Composition as nonlinear combination in semantic space:

Exploring the effect of compositionality on Chinese compound recognition

Tianqi Wang[1], Xu Xu[2]

[1] The University of Hong Kong [2] Shanghai Jiao Tong University

Introduction

- Most Chinese words are formed by the combination of characters (e.g., 冰箱 refrigerator = 冰 ice + 箱 box)
- Characters are highly salient perceptual units, making morphological segmentation executed without effort
- The role played by constituents in compound processing has been studied via **semantic transparency** (ST; e.g., *bedroom* vs. *hogwash*), which produced inconsistent results
- Psycholinguists started to reconceptualize ST from the ***compositional perspective***, i.e., the predictability of the compound meaning given the combination of the constituents' meaning
- It is unclear how this combinatorial process modulates compound processing 🤔❓

A quick view of this work

- We built a computational model to learn the compounding rules
- Using the model, we generated the compositional meaning representation and characterized its two attributes
- We examined how these attributes affected Chinese compound processing

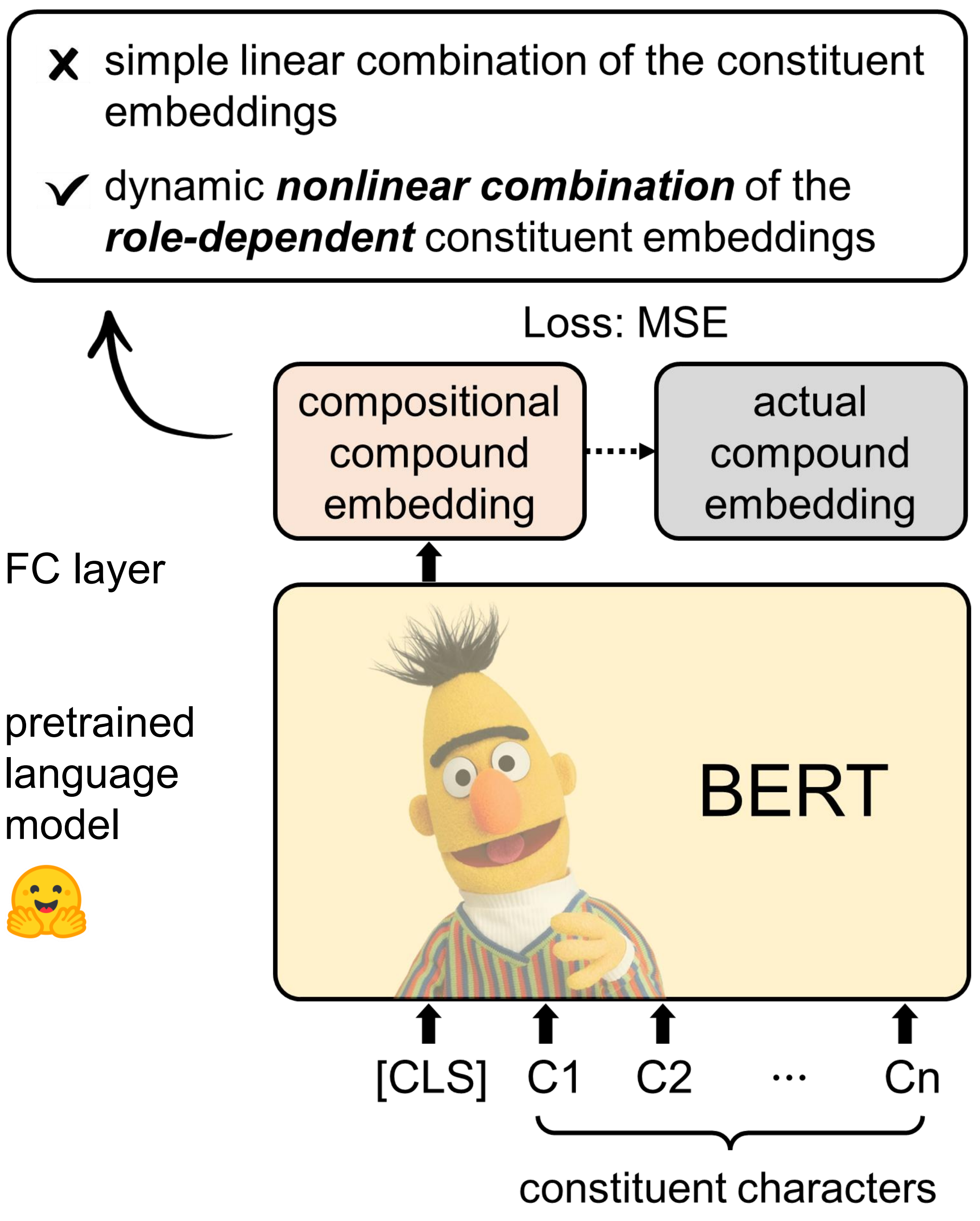
Method

Computational model

A transformer-based deep neural network is trained to optimally predict the actual compound embedding so as to acquire the ***compounding rules***.

Why do we need such a model?

Because the relationship between Chinese constituent characters and the compound words is ***less than systematic***.



✗ simple linear combination of the constituent embeddings

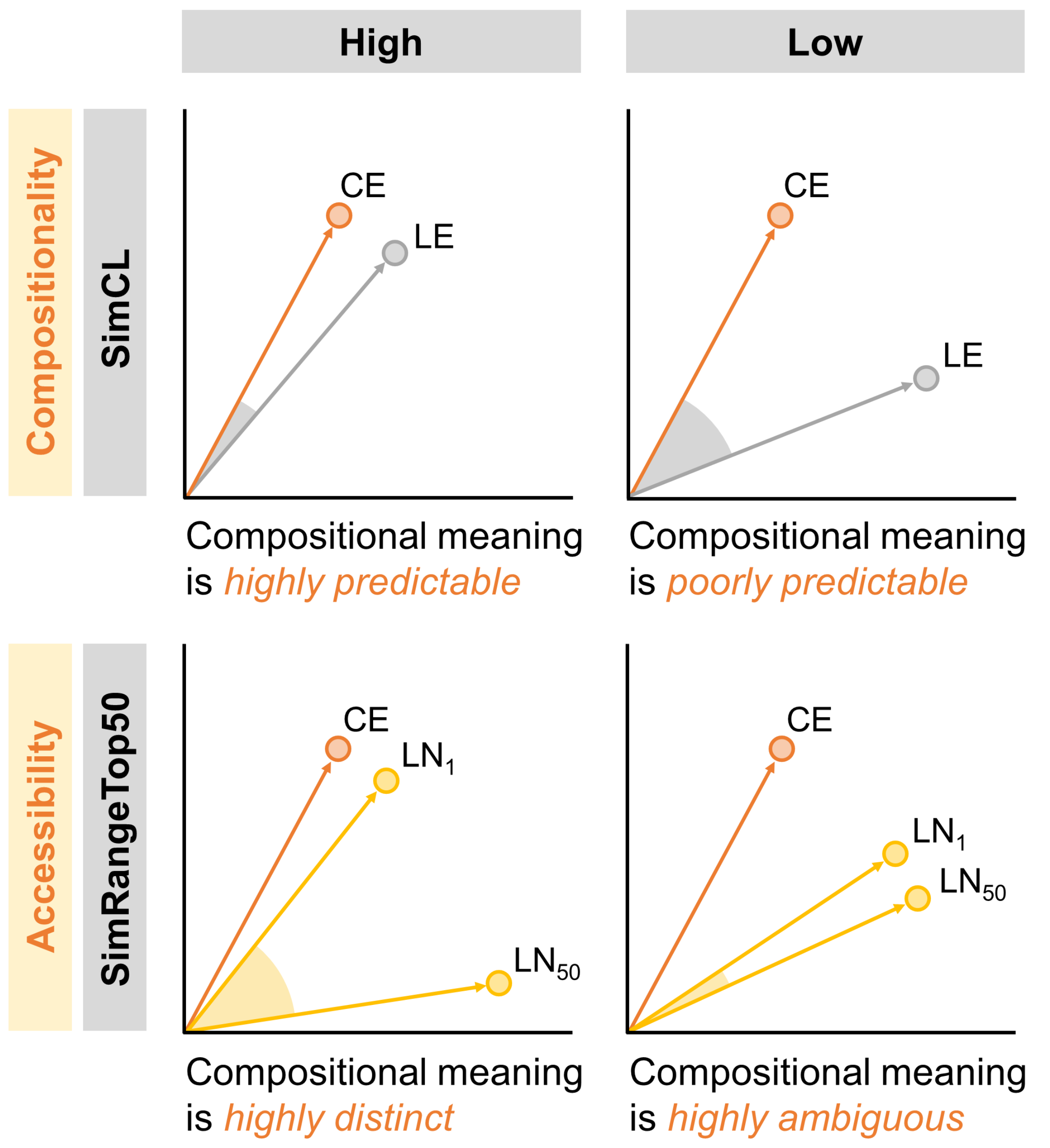
✓ dynamic ***nonlinear combination*** of the ***role-dependent*** constituent embeddings

