Assessing Minimal Pairs of Chinese Verb-Resultative Complement Constructions: Insights from Language Models

Xinyao Huang
ECNU Shanghai &
HHU Düsseldorf
huangxinyao_23
@stu.ecnu.edu.cn

Yue Pan
ECNU Shanghai
51270400014
@stu.ecnu.edu.cn

Stefan Hartmann
HHU Düsseldorf
hartmast@hhu.de

Yanning Yang
ECNU Shanghai
ynyang@english
.ecnu.edu.cn

Abstract

Chinese verb-resultative complement construction (VRCC), constitute a distinctive syntactic-semantic pattern in Chinese that integrates agent-patient dynamics with real-world state changes; yet widely used benchmarks such as CLiMP and ZhoBLiMP provide few minimal-pair probes tailored to these constructions. We introduce ZhVrcMP, a 1,204 pair dataset spanning two paradigms: resultative complement presence versus absence, and verb–complement order. The examples are drawn from *Modern Chinese* and are annotated for linguistic validity. Using mean log probability scoring, we evaluate Zh-Pythia models (14M-1.4B) and Mistral-7B-Instruct-v0.3. Larger Zh-Pythia models perform strongly, especially on the order paradigm, reaching 89.87% accuracy. Mistral-7B-Instruct-v0.3 shows lower perplexity yet overall weaker accuracy, underscoring the remaining difficulty of modeling constructional semantics in Chinese.

1 Introduction

Chinese verb-resultative complement constructions (VRCC) stand out as one of the distinctive and challenging features in syntax and semantics. They feature a complex interplay of elements like agent-patient dynamics, resultative states, and real-world state changes. Any syntactic or semantic mismatch in these constructions can sharply reduce sentence acceptability (often marked with *), as it diminishes the likelihood of such events occurring in reality. For illustration, example (a) shows a clear relation between agent and patient. The agent "I (我)" performs the action "broke (打)" on the patient "vase", which causes the state change "up (碎)" and yields a complete resultative event. The physical properties of the patient constrain the result: a vase can plausibly become "broke up (碎)" but not "into two pieces (断)", so (b) is well formed in syntax but infelicitous in meaning. VRCC also respect event order, causing action must come first and the result must follow, so (c) violates this sequence and is semantically unacceptable. Capturing VRCC requires balancing the individual semantics of components with their overall integration, which poses significant hurdles for grammatical annotation, semantic parsing, and broader NLP applications.

a. 我打碎了花瓶。
wǒ dǎ suì le huāpíng
I broke up the vase.

b. * 我打断了花瓶。
wǒ dǎ duàn le huāpíng
I broke the vase into two pieces.

c. * 我碎打了花瓶。
wǒ suì dǎ le huāpíng
I up broke the vase.

Beyond computational capacity and data scale, the capability of language models to handle complex grammatical structures significantly

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 144–150

impacts their performance in 'understanding' and generating natural language. The minimal pair (MP) method, a foundational linguistic paradigm for testing human language aptitude, has been widely adopted to evaluate language models (LMs) (Xiang et al., 2021; Song et al., 2022; Someya and Oseki, 2023; Warstadt et al., 2023; Capone et al., 2024; Liu et al., 2024). This method generates sentence pairs differing in a single grammatical feature (e.g., word order, morphology, syntax) to assess model comprehension of specific grammatical phenomena. An effective LM should assign higher acceptability probabilities to grammatically and semantically valid sentences in MPs.

With advantages in rigorous variable control, scalable automated design, cross-lingual applicability, and prompt-interference immunity, MPs-based benchmarks exist for multiple languages, including Chinese-specific CLiMP (Xiang et al., 2021), SLING (Song et al., 2022), and ZhoBLiMP (Liu et al., 2024).

Although these datasets excel in broad syntactic paradigm coverage, they lack in-depth exploration of linguistic phenomena through a constructional lens as well as semantic minimal pair design. Poor differentiation between formal and functional competencies leaves model comprehension of semantic relations unaddressed (Mahowald et al., 2024), weakening evaluation interpretability.

