13.67)	的 (35.19)	困难 (13.38)
企业 (9.69)	刚刚 (10.58)	在 (21.67)	令 (10.58)
所有 (6.92)	上 (10.27)	上述 (10.58)	严重 (9.09)
次 (6.81)	却 (8.90)	失误 (7.77	紧张 (7.77)
已 (5.67)	能 (6.81)	新 (6.92)	有的 (7.77)

Second, three colligation patterns and examples of "出现" are summarized as below.

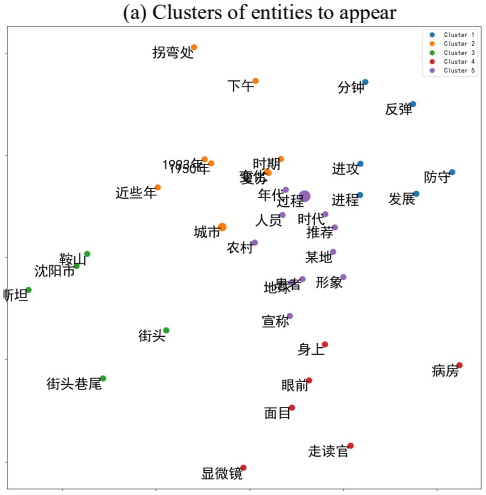
1) noun + 出现 + preposition + noun
   Example: 因为父亲喜欢看到我们双双出现在他的眼前。
   (Translation: Because father cherished having both of us right in front of him.)
2) noun + 出现 + noun
   Example: 苏瓦出现了纵火和抢掠。
   (Translation: There was arson and looting in Suva.)
3) noun + 出现 + 的 (de4, as noun modifier marker) + noun
   Example: 沈阳市出现的春夏连旱的严重程度是百年来没有过的。
   (Translation: The severity of the consecutive spring and summer droughts that occurred in Shenyang was unprecedented in a century.)

Colligation patterns show that "出现" has two frequently co-occurring grammatical categories: nouns and prepositions. Nouns can be further divided into two types: entities to appear and location nouns where entities appear. Therefore, the semantic sets of these two types of nouns are crucial to describe the usage of "出现".

Third, the collocates appearing in these two grammatical categories were retrieved by dependency parsing under the dependency relation labels of nsubj, pobj, and dobj. They denote the noun subject, the object of prepositions, and the direct object. The collocates were represented as vectors and clustered for the projection onto two-dimensional panels (as in the following Fig. 2).



(a) Clusters of entities to appear



(b) Clusters of location nouns

Fig. 2. Clusters of noun collocates of "出现".

According to the clusters of nouns collocates, entities to appear can be summarized into five categories as in Fig. 2(a). The first semantic set (orange, at the top of the panel) encompasses terms related to change, decrease, or decline. The second semantic set (red, on the left side of the panel) pertains to people, social roles, organizations, or places. The third set (blue, in the center) describes situations, conditions, and phenomena. The fourth set (purple, on the right side) represents trends or directions. The final semantic set (green, at the bottom) includes words that describe problems, difficulties, and risks. Location nouns can be categorized into four semantic sets as illustrated in Fig. 2(b). Semantic set 1 (upper middle, colored in orange) denotes time or eras. Semantic set 2 (center, colored in purple) includes words describing processes. Semantic set 3 (left, colored in green) denotes geographical locations. Semantic set 4 (bottom, colored in red) comprises words referring to parts of the body.

Next, the attitudinal meaning of concordances was predicted by the fine-tuned model for the sentiment analysis task: "Erlangshen-Roberta-330M-Sentiment". There are totally 200 Chinese concordances. Results show that 34.5% concordances are negative; 61.5% concordances are positive; 4% are neutral or not prominent in attitudinal polarity.

