The 15th edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL)@ LREC 2026, Mallorca, Spain, 6 May 2026

Semantic Norms and Structural Relations in Chinese Semantic Composition: A Dataset and Analysis

He Zhou, Emmanuele Chersoni, Yu-Yin Hsu
Department of Language Science and Technology
The Hong Kong Polytechnic University
{he.zhou, emmanuele.chersoni, yu-yin.hsu}@polyu.edu.hk

Abstract

Word meanings can combine, and they can be represented in many different ways. In our study, we present an exploration of the relationship between linguistic structural relations and semantic norm features in the semantic composition process of Chinese phrasal structures. To achieve this goal, we constructed a dataset containing 180 compositional structures, covering six types of linguistic structural relations. Subsequently, we recruited human raters and utilized Large Language Models (LLMs) to rate the relevance of these 180 compositional structures on the dimensions of Binder-style semantic norm features.

We found that the prediction of LLMs are generally close to human ratings. Using this dataset, we further investigated whether the semantic norm feature representations of compositional structures can be approximated through composition operations using the two constituents. The results show that the semantic composition process of Chinese compositional structures is similar to that of English, being influenced by a combination of linguistic relations and semantic norm features, and exhibits non-linear compositional patterns. However, unlike English, Chinese compositional structures tend to be better approximated when their semantic norm features are represented in a lower-dimensional space. This not only results in a closer approximation to human ratings but also reveals the interactions among features. Although such a pattern may reflect the paratactic nature of Chinese, further empirical validation is needed to validate this hypothesis.

Keywords: Linguistic relations, Semantic norms, Semantic composition

1. Introduction

How the meaning of a compositional structure (either a compound or a phrase) is composed of the semantics of its constituents has been a focal point of research in lexical semantics. The composition process is complex, involving not only the meanings of the individual words, but also their structural and semantic relations. Various linguistic subfields have investigated the composition of meaning from different perspectives.

Formal linguistics is concerned with analyzing the relationships formed by the constituents within such compositional structures. Syntactically, the relations include the modifier-head structure in complex nominals and the government structure in verbal phrases. In semantics, there are diverse inventories characterized by varying levels of granularity (Nastase et al., 2022). For example, the Recoverably Deletable Predicates (RDP) schema, introduced by Levi (1978) for English noun-noun compounds, has served as a pivotal foundation for subsequent semantic relation classification efforts. These studies typically support downstream NLP tasks, such as information retrieval and machine translation. They mainly rely on the immediate constituents of complex linguistic structures and the semantic relations between them to classify or interpret these structure, but often lack analysis and utilization of deep semantic representations of the individual constituents. While word embeddings trained using neural architectures are extensively utilized for semantic representation, they suffer from a lack of interpretability, making it difficult to analyze them in terms of human-readable components of word meaning (Chersoni et al., 2021).

In contrast, in cognitive linguistics and psycholinguistics, lexical semantic analysis is predominantly grounded in componential theories, which posit that word meanings can be decomposed into sets of semantic features. Although these approaches primarily focus on individual words, the semantic attributes of constituents play a crucial role in shaping the meaning of complex nominals and phrases. In recent years, one of the most influential analyses has been the study conducted by Binder et al. (2016). Based on the functional divisions in the human brain, they developed a set of 65 experiential norms of 14 domains, and used these features to describe English nouns, adjectives, and verbs. In a similar vein, Peng et al. (2025) introduced a dataset for Chinese, BINDER-ZH 2.0, which includes Binder-style norm inventories. The multidimensional description of word meanings offers a promising approach for investigating the composition of meanings in compositional structures.

The syntactic and semantic relations of the compositional structures, as well as the semantic norms in cognitive linguistics, are both essential for understanding the semantic composition process of

compositional structures. Although these two aspects are distinct, their interaction is a question that is worthy of further investigation. Wang et al. (2025) investigated the interplay between linguistic relations and semantic features in English phrasal composition, utilizing the Binder dataset. Building on this pioneering work, the present study introduces a Chinese dataset specifically designed for investigating this question. The dataset involves 180 items in six linguistic relations, with human ratings for 68 semantic norms. By employing the same analytical methods used for English, we aim to explore whether similar tendencies can be observed in the composition of meaning structures. Additionally, given that Chinese is a paratactic language, where meaning is conveyed through the interaction of concepts rather than through morphological markers or rigid syntactic rules (Yu, 1993; Tse, 2010), we also examine whether Chinese displays distinctive paratactic characteristics that differ from English.

Based on this dataset, we investigate the relationship between linguistic structures and semantic norm features in the semantic composition process of Chinese. We also compare the performance of human raters and LLMs. The current research addresses the following questions:

1. Do LLMs possess the capability to perform semantic norm rating for compositional structures in Chinese?

2. Is Chinese semantic composition influenced solely by linguistic structural relations, solely by semantic features, or by a combination of both factors?

3. Can the semantic norm ratings of compositional structures be approximated from the two constituents through a specific composition operation? If that is the case, which operation is the best fit?