CLiMP covers five VRCC paradigms (51000 pairs) but relies solely on complement alteration, with non-random sampling and limited variation compromising validity. In contrast, SLING addresses 38 linguistic phenomena, but omits explicit VRCC. ZhoBLiMP includes partial VRCC in its 14 verb phrase paradigms (14×300 pairs) but lacks a dedicated design and has severely restricted lexis.

To fill this gap, we present ZhVrcMP, a MP dataset for Chinese VRCC, comprising two paradigms and 1,204 total MPs. Words in ZhVrcMP are linguistically selected from the *Modern Chinese*, with partial lexicon adaptation from ZhoBLiMP (Section 3). We tested two types of language models on ZhoBLiMP, our benchmark for assessing how well these models handle Chinese grammar through pairs of sentences with only one grammatical or semantical difference. The first is Zh-Pythia, a set of models adapted for Chinese based on the Pythia framework (Liu et al., 2024), with sizes ranging from 14 million to 1.4 billion parameters (a measure of each model's complexity and capacity). The second is Mistral-7B-Instruct-v0.3, a leading model that has been specially adjusted to follow user instructions effectively; it uses a Transformer design with a 32,768 vocabulary. (Section 4).

Results are detailed in Section 5, along with the part of ZhVrcMP, and model evaluation scripts.

2 Related Work

2.1 Verb-Resultative Complement Construction in Chinese

VRCC, a major subtype of Chinese verb–complement patterns, has the form V + RC where the complement encodes the resultant state caused by the event. This tight coupling of lexical semantics and causation makes VRCC an informative minimal-pair testbed for evaluating LMs' syntactic–semantic processing(Marvin and Linzen, 2018; Kuribayashi et al., 2024).

Construction grammar (CxG)'s form-meaning pairing principle guides the design of minimal pairs (MPs) to probe language models' (LMs) capabilities in semantic role labeling and constructional structure recognition (Weissweiler et al., 2023). As reviewed in recent computational syntheses (Doumen et al., 2024), these principles are operated through unsupervised learning methods (e.g., word embedding clustering) for automatic VRCC identification and

association-based algorithms (e.g., the ΔP metric) for selecting representative MPs (Dunn, 2022). By association-based methods we mean corpus measures that quantify the strength of pairing between verbs and resultative complements, such as ΔP, PMI, and log-likelihood (Stefanowitsch and Gries, 2003; Dunn, 2024). In this paper we construct ZhVrcMP via controlled grammatical manipulations and manual validation rather than corpus-based association scores, although the two approaches are complementary.

Cognitive studies show that VRCC processing involves real-time structure-meaning mapping, with type-shifting complements prolonging model inference (Xue et al., 2021). Thus, VRCC's markedness (e.g., grammaticality constraints) and semantic subtypes (e.g., resultative/stative) enable controlled MPs to assess LMs' grammaticality judgment and low-frequency construction learning (Someya and Oseki, 2023; Warstadt et al., 2023).

2.2 Construction Grammar in Evaluation of Language Capabilities of LMs

CxG grounds LM research through its form-meaning pairing principle, which in turn allows for addressing traditional models' failure to capture implicit constructional information in Chinese VRCC (Zhan, 2017). For instance, Weissweiler (2023) demonstrates that Transformer self-attention aligns with CxG's gestalt cognition, thereby enabling more effective encoding of constructional knowledge and ultimately improving recognition of Chinese VRCC.

These CxG-inspired LM approaches (e.g., Tseng (2022)'s 17.6% accuracy boost in structured tasks) enhance low-frequency construction learning. Despite their importance for LM assessment, the acquisition of constructional knowledge still lacks standardized benchmarks. Existing models focus on form-meaning pattern extraction (Dunn, 2023) and verb argument structure learning (Dominey, 2005; Alishahi and Stevenson, 2008), which form the basis for MP design in VRCC evaluation.

3 Data

ZhVrcMP includes two paradigms: resultative complement presence/absence (Para 1) and verb-complement order inversion (Para 2) with 602 MPs each and 1,204 in total. Curated from the authoritative grammar book *Modern Chinese* (Huang and Li, 2012), which provides comprehensive explanations and examples of Chinese syntax, it adapts the lexicon from ZhoBLiMP. Linguists annotated noun/verb/complement features, generated matching lists (3.1). MPs were automatically generated using an algorithm and manually validated afterwards (3.2).