Loss: MSE
compositional compound embedding
actual compound embedding
FC layer
pretrained language model
BERT
[CLS]
C1
C2
···
Cn
constituent characters

Computed metrics

Two computed metrics are defined to characterize the end product of the combinatorial route.

- **SimCL:** the cosine between the compositional and lexicalized (actual) compound embeddings
- **SimRangeTop50:** the range of the cosine distances of the 50 lexicalized neighbors that are closest to the compositional embedding



Compositional meaning is *highly predictable*

Compositional meaning is *poorly predictable*

Compositional meaning is *highly distinct*

Compositional meaning is *highly ambiguous*

CE = compositional embedding LE = lexicalized embedding
LN = lexicalized neighbor

How lexical decision times are influenced by the computed metrics

- **Dataset:** megastudy of lexical decision (Tsang et al., 2018) with 10,022 two-character compounds
- **Statistical analysis:** forward analysis with the computed metrics and potential interactions added to the linear mixed effects model over and above the lexical, semantic, and phonological variables (baseline)

Parameter	Estimate	*SE*	*t*	*df*	*p*	% ΔR^2	R^2
Intercept	-0.30	0.003	-90.82	943	< 0.001		
LogWF	-0.21	0.003	-68.17	9149	< 0.001		
Stroke	0.004	0.001	6.25	3044	< 0.001		
C1.LogCF	0.03	0.005	5.53	3028	< 0.001		
C2.LogCF	0.03	0.005	4.74	2220	< 0.001		
C1.LogFS	-0.06	0.009	-6.84	1788	< 0.001		
C2.LogFS	-0.06	0.010	-6.42	1318	< 0.001		
C2.LogNoM	0.06	0.014	4.24	1166	< 0.001		
C1.LogNoP	0.09	0.031	2.84	1323	0.005		
Baseline model							0.435
SimCL	-0.42	0.048	-8.65	9245	< 0.001	2.67	0.446
SimRangeTop50	-0.29	0.044	-6.72	8081	< 0.001	0.63	0.449
SimCL × LogWF	0.19	0.046	4.07	9111	< 0.001	0.27	0.451
SimRangeTop50 × C2.LogCF	0.19	0.043	4.45	7747	< 0.001	0.19	0.451
Computed metrics						3.80	0.451