2. Related Work

A relevant line of research in classical distributional semantics consists of using property norms for grounding distributional models in perceptual data, by mapping them onto interpretable representations (Fagarasan et al., 2015; Bulat et al., 2016). The approach has been proven useful, among the other things, also for the detection of cross-domain mappings in metaphors (Bulat et al., 2017).

Although mapping word vectors onto discrete and interpretable semantic features has been a goal for many studies, identifying a set of basic features that can be efficiently used as primitives is not an easy task (Jackendoff, 1992; Zeldes, 2013).

A relatively successful proposal is the so-called *brain-based semantics* (Binder et al., 2016), in which conceptual primitives are defined in terms of *modalities of neural information processing*: human experience is organized in different domains, each one corresponding to a variable number of features for which a specialized neural processor had been identified and described in the neuroscience literature. The original set of features by Binder and colleagues has been used in several studies on word embeddings interpretability (Utsumi, 2020; Turton et al., 2020; Chersoni et al., 2021; Qiu et al., 2023; Peng et al., 2025), mostly to analyze word meanings in isolation.

One exception was the work of Wang et al. (2025), who collected human ratings of Binder features both for individual words and for phrases composed of the same words. They compared the ability of different combination rules to explain the phrase ratings based on the ratings of the individual words, finding that different rules can be more or less effective depending on feature type and on the relation between the constituents.

3. A Dataset of Chinese Compositional Structures with Semantic Norm Ratings

3.1. Compositional Structure Stimuli

To be consistent with the linguistic relation types addressed by Wang et al. (2025), we also construct six types of linguistic relations, listed in Table 1. These six relational types consist of five modification relations and one governance relation. The modification relations involve two "adjective-noun" structures and three "noun-noun" structures. Among the adjectives, we distinguish between scaler adjectives (Scaler) and intersective adjectives (Intersective). The "noun-noun" structures include three categories based on their semantic relations: For (indicating purpose), MadeOf (referring to material or manufacturing), and Has (denoting possession). In addition, the compositional structures of the governance relationship type are verb-object phrases.

To facilitate subsequent experiments exploring the relationships between single word norms and compositional structure norms, we use words in Binder-zh 2.0 as the basic units for constructing our dataset. Binder-zh 2.0 contains 535 words, including 434 nouns, 62 verbs and 39 adjectives. We first randomly select two words from the Binder-zh 2.0 dataset based on their part-of-speech to form all possible candidate structures of "adjective-noun", "noun-noun" and "verb-noun". It should be noted that the concept of adjectives in Chinese is less clearly defined compared to other languages.

Syntactic Rel	Structure	Semantic Type	Example
Modification	Adjective+Noun	Scaler	大␣船 (big ship)
		Intersective	綠␣芥末 (green mustard)
	Noun+Noun	For	家庭␣票 (family ticket)
		Madeof	藍莓␣果醬 (blueberry jam)
		Has	茄子␣餡餅 (eggplant pie)
Government	Verb+Noun	Verb	吃␣西蘭花 (eat broccoli)

Table 1: Linguistic relation types involved in this study

In the Binder-zh 2.0 dataset, the adjectival marker -的 (-de) was added to all adjectives to distinguish their adjectival usage from verbal usage. However, when we construct the "adjective+noun" structures, we consistently removed the suffix.

Subsequently, we query the frequency of these candidate structures in the BCC corpus [1] (Xun et al., 2016). If the frequency of a compositional structure candidate is zero, it indicates that the structure does not appear in natural language and will therefore be excluded. Then, from these compositional structures that have a certain frequency of occurrence in the corpus, we selected 30 for each category shown in Table 1, ultimately forming a total of 180 compositional structures.

3.2. Rating Procedure

In the Binder-zh 2.0 dataset, each word is scored across 68 dimensions within 14 domains. For our dataset, we also adopted the same domains and dimensions for scoring, which are listed in Table 2. We shuffled the 180 compositional structures and divided them into 60 groups, with each group containing three compositional structure types. For each compositional structure, we designed 68 rating questions based on the 68 norm features. Raters were asked to rate the relevance of each feature to the semantics of the compositional structure on a 0-6 Likert scale, where 0 means completely irrelevant and 6 means highly relevant. In addition, to ensure the reliability of the ratings, we incorporated six filler questions into the questionnaire, each with an unambiguous answer of either 0 or 6. Consequently, each questionnaire set contained a total of 210 questions.

Raters are required to complete the questionnaire using a computer browser. Figure 1 illustrates the interface of a sample question, which is what raters see during the experiment. For each rating question, we provide a pair of examples related to the feature mentioned in the question to help raters better understand the meaning of the feature. In addition, we record not only the ratings but also the raters' reaction time (RT), which is the duration the rater takes to complete the final rating. This measurement allows us to investigate how rater responses relate to the test items and semantic norms.