3.1 Minimal Pairs Generation

3.1.1 Data Sources

As mentioned above, ZhVrcMP sources two main datasets:

1. examples from *Modern Chinese* (pp.78–83);

2. the lexicon of ZhoBLiMP.

For *Modern Chinese* sentences, we parsed components into nouns, verbs, and resultative complements, systematically identifying all verb-complement pairings to ensure dataset richness in capturing VRCC syntactic-semantic relationships.

3.1.2 Vocabulary

ZhVrcMP's noun lexicon has 342 entries with POS, subcategory, gender, animacy, and number annotations. The verb set includes 53 verbs annotated for compatible resultative complements, subject/object subtypes, transitivity, and animacy constraints, matched with 66 unique complements. Using a "subject + verb + complement + (aspect

marker 了) + object" structure, a Python script generated 24,000+ MPs. To minimize unnecessary variation in subjects (which would increase generation workload without adding evaluative value), , "张三 (Zhang San)" was fixed as the sole subject for consistent evaluation.

3.2 Manual Validation

Two annotators with a background in linguistics conducted a double-blind verification of lexical annotations and the automatically generated MPs, yielding an initial inter-annotator agreement rate of 62.6%. After revising 99 pairs (adding indirect objects, aspect markers, etc.) and re-verifying, 602 sentence pairs were selected for each category with a 98% agreement rate (Table 1). We binary-labeled all sentences as GOOD if they were grammatically and semantically well-formed, or as BAD if they differed minimally from the GOOD sentences in one aspect, making them grammatically or semantically invalid. Chi-square tests confirmed statistical equivalence of auxiliary features across label groups (all $p > 0.05$), demonstrating no significant association between exogenous feature distributions and VRCC to isolate the core test variable (Table 2).

4 Models and Methods

We evaluated Zh-Pythia (14M-1.4B) and Mistral-7B-Instruct-v0.3 (4.1) using mean log probability to compare GOOD/BAD sentence probabilities (4.2).

4.1 Models

We evaluated two models: Zh-Pythia (from the ZhoBLiMP study) and Mistral-7B-Instruct-v0.3. Zh-Pythia consists of 20 Chinese-focused Transformer models, trained from scratch on 3B tokens with GPT-NeoX architecture and a Chinese tokenizer to analyze scaling effects on Chinese linguistic phenomena in ZhoBLiMP. Mistral is a commercial 7B-parameter English model optimized for instruction tasks, pre-trained without Chinese adaptation (Table 3).

The models were selected for their contrasting attributes: Zh-Pythia is a Chinese-specific, scalable design evaluated on ZhoBLiMP, while Mistral features a fixed-scale, English-oriented architecture.

4.2 Evaluation Methods

To evaluate the model, we devised a score based on the mean log probability P_{ML}.

$$P_{ML} = \frac{\log P_m(\gamma)}{n_\gamma} \tag{1}$$

In (1), P_{ML} is the mean log probability, $logP_m(\gamma)$ is the mean log probability of model m for γ, n_γ is the sentence count in γ.

Based on the mean log probability obtained above, for each pair set p, we calculated the evaluation score via (2).

$$S(p) = \frac{1}{|p|} \sum_{g,u \in p} \mathbf{1}_{[0,+\infty)}\left(\log \frac{P_{ML}(g)}{P_{ML}(u)}\right) \tag{2}$$

In (2), $S(p)$ is the score for pair set p, $|p|$ is the pair count, and g, u represent the GOOD and BAD sentences, respectively. An indicator function counts the number of valid ratios where $P_{\mathrm{ML}}(g)/P_{\mathrm{ML}}(u) > 1$ (i.e., the model assigns higher probability to the GOOD sentence), and this count is then averaged across pairs to measure the model's ability to capture linguistic capabilities.

Finally, we computed perplexity via mean log probability to quantify model prediction uncertainty, where lower values indicate better data fit.

$$P_{PL} = e^{-P_{ML}} \tag{3}$$

P_{ML} is the mean log probability obtained above. $e^{-P_{ML}}$ converts the log probability to perplexity, a standard metric for assessing LM performance.