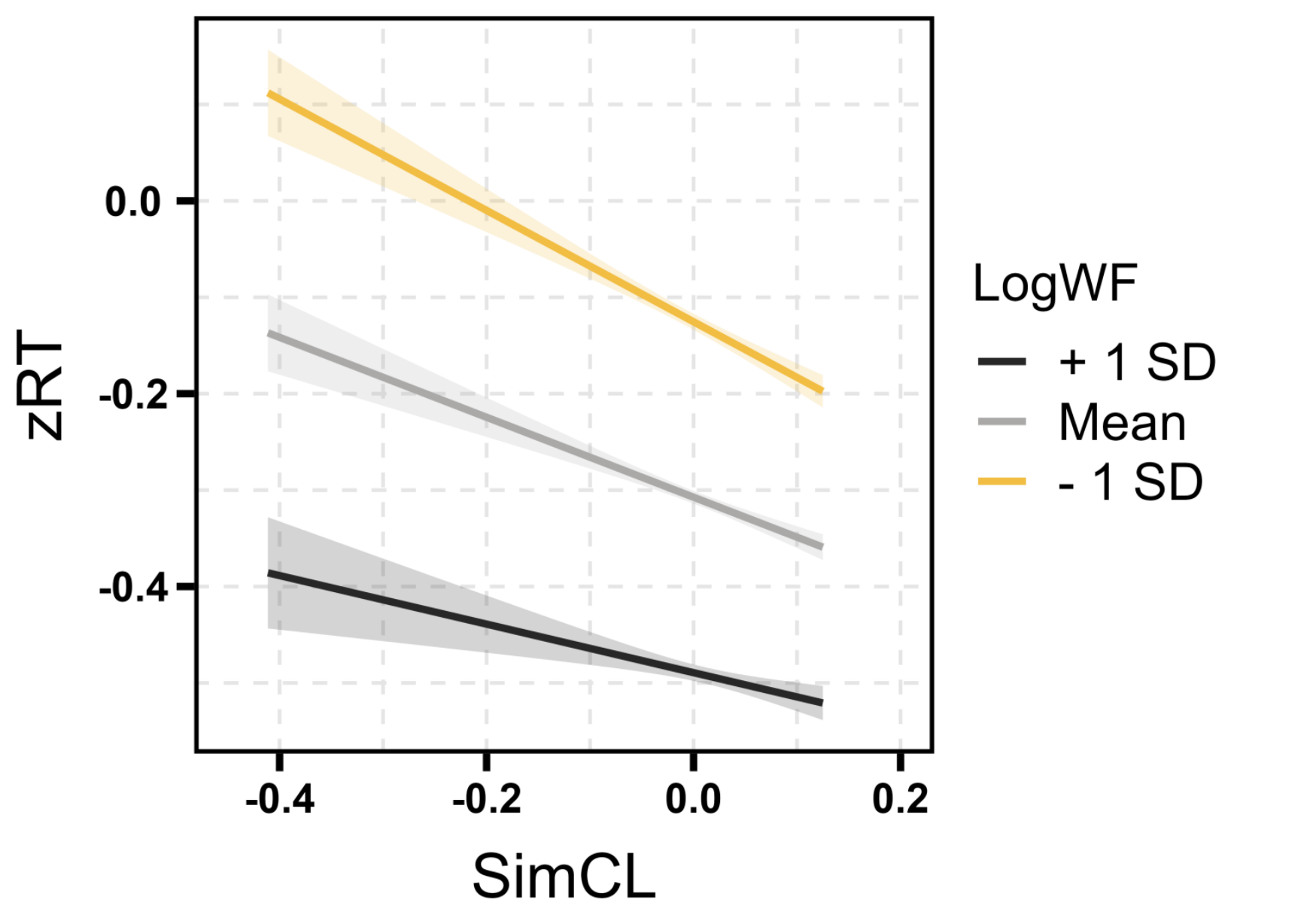
WF = word frequency CF = character frequency FS = family size NoM = number of meanings NoP = number of pronunciations
C1 = first character C2 = second character