Domain	Features
Vision (15)	Vision, Bright, Dark, Color, Pattern, Large, Small, Motion, Biomotion, Fast, Slow, Shape, Complexity, Face, Body
Somatic (8)	Touch, Hot, Cold, Smooth, Rough, Light, Heavy, Pain
Audition (7)	Audition, Loud, Low, High, Sound, Music, Speech
Gustation (1)	Taste
Olfaction (1)	Smell
Motor (6)	Head, Upper Limb, Lower Limb, Practice
Spatial (7)	LandMark, Path, Scene, Near, Toward, Away, Number
Temporal (4)	Time, Duration, Long, Short
Causal (2)	Caused, Consequential
Social (4)	Social, Human, Communication, Self
Cognition (1)	Cognition
Emotion (10)	Benefit, Harm, Pleasant, Unpleasant, Happy, Sad, Angry, Disgusted, Fearful, Surprised
Drive	Drive, Needs
Attention (2)	Attention, Arousal

Table 2: 68 semantic norms (features) in 14 domains utilized in Peng et al. (2025)

We released the 60 sets of questionnaires on the Ibex platform of PennController[2] (Zehr and Schwarz, 2018). Each rater is randomly assigned one set at a time. We allow each rater to complete multiple sets of questionnaires without repetition. Consequently, each set of questionnaires is rated by 15 different raters. Each rater has been paid 30 RMB for completing one set.

The dataset is available at https://osf.io/g25ey.

[1] https://bcc.blcu.edu.cn/

[2] https://farm.pcibex.net/

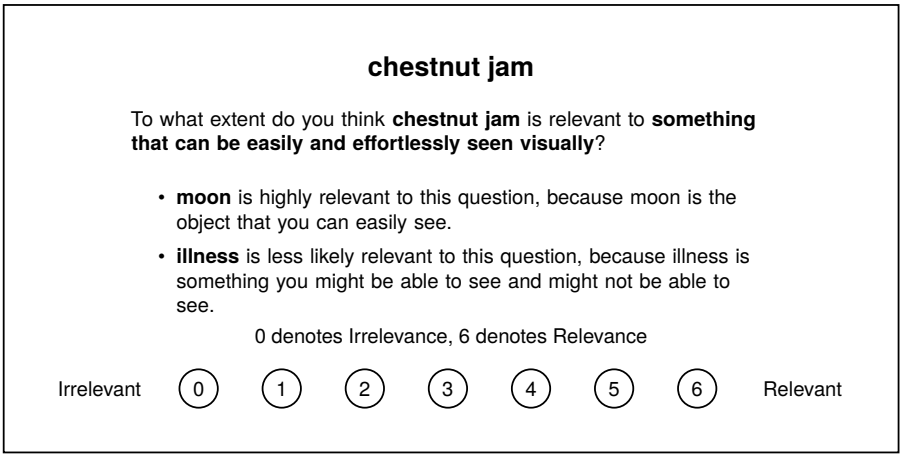

Figure 1: An example rating question. Raters are asked to score the relevance of *chestnut jam* to the *Vision* feature. Note that all questions are in Mandarin Chinese in the real experimental setting. Here, we provide the English version primarily for reviewers' convenience.

3.3. Participants and Inter-rater Agreement

A total of 100 native Mandarin speakers participated in this study, comprising 90 females and 10 males, all of whom were undergraduates or above. All the participants were recruited online, with an average age of 23.68 years old (max=37, min=19). To ensure the reliability of the ratings, we employ Fleiss's kappa κ (Fleiss, 1971) to measure the inter-rater agreement.

We calculated the inter-rater agreement based on both the test items and the features. When we calculated solely based on the features, the resulting κ score was only 0.071, indicating that the raters achieved only slight agreement. When we calculated the inter-rater agreement for each test item across all features and took the average across all items, the resulting average item-wise agreement score was 0.344, indicating that the raters achieved a fair agreement.

During the rating process, we also informed the raters that the 0-6 Likert scale can be mapped to three categories: 0-2 for "not likely to be relevant", 3-4 for "possibly relevant", and 5-6 for "highly likely to be relevant". Therefore, we further integrated the rating results by category and calculated the inter-rater agreement on these three categories. The results showed that the feature-only κ score was 0.154, while the item-wise κ score was 0.598. The lower feature-only agreement score indicates that raters had disagreements on the ratings of a feature across different test items. In contrast, raters showed higher agreement on the ratings of features for each specific test item.

The agreement scores based on both the ratings and the broader categories suggest that, although raters agree on the overall semantic relevance between test items and features, their consensus weakens when determining precise scores within the broader categories.