	Para 1	Para 2
Number	602	602
GOOD	张三摔破额头。	张三搞错观点。
	Zhāng Sān shuāi pò é tóu	Zhāng Sān gǎo cuò guāndiǎn
	Zhang San fell and broke his forehead.	Zhang San got the wrong point.
BAD	* 张三摔额头。	* 张三错搞观点。
	Zhāng Sān shuāi é tóu	Zhāng Sān cuò gǎo guāndiǎn
	Zhang San fell his forehead.	Zhang San wrong got the point.

Table 1: ZhVrcMP paradigms and example minimal pairs. Para 1 tests resultative complement presence versus absence; Para 2 tests verb–complement order. Each paradigm contains 602 minimal pairs (1,204 total). GOOD is grammatical and semantically plausible; BAD differs only in the targeted constructional feature and is unacceptable (marked with *).

Feature	χ^2	p-value	Conclusion
AM	0.013	0.910	Indep
CL	0.001	0.972	Indep
IO	0.066	0.797	Indep
PASS	0.066	0.797	Indep
MOD	0.639	0.424	Indep

Table 2: Chi-Square Test for Auxiliary features in ZhVrcMP. AM: Aspect Marker, CL: Classifer, IO: Indirect Object, PASS: Passive Voice, MOD: Modifier. Independence is abbreviated as Indep.

Models	Zh-Pythia	Mist
Parameters	14M, 70M, 160M, 410M, 1.4B	7B

Table 3: Evaluation LMs. Mist indicates Mistral-7B-Instruct-v0.3.

5 Results

Results for Zh-Pythia and Mistral-7B-Instruct-v0.3 in the two paradigms are presented in Tables 4, 5, 6. In general, the correct evaluation counts of the models cluster between 400-570 pairs, with scores ranging from 60 to 95, showing a relatively wide range. The perplexity sheds light on the uncertainty of the models in VRCC processing. Although overall scores are high, indicating a notable uncertainty in distinguishing VRCC, substantial differences in perplexity between GOOD and BAD sentences allow effective differentiation.

5.1 Paradigm Results

In Table 4, we analyze performance as a function of model size (number of parameters). In Para 1, Zh-Pythia shows a positive parameter-scaling trend: the number of correctly judged pairs and the mean score both rise with increasing model size. In Para 2, we observe no clear parameter-scaling effect; performance fluctuates slightly across sizes. Overall, Para 2 outperforms Para 1, suggesting that verb–complement order inversion is easier for these models than resultative-complement presence/absence. Performance peaks at 160M (574 correct pairs; mean score 95.19) and declines for larger models, indicating non-monotonic (inverse) scaling beyond 160M. We hypothesize this may relate to training budget or regularization rather than an inherent property of Transformers; order sensitivity can often be captured by local attention patterns, whereas presence/absence relies more on lexical compatibility.

	Zh-Pythia					Mist	Human
Parameter	14M	70M	160M	410M	1.4B	7B	
Para 1	414	487	500	499	517	382	590
Para 2	522	558	574	560	566	504	590
Overall	468	522.5	537	529.5	541.5	443	590

Table 4: Correct Evaluation Numbers of all LMs and human on ZhVrcMP. Human indicates linguistics experts annotation results.

	Zh-Pythia					Mist	Human
	14M	70M	160M	410M	1.4B	7B	
Para 1	68.77	80.90	83.06	82.89	85.88	63.46	98.00
Para 2	86.57	92.54	95.19	92.87	93.86	83.58	98.00
Overall	77.67	86.72	89.13	87.88	89.87	73.52	98.00

Table 5: Percentage Score of all LMs and human on ZhVrcMP

Parameter	**Para 1**		**Para 2**		**Overall**	
	GOOD	BAD	GOOD	BAD	GOOD	BAD
14M	1931.78	2866.34	1928.92	4583.57	1930.35	3724.96
70M	1512.07	2690.21	1511.60	4389.68	1511.84	3539.95
160M	1277.36	2080.77	1278.03	4246.92	1277.70	3163.85
410M	1468.29	2759.91	1469.66	4339.49	1468.98	3549.7
1.4B	2115.22	3902.21	2115.75	6672.02	2115.49	5287.12
7B*	520.36	760.97	524.48	753.70	522.42	757.34

Table 6: Perplexity of all LMs on ZhVrcMP. 7B indicates Mistral-7B-Instruct-v0.3.