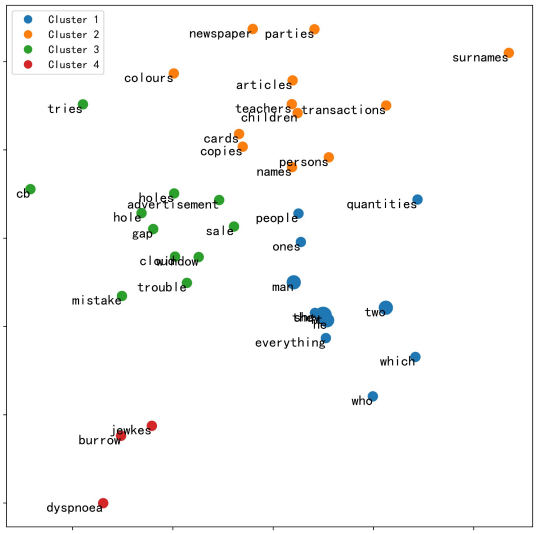
C. The Extended Units of Meaning of "Appear"

100 concordances of "appear" were randomly collected from BNC. After the filtering of senses, forty concordances were identified to express the meaning of appearance. The following analysis was performed based on these forty concordances.

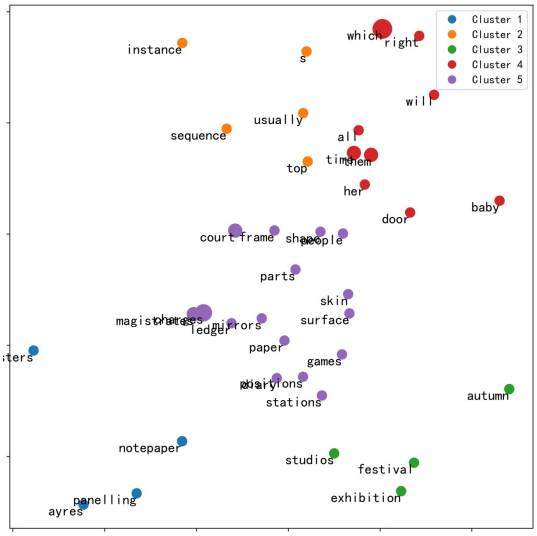
In terms of collocation and colligation, "which" (4.01), "had" (15.15), "in" (25.42), and "the" (19.50) are the most significant collocates in the position of two tokens to the left and right of "appear", respectively. By analyzing the concordances, the colligation patterns of "appear" are listed below.

1) noun/pronoun + APPEAR + preposition + noun
   Example: At the station appears cards.
2) noun + in which + noun/pronoun + APPEAR
   Example: … the alphabetical sequence in which surnames appear.

Similar to the colligation of "出现", the semantic preferences of "appear" are mainly undertaken by noun collocates. The noun collocates can then be divided into two types: entities to appear and locations where entities to appear. Then, these collocates were retrieved by dependency relation of nsubj, pobj, and relcl (relative clause modifier). These nominal collocates are projected as in Fig. 3.



(a) Clusters of entities to appear



(b) Clusters of location nouns

Fig. 3. Clusters of noun collocates of "appear".

The semantic sets of entities to appear can be categorized into four groups, as illustrated in Fig. 3(a). The first set, colored blue in the bottom right, includes pronouns and quantitative concepts. The second set, shown in orange at the top, encompasses social roles, communication, and commerce. The third set, in green on the left, consists of words related to obstacles or physical objects. The fourth set, in red at the bottom left, contains names or terms. The semantic sets of location nouns can be divided into four groups, as depicted in Fig. 3(b). The first set, in orange at the upper middle, describes the order of items and temporal concepts. The second set, in green at the bottom right, includes words related to events and seasons. The third set, in red at the upper right, primarily denotes personal entities. The fourth set, in purple at the center, refers to tangible items or specific locations.

The attitudinal meaning of "appear" concordances was identified by distilbert-base-uncased-finetuned-sst-2-English. The results of sentence polarity show that 67.5% concordances are negative while 32.5% are positive.

D. The Comparison between "出现" and "Appear"

The comparison of the extended meanings of "出现" and "appear" reveals both similarities and differences. In terms of similarities, both words tend to occur in the past tense and frequently co-occur with prepositions and nouns. They connect two types of nouns: entities that appear and the locations where these entities appear. Additionally, both verbs can function as modifiers in relative clauses. Their semantic preferences for collocates also show commonality, as both verbs describe the appearance of people, physical objects, and problems in specific places or during certain periods. The verb patterns summarized in previous sections serve a reportative function, depicting the existence of people, things, or events.