3.4. Correlation between Ratings and Reaction Times

To examine how items and features might influence raters' decision-making, we first fitted a linear mixed-effects model predicting log (RT) (the logarithm of reaction time) with rating as a fixed effect and a random intercept for the 180 items (N=3,066,732 observations). The point of reporting reaction time in collecting the ratings is to serve as a measure of processing difficulty and cognitive efforts. Ratings showed a small but highly significant positive effect on log(RT) (β=0.026, p<0.001), corresponding to an approximately 2.63% increase in reaction time per 1-point increase in rating (approximately 16.9% increase over a 6-point change). We then fitted a linear mixed-effects model predicting log(RT) with rating as fixed effect and random intercept by the 68 features. On the original reaction time scale this corresponds to an approximately 0.40% increase in reaction time per 1-point increase in rating (approximately 2.4% increase over a 6-point scale). This effect is also significant but small (β=0.004, p<0.001).

From the analysis across the two dimensions, we can observe that when rating relevance, the higher the relevance between a feature and the semantics of a test item, the longer time raters tend to spend on thinking and judgment, and the more cautious they are in rating the item.

3.5. LLM Raters

Moreover, we employed three Large Language Models (LLMs) as raters to score the compositional structures as human raters did, in order to examine whether LLMs possess the capability to rate semantic norms based on cognition in a way comparable to humans. The three LLM raters included DeepSeek-chat (Liu et al., 2024), Qwen3-30b (Yang et al., 2025), and GPT-5 [3]. We prompted the models to score each item strictly according to the method shown in Figure 1. In order to ensure that the models provide a consistent integer score on the scale from 0 to 6, we set the temperature parameter to 0.1.

After the three models completed their rating, we calculated their average scores, which we used as the overall performance for the LLMs.

In the following sections, we explore whether the representations of Chinese compositional structures based on semantic norms can be approximated through different composition operations using the representations of their constituents. Additionally, we will investigate whether the ratings given by LLMs for compositional structures exhibit a distribution similar to that of human ratings, as well as the relationship with the ratings of the constituents.

4. Conceptual Composition Operations

4.1. Basic Composition Operations

Given a compositional structure c composed of w_1 and w_2, $f(w_1)$ and $f(w_2)$ are the 68-dimensional feature vectors extracted from ZH-BINDER 2.0 based on human ratings. $f^h(c)$ represents the human ratings we collected fro the compositional structure c, while $f^m(c)$ is the average of the ratings given by the three large language models, and both of these vectors are also 68-dimensional. We employed the four semantic composition operation methods proposed by Wang et al. (2025) to calculate the composed semantic representation denoted as $\hat{f}(c)$. These four operation methods reflect different patterns of semantic composition:

1. Addition: $[f(w_1) + f(w_2)]/2$
2. Multiplication: $[f(w_1) \times f(w_2)]/6$
3. Word1: $f(w_1)$
4. Word2: $f(w_2)$

[3] https://cdn.openai.com/gpt-5-system-card.pdf

Then, we calculated the feature-wise composition error, that is, the difference between $\hat{f}(c)$ and both $f^h(c)$ and $f^m(c)$. The smaller the difference, the more feasible the corresponding composition operation is for the computation of compositional meaning.

4.2. Parameterized Composition Operations

Weighted operation The weighted operation measures the extent to which the compositional structure c can be reconstructed across each of the 68 dimensions (i.e., each feature) using a linear combination of weights on w_1 and w_2. The weights are optimized through training a linear regression model. The reconstruction performance is evaluated by the mean absolute error (MAE) between the predicted values and the human ratings of c, and the final result is averaged across the 68 features, denoted as wtd in Figure 3. This operation focuses on a linear combination of the two constituents on each dimension independently, but cannot capture the interactions between dimensions.

Interactive operation The interactive operation complementarily explores the interactions between features. Given the complexity of the 68-dimensional feature space, we applied Principal Component Analysis (PCA) to reduce dimensionality. Specifically, the 68 features were projected onto 10 principal components, which capture over 80% of the total variance in the data. Each word w_i are thus represented as $f'(w_i) = [f_i^1, f_i^2, ..., f_i^{10}]$.

For a compositional structure c, where the constituents w_1 and w_2 have been dimensionally reduced to $f'(w_1)$ and $f'(w_2)$, we concatenate these two vectors. Linear regression is then applied to each principle component of $f'(c)$, and the MAE is computed per component. The final result is obtained by averaging the MAEs across all 10 principle components, denoted as $interact$ in Figure 3. This operation assesses how well the combined information of w_1 and w_2 explains the compositional structure c in a lower-dimensional space.

Ridge regression with 5-fold cross validation To avoid overfitting, we also leveraged the Ridge Regression model with L2 regularization to adjust the penalty coefficient in the loss function. The regularization strength parameter α was searched within the range of 0.001 to 1000. To build a more robust model, we utilized 5-fold cross-validation, dividing the dataset into five equal parts. Each part was used as a validation set once, while the remaining parts were used for training. This process ensured a more reliable estimation of the model's generalization ability.

Subsequently, this model was employed to learn the weights for the aforementioned weighted operation and interactive operation. The MAEs obtained from the two operations were denoted as $wtd(cv)$ and $interact(cv)$ in Figure 3.