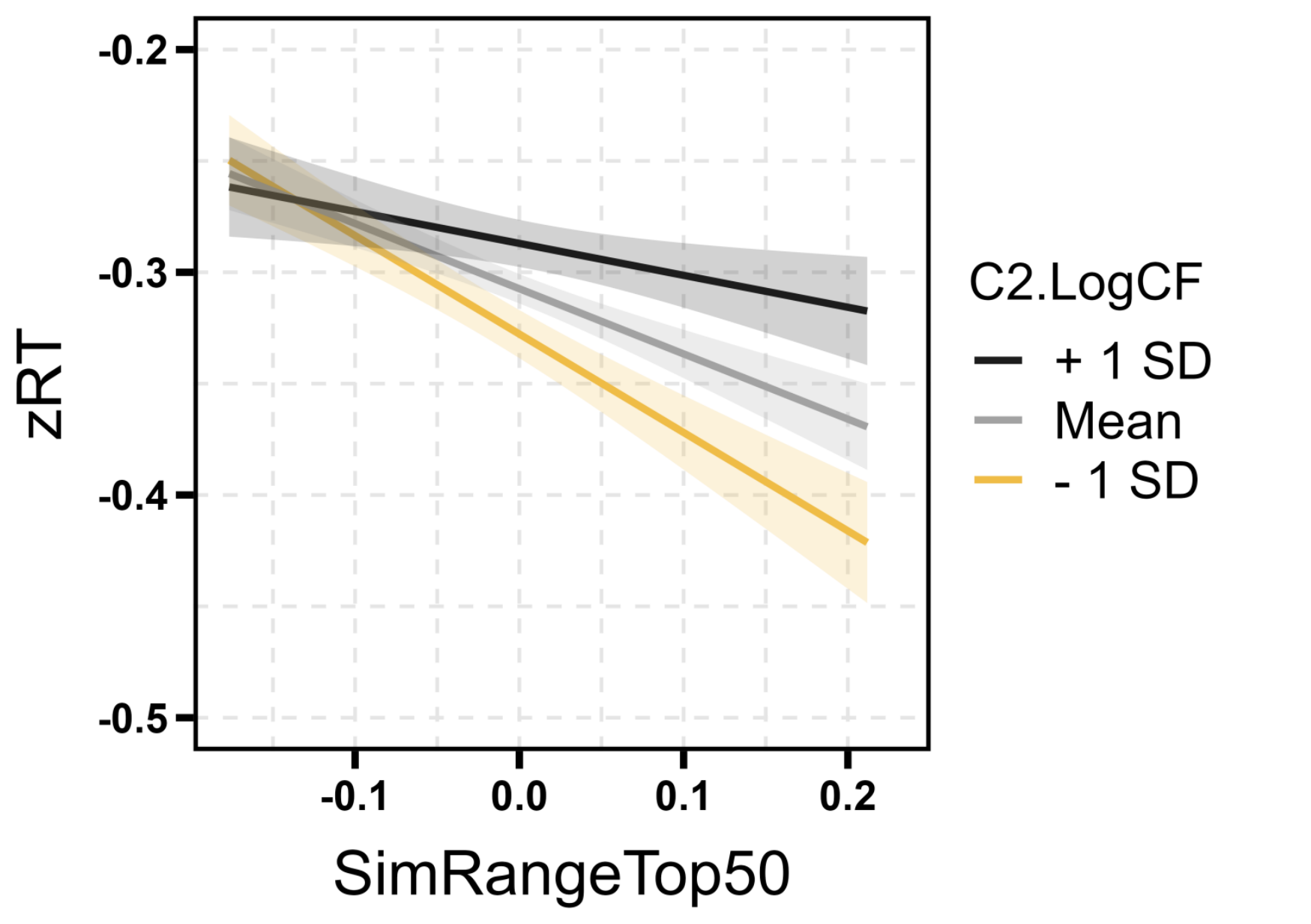
Results

Efficacy of the computed metrics

- The inclusion of the two metrics, SimCL and SimRangeTop50, significantly improved the fit of the baseline model, $\chi^2(2) = 233.59$, $p < 0.001$
- Both metrics showed ***facilitatory effect*** on lexical decision times

Interactions with other variables





Take-home Message

- A ***combinatorial process*** is actively involved in Chinese compound processing, which is moderated by word frequency, i.e., an indicator on whether the ***holistic route*** is likely to prevail
- Two attributes associated with the end product of the combinatorial route, i.e., ***compositionality*** and ***accessibility*** of the compositional representation, can affect the efficiency of compound processing
- The computational characterization of the dual-route framework sheds light on the ***universal process*** of compound comprehension

Article Code Contact

📢 This is not the end of the story. We recently obtained encouraging evidence from behavioral response for nonword rejection, ERP response for Chinese word recognition, as well as 👁 eye-tracking data in sentence reading. ❤ Follow us and we will keep you in the loop!