5.2 Model Results

Zh-Pythia demonstrates superior performance with smaller parameter sizes compared to Mistral-7B-Instruct-v0.3 (Table 5). Despite Mistral's larger parameters, its correct evaluation counts and scores trail behind Zh-Pythia, particularly in Para 1. However, Mistral exhibits significantly lower perplexity values than Zh-Pythia across both paradigms (Table 6). This duality suggests that while Zh-Pythia's parameter——scaling efficiency aligns more closely with VRCC, Mistral's larger model capacity enhances confidence in distinguishing grammatical and ungrammatical sentences, as reflected by its lower perplexity. The contrasting trends in accuracy and perplexity underscore the interplay between training data relevance and model architectural inductive biases, where Mistral's Transformer design excels in capturing sequence dependencies, thereby reducing perplexity.

6 Conclusion

This paper has introduced ZhVrcMP, a Chinese verb-resultative construction dataset, to assess LMs' semantic-grammatical comprehension. Comprising 1,204 minimal pairs across two paradigms, we have evaluated Zh-Pythia (14M–

1.4B) and Mistral-7B-Instruct-v0.3. The models excel more at verb-complement order than resultative complement presence/absence. Zh-Pythia shows parameter-performance correlation in the latter, peaking at 160M for the former. Mistral lags behind Zh-Pythia, especially in resultative complement tasks. Both models trail human performance, highlighting that construction-semantic processing still has room to improve.

Acknowledgments

This research was funded in part by East China Normal University, 2025 Program for Outstanding Doctoral Student Academic Innovation (Grant No. YBNLTS2025-049).

References

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive science*, 32(5):789–834.

Luca Capone, Alice Suozzi, Gianluca E Lebani, Alessandro Lenci, et al. 2024. BaBIEs: A benchmark for the linguistic evaluation of italian baby language models. In *Ceur Workshop Proceedings*. CEUR-WS.

Peter Ford Dominey. 2005. From sensorimotor sequence to grammatical construction: Evidence from simulation and neurophysiology. *Adaptive Behavior*, 13(4):347–361.

Jonas Doumen, Veronica Juliana Schmalz, Katrien Beuls, and Paul Van Eecke. 2024. The computational learning of construction grammars: State of the art and prospective roadmap. *Constructions and Frames*.

Jonathan Dunn. 2022. Cognitive linguistics meets computational linguistics: Construction grammar, dialectology, and linguistic diversity. *Data Analytics in Cognitive Linguistics: Methods and Insights*, 41:273.

Jonathan Dunn. 2023. Exploring the constructicon: linguistic analysis of a computational cxg. *arXiv preprint arXiv:2301.12642*.

Jonathan Dunn. 2024. *Computational construction grammar: A usage-based approach*. Cambridge University Press.

Borong Huang and Wei Li. 2012. *Modern Chinese (Volumn 2)*. BEIJING BOOK CO. INC.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. *arXiv preprint arXiv:2402.12363*.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024. ZhoBLiMP: A Systematic Assessment of Language Models with Linguistic Minimal Pairs in Chinese.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. Sling: Sino linguistic evaluation of large language models.

Anatol Stefanowitsch and Stefan Th Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. Cxlm: A construction and context-aware language model. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6361–6369.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023. BLiMP: The Benchmark of Linguistic Minimal Pairs for English.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2023. Explaining pretrained language models' understanding of linguistic structures using construction grammar. *Frontiers in Artificial Intelligence*, 6:1225791.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A Benchmark for Chinese Language Model Evaluation.

Wenting Xue, Meichun Liu, and Stephen Politzer-Ahles. 2021. Processing of complement coercion with aspectual verbs in mandarin chinese: Evidence from a self-paced reading study. *Frontiers in psychology*, 12:643571.

Weidong Zhan. 2017. On theoretical issues in building a knowledge database of chinese constructions. *J. Chinese Inform. Process*, 31:230–238.