5. Results and Analysis

5.1. LLMs exhibit human-like ratings

With human ratings of the relevance between the 180 test items across the 68-dimension features in 14 domains, as well as those obtained from the LLMs, we first calculated the item-wise Euclidean distance (with min-max normalization) between the ratings of the three LLMs and the human average ratings. This yielded the average distance between each model and human ratings: 1.883 for GPT5 (std=0.247), 1.922 for DeepSeek (std=0.306), and 2.208 for Qwen3 (std=0.289). We then calculated the distance of the average ratings of the three models and the human average ratings, which is 1.465 (std=0.221). This generally reveals the extent to which the three models' ratings are close to human ratings. Moreover, the average ratings of the three LLMs is even closer to human ratings.

Using the average ratings of humans and LLMs, we visualized the domain-specific patterns for each category of compositional structures with radar charts. In addition, we incorporated human ratings of the constituents that make up the 180 compositional structures of BINDER-ZH 2.0 (Peng et al., 2025), integrating them with our results for comparison, illustrated in Figure 2.

At first glance, a clear pattern is that ratings for compositional structures, both from human raters and LLMs, tend to be lower than those for their individual constituent words. Despite this, the rating distribution of each type of compositional structure across domains remains largely consistent. This indicates that the cognitive perception of compositional structures across all domains is largely constrained by the features of their two constituents. Notably, for the HAS and MADEOF types of "noun-noun" structures, ratings in the olfaction and gustation domains are close to those of the modifiers. This implies that in such domains the features of the modifiers are more decisive in composing the semantics of the entire structure. In "adjective-noun" structures, regardless of SCALER or INTERSECTIVE, the head noun is more decisive in the causal domain. In "verb-noun" structures, the modifier is more decisive in the causal domain, while the head noun plays a more important role in the motor domain. Interestingly, we observed that LLM ratings tend to follow the overall shape of human ratings but are usually slightly lower. This is especially true for "noun-noun" and "adjective-noun" structures. However, for "verb-noun" structures, human and LLMs ratings overlap in most domains.

These results suggest that LLMs' pre-trained knowledge contain conceptual norm knowledge, driving their ratings to align with human ratings. In addition, LLMs can highlight highly relevant domains and lower less relevant domains. For example, in HAS and MADEOF categories, LLMs ratings closely match human ratings in olfaction and gustation domains. But in less relevant domains like causal, temporal, and spatial, LLMs ratings are even lower than human ratings. Thus, using LLMs can more effectively reveal semantic norm features embedded in word meanings.

5.2. Feature and Relation Interaction on Semantic Composition

In this study, we aim to investigate the relationship between the semantic norm representations of compositional structures and their constituents components. By applying the four compositional operations introduced in Section 4.1 to the two constituents that form a compositional structure, we obtained the semantic representation of it and compared it with the ratings given by humans or LLMs for the compositional structure across 68 features.

In this section, we discuss whether the semantic composition process is influenced solely by linguistic relations, solely by semantic features, or by a combination of both factors. If semantic composition is primarily impacted by linguistic relations, it mainly relies on the syntactic head of the structure. For example, in "adjective-noun" structures, the semantics is mainly determined by the head nouns, while in "verb-noun" structures, the semantics is mainly determined by the verb. In contrast, if semantic composition is solely driven by semantic features, it is primarily driven by the core features of the words, regardless which syntactic structure a structure belongs to.

To examine the influence of the two factors, we calculated the compositional error for each group according to the combination of semantic features and linguistic relations (i.e., 68$\times$6=408). The compositional error refers to the difference between the rating by humans $f^h(c)$ or LLMs $f^m(c)$ and the rating obtained through a specific operation $f(w_1\#w_2)$.

We first fitted the data calculated based on human ratings into a two-way ANOVA with considering linguistic features and semantic norm features as independent variables, with the results shown in Table 3. Regardless of which of the four compositional operations was utilized, the analysis showed a significant combined influence of both factors (p<0.001). The finding aligns with the re-