In terms of differences, the semantic preferences for the entities and locations associated with "出现" and "appear" display subtle distinctions. "出现" tends to collocate with abstract concepts, particularly those related to change, increase, or decline, whereas "appear" does not typically collocate with these words. Additionally, their semantic polarity differs: "出现" often carries a positive prosody, while "appear" predominantly has a negative prosody. According to concordance data, "出现" is frequently used to report newness, whereas "appear" tends to convey unexpectedness. This difference may be attributed to the role of evaluative meaning in communication, which helps express the speaker's stance or attitude. It is important to acknowledge that the data in this research may not be entirely sufficient. This limitation highlights an opportunity for further study and data collection to strengthen or compare the findings.

IV. Conclusion

The contrastive phraseological analysis of appearance verbs, using the case of "出现" and "appear", illuminates the patterning features of these verbs. By examining the extended units of meaning, the study reveals that while "出现" and "appear" share similar colligation patterns, their semantic preferences for nominal collocates diverge. Furthermore, the

evaluative nuances conveyed through their verb patterns differ as well. This research also highlights the practical significance of applying pre-trained language models to investigate semantic preferences and semantic polarity. Especially, embedding models are shown to be highly efficient and convenient for researchers analyzing semantic sets and attitudinal meaning. These models streamline the process of identifying and interpreting complex semantic relationships and evaluative meaning, making them invaluable tools in linguistic research. By using these models, scholars can gain valuable insights into cross-linguistic comparison and enhance our understanding of phraseology, translation studies, and foreign language pedagogy.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Maocheng Liang contributed to supervision, project administration, methodology, writing review and editing, and funding acquisition; Siwen Guo was responsible for conceptualization, research design, data collection and analysis, writing the original draft, and visualization; both authors had approved the final version.

Funding

This work was supported in part by the National Social Science Fund of China under Grant No. 19BYY082.

References

[1] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press, 1993.

[2] A. Partington, *Patterns and Meanings: Using Corpora for English Language Research and Teaching*, Amsterdam: John Benjamins, 1998.

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

[4] J. Sinclair, "The search for units of meaning," *Textus*, vol. 9, pp. 75–106, 1996.

[5] N. Wei and X. Li, "Exploring semantic preference and semantic prosody across English and Chinese: Their roles for cross-linguistic equivalence," *Corpus Linguistics and Linguistic Theory*, vol. 10, no. 1, pp. 103–138, 2014.

[6] W. Zhan, R. Guo, and Y. Chen. *CCL Corpus (Center for Chinese Linguistics Peking University)*. (2003) [Online]. Available: http://ccl.pku.edu.cn:8080/ccl_corpus.

[7] BNC Consortium. *The British National Corpus (XML Edition)*. (2007) [Online]. Available: http://www.natcorp.ox.ac.uk/XMLedition/.

[8] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[9] S. Hunston, "Semantic prosody revisited," *International Journal of Corpus Linguistics*, vol. 12, no. 2, pp. 249–268, 2007.

[10] D. Stewart, *Semantic Prosody: A Critic Evaluation*. New York: Routledge, 2010.

[11] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

[12] W. de Vries, A. van Cranenburgh, and M. Nissim, "What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models," presented at the EMNLP, Online, Nov. 16–20, 2020.

[13] P. Lin, "ChatGPT: Friend or foe (to corpus linguists)?" *Applied Corpus Linguistics*, vol. 3, no. 3, 100065, 2023.

[14] Y. Cui and M. Liang, "Automated scoring of translations with BERT models: Chinese and English language case study," *Applied Sciences*, vol. 14, no. 5, p. 1925, 2024.

[15] W. Geng and M. Liang, "From words to senses: A sense-based approach for quantitative polysemy detection across disciplines," *Journal of English for Academic Purposes*, (forthcoming), 101449, 2024.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, eds. C.J. Burges, *et al.*, vol. 26, 2013.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," presented at the International Conference on Learning Representations, Scottsdale, Arizona, May, 2013.

[18] spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). [Online]. Available: https://spacy.io