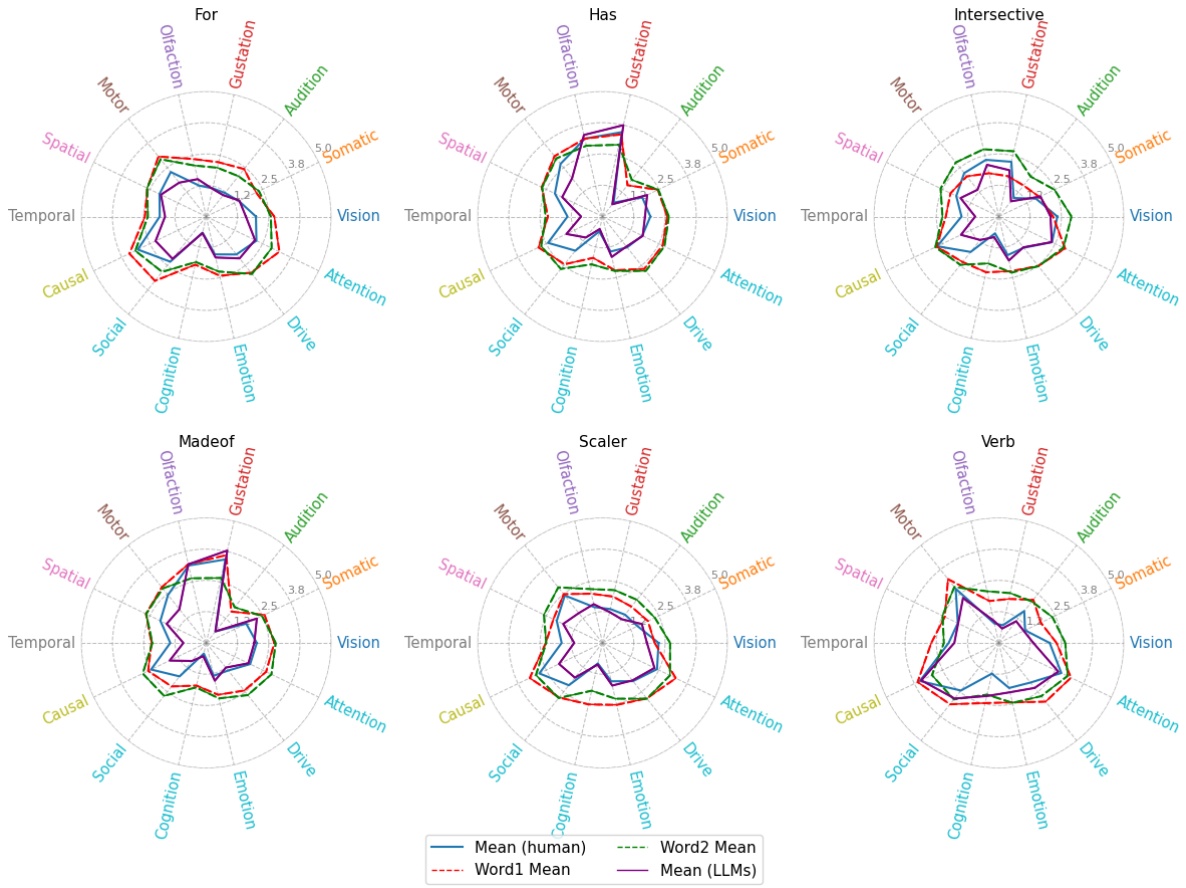

Figure 2: Average norm ratings of the compositional structures and their constituents based on 14 semantic domains

feature∼relation	Add		Multiply		Word1		Word2	
	F	p	F	p	F	p	F	p
$f^h(c) - f(w_1 \# w_2)$	2.158	2.84e-29	2.997	4.39e-64	2.562	3.04e-45	1.855	1.09e-18
$f^m(c) - f(w_1 \# w_2)$	2.978	2.88e-63	3.062	4.94e-67	3.369	3.77e-81	2.388	3.99e-38

Table 3: Two-way ANOVA results, where $\#$ denotes one of the four composition operations, F refers to the F statistic, and p refers to the p value.

sults reported by Wang et al. (2025), suggesting a cross-lingual universal that the semantic composition process is influenced not only by linguistic relations but also by conceptual semantic features.

Similarly, we performed a two-way ANOVA on the data computed from the ratings of LLMs. The results also showed the combined influence of two factors. This confirmed that LLMs exhibit human-like rating behavior when rating the conceptual semantic norms. It also indicates that LLMs not only generalize syntactic and semantic structural information but also acquire knowledge of conceptual semantic norms. This knowledge is activated during the semantic composition process. However, it is still unclear which types of features the models learned. This requires more in-depth analysis.

5.3. Which operation is a better fit for semantic composition?

Using the composition operations introduced in Section 4.1, we investigated which operation could better fit the ratings of compositional structures based on the ratings of their constituents. For each operation, we calculated the composition error between the actual ratings of the compositional structures and the fitted ratings obtained through operations on their constituents. The smaller the error, the better the operation performed in fitting the process of semantic composition. The calculation of composition errors was conducted for each type of linguistic relations. We first computed the composition error for each compositional structures within each category, and then obtained the average value. The results are illustrated in Figure 3.

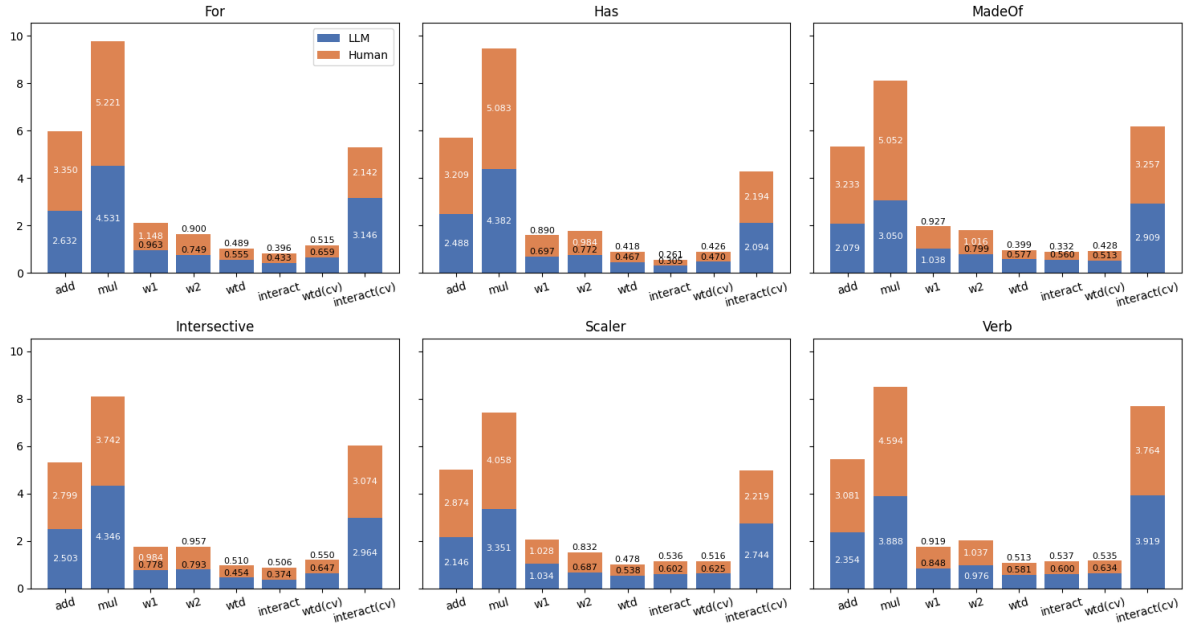

Figure 3: Composition errors obtained using different composition operations, where add, mul, w_1 and w_2 refer to the four basic operations, wtd refers to the weighted operation, $interact$ refers to the interactive operation, and $wtd(cv)$ and $interact(cv)$ indicate parameters learned from the ridge regressor using five-fold cross validation.

Overall, all linguistic relations exhibit a basically consistent pattern. Simply adding or multiplying the ratings of two words does not sufficiently fit the ratings of compositional structures, as it results in relatively higher composition errors. In contrast, relying one of the constituents (either w_1 or w_2) to fit the ratings of the entire compositional structure yields smaller composition errors, though the difference is subtle and varies depending on the type of linguistic relations. Among the parameterized operations, the interactive operation using a ridge regressor with five-fold cross validation generates relatively larger composition errors, while the composition errors obtained from the other three parameterized operations are significantly smaller.

In terms of linguistic relations, the operations with the smallest composition error are mainly two. For the three "noun-noun" compositional structures, as well as the intersective "adjective-noun" structures, the interactive operation achieved the smallest composition error. In contrast, for the scaler "adjective-noun" and "verb-noun" structures, the weighted operation achieved the smallest composition errors. Our results differ from Wang et al. (2025): in their experiments, the weighted operation yielded the smallest composition errors in most cases, whereas in our study the interactive operation outperformed the others in most cases. As mentioned in Section 4.1, in the weighted operation, the 68 dimensions are independent of each other; while the interactive operation reduced the 68 dimensions to 10 principal components, focusing more on the interactions among features. Projecting high-dimensional semantic features into a lower-dimensional space yielded improved approximation of semantic composition, suggesting that semantic composition in Chinese may place greater emphasis on interactions among semantic features. This observation could potentially be explained by the paratactic nature of Chinese, where meaning is constructed primarily through the semantic and logical interaction between concepts rather than through explicit morphological markers or rigid syntactic rules (Yu, 1993; Tse, 2010). However, definitive conclusions cannot be drawn from the current study, and further investigation is required to valid this hypothesis.

Figure 3 stacks the composition errors based on human ratings and those based on LLMs. Taking a closer comparison of the operations derived from the two types of ratings, we observed that across all linguistic structure categories, for the four basic unparameterized operations, the composition errors computed from LLMs' ratings were consistently smaller than those from human ratings, while the four parameterized operations demonstrated a conversed picture. This reveals a potential issue in using LLMs for semantic norm rating. Although our experiments have shown that the ratings given by LLMs are largely human-like, the models' ratings may still contain unreasonable or irregular ratings at times. Therefore, when learning weights

based on the LLMs' ratings, these irregularities can introduce noise, leading to larger errors in the estimation of the ratings. We leave to future research to examine which models exhibit irregularities in rating which features.

6. Conclusion

In this study, we present a novel dataset of 180 Chinese phrasal compositional structures, covering six linguistic relations. The compositional structures are rated and evaluated across 68 Binder-style semantic norm features by human raters and three LLMs. The rating results demonstrate that LLMs exhibit human-like semantic ratings, particularly in domain-specific patterns. A further investigation of the factors affecting semantic composition reveals that the process is influenced by both linguistic structural relations and semantic norm features, aligning with previous studies in English. Moreover, we show that simple compositional operations, such as addition or multiplication, are insufficient to fully capture the semantic norms of compositional structures. Instead, parameterized operations, especially those incorporating dimensionality reduction and feature interaction, yield more accurate approximations, indicating that the semantic composition of compounds may not depend on simple linear combination along but rather involves non-linear operations. In conclusion, differently from English, Chinese compositional semantics seems to benefit from lower-dimensional representations, reflecting the nuanced interplay between constituent features.

Limitations

This study may have the following two limitations. First, our dataset is constructed based on the 535 words in ZH-BINDER 2.0, with only 30 compositional structures for each type of the six linguistic structures. While this provides a solid foundation for our initial investigation, the relatively constrained lexicon and linguistic structure categories may limit the generalizability of our findings to the full understanding of Chinese compositional structures. We will expand the dataset to include a broader vocabulary and linguistic structural types. Moreover, due to the strict conditions set during the data construction process, the data is limited in size from an NLP perspective. However, these data can help us gain a deeper understanding of semantic composition in Chinese. Second, due to the space limitations, our analysis of the dataset mainly focuses on the analysis based on 14 domains (as shown in Figure 2) and the analysis based on 6 types of linguistic structures (as shown in Figure 3), without further in-depth investigation based on the 68 semantic norm features. A more detailed feature-by-feature analysis will be conducted to uncover a more fine-grained semantic composition process.

Ethics Statement

This study collected semantic feature norm ratings from human participants and received approval from the ethical committee of the university. Along the ratings, we also collected demographic data from the raters, including gender, year of birth, languages, and so on. By incorporating demographic information from raters, we aim to enhance our understanding how different groups perceive the semantic norms. All raters provided informed consent for the participation and were paid 30 RMB for completing a questionnaire set. Upon acceptance of this paper, this dataset will be made available, which will not contain any of the demographic information of the raters.

Acknowledgements

We acknowledge The Hong Kong Polytechnic University for financial support through the Postdoc Matching Fund (1-W28X) to He Zhou, the Start-up Fund for New Recruits, Faculty of Humanities (1-BE8G) to Emmanuele Chersoni, and the Large Equipment Fund (LEF2223-008) to Yu-Yin Hsu.

References

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a Brain-Based Componential Semantic Representation. *Cognitive Neuropsychology*, 33(3-4):130–174.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling Metaphor with Attribute-Based Semantics. In *Proceedings of EACL*.

Luana Bulat, Douwe Kiela, and Stephen Christopher Clark. 2016. Vision and Feature Norms: Improving Automatic Feature Norm Learning Through Cross-Modal Maps. In *Proceedings of NAACL-HLT*.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding Word Embeddings with Brain-based Semantic Features. *Computational Linguistics*, 47(3):663–698.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From Distributional Semantics to Feature Norms: Grounding Semantic Models in Human Perceptual Data. In *Proceedings of IWCS*.

Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5):378.

Ray Jackendoff. 1992. *Semantic Structures*, volume 18. MIT Press.

Judith N Levi. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séagdha. 2022. *Semantic Relations between Nominals*. Springer Nature.

Bo Peng, Yu-yin Hsu, Emmanuele Chersoni, Le Qiu, and Chu-Ren Huang. 2025. Multilingual Prediction of Semantic Norms with Language Models: A Study on English and Chinese. *Language Resources and Evaluation*, pages 1–27.

Le Qiu, Yu-Yin Hsu, and Emmanuele Chersoni. 2023. Collecting and Predicting Neurocognitive Norms for Mandarin Chinese. In *Proceedings of IWCS*.

Yiu-Kay Tse. 2010. Parataxis and Hypotaxis in the Chinese Language. *International Journal of Arts and Sciences*, 3(16):351–359.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating Binder Style Word Embeddings to New Words. In *Proceedings of the Workshop on Linguistic and Neurocognitive Resources (LiNCR)*.

Akira Utsumi. 2020. Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, 44(6):e12844.

Shaonan Wang, Songhee Kim, Jeffrey R Binder, and Liina Pylkkänen. 2025. Unlocking the Complexity of Phrasal Composition: An Interplay between Semantic Features and Linguistic Relations. *Cognition*, 254:105986.

Endong Xun, Gaoqi Rao, Xiaoyue Xiao, and Jiaojiao Zang. 2016. The Construction of the BCC Corpus in the Age of Big Data. *Corpus Linguistics*, 3(1):93–109. In Chinese.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Ning Yu. 1993. Chinese as a Paratactic Language. *Journal of Second Language Acquisition and Teaching*, 1:1–15.

Jeremy Zehr and Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX).

Amir Zeldes. 2013. Productive Argument Selection: Is Lexical Semantics Enough? *Corpus Linguistics and Linguistic Theory*, 9(2):263–291.

7. Language Resource References

Peng et al. 2025. *Binder-zh 2.0*. PID https://anonymous.4open.science/r/norms_correlation-F185/README.md.