

## Can Semantic Similarity Capture Headedness Distributions? Evidence from Chinese Disyllabic Compounds

**INTRODUCTION** Headedness in Chinese compounding has long been debated. Building on compound typologies and Chinese-specific classifications (Bisetto & Scalise, 2005; Ceccagno & Scalise, 2006; Ceccagno & Basciano, 2007, 2008), we create a lexicon-based dataset of 5,024 Chinese disyllabic compounds (see Appendix A) to first replicate the classification and distributional patterns reported for the four headedness patterns (left, right, two, and exocentric) and three structural classes (subordinate, attributive, and coordinate). We then compute semantic similarity between compounds and their constituents using word embeddings, defining two metrics,  $\Delta$  and  $\tau$ , that quantify semantic asymmetry and semantic proximity in the embedding space. By applying statistical tests, we evaluate whether distributional semantic measures can provide a valid quantitative approximation of headedness, and more broadly, whether semantic similarity can capture headedness distributions in Chinese disyllabic compounds.

**REPLICATION ANALYSIS** We first replicate generalizations previously established on neologism corpora (Ceccagno & Basciano, 2007, 2008) using a broader lexicon. Comparisons across (i) class and category distributions, (ii) category and headedness distributions across classes, (iii) category distribution across headedness, and (iv) the most frequent constructions across classes all point to a high degree of consistency between our dataset and that of the original study. This analysis provides large-scale empirical support for the distributional generalizations proposed for Chinese disyllabic compounds and confirms the reliability and internal consistency of our dataset with respect to the examined features.

**SIMILARITY ANALYSIS** Using Tencent Chinese word embeddings (Song et al., 2018), we define and compute  $\Delta$  and  $\tau$  based on  $\text{sim}(A)$  and  $\text{sim}(B)$  (see Appendix B). The two derived metrics target two distinct theoretical questions. First,  $\Delta$  operationalizes directional semantic attraction: if headedness corresponds to the constituent that contributes more strongly to the compound's distributional meaning, then left-headed compounds should yield  $\Delta > 0$ , right-headed compounds  $\Delta < 0$ , and two-headed compounds should concentrate near  $\Delta \approx 0$ . Second,  $\tau$  operationalizes constituent anchoring: if exocentric compounds are semantically "external" to both constituents, neither  $\text{sim}(A)$  nor  $\text{sim}(B)$  should be high, so  $\tau$  should be systematically lower for exocentrics than for endocentric compounds. Because  $\Delta$  encodes relative dominance while  $\tau$  encodes absolute alignment, they are not expected to correlate strongly; we check this empirically before interpreting them as complementary signals.

**RESULTS** Based on three statistical tests, we find that (i) the sign and magnitude of  $\Delta$  reliably track the expected directional asymmetry: left-headed compounds show  $\Delta > 0$ , right-headed compounds  $\Delta < 0$ , and two-headed compounds cluster near zero, though not significantly (see Figure 1), and (ii) exocentric compounds exhibit markedly lower  $\tau$  values than non-exocentric compounds (see Figure 2), providing strong evidence that  $\tau$  effectively captures exocentricity. Our results therefore suggest that semantic similarity does provide a meaningful approximation of headedness distributions in Chinese disyllabic compounds: it captures right-headedness particularly well and left-headedness and exocentricity robustly, and reflects two-headedness in the expected direction, though more weakly. Future work may apply the same approach cross-linguistically to help determine whether the semantic correlates of headedness identified here are specific to Chinese or reflect more general properties of compounding across languages. (493 words)

## APPENDIX

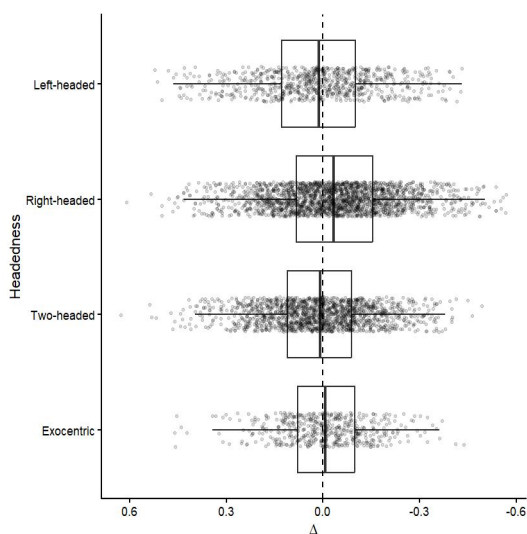
**A. Eight annotated examples. Glosses in the last column are provided for reference only and are not part of the annotated dataset.**

Compound	Class	Construction	Category	Head	Gloss
攀高	SUB	V+A	V	left	climb + high = climb upward
开会	SUB	V+N	V	left	hold + meeting = hold a meeting
品读	ATT	V+V	V	right	savor + read = read with appreciation
高考	ATT	A+N	N	right	high + exam = college entrance exam
打压	CRD	V+V	V	two	strike + suppress = crack down on
酸甜	CRD	A+A	A	two	sour + sweet = sweet-and-sour
网虫	SUB	N+N	N	exo	net + insect = web enthusiast
花心	ATT	A+N	A	exo	flashy + heart = unfaithful

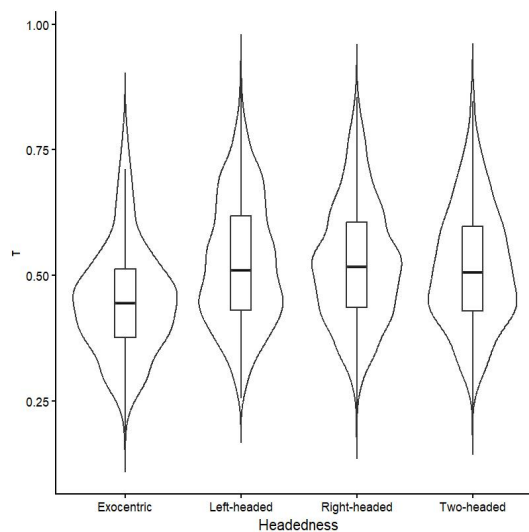
## B. Formulae used

$$\begin{aligned} \text{sim}(A) &= \cos(A, AB) \in [-1, 1] \\ \text{sim}(B) &= \cos(B, AB) \in [-1, 1] \\ \Delta &= \text{sim}(A) - \text{sim}(B) \in [-1, 1] \\ \tau &= \max(\text{sim}(A), \text{sim}(B)) \in [-1, 1] \end{aligned}$$

## C. Figures



**Figure 1:  $\Delta$  by headedness**



**Figure 2:  $\tau$  by headedness**

**REFERENCES** [1] Bisetto, A., & Scalise, S. (2005). The classification of compounds. *Lingue e Linguaggio*, 4(2), 319–332. [2] Ceccagno, A., & Scalise, S. (2006). Classification, structure and headedness of Chinese compounds. *Lingue e Linguaggio*, 5(2), 233–260. [3] Ceccagno, A., & Basciano, B. (2007). Compound headedness in Chinese: An analysis of neologisms. *Morphology*, 17(2), 207–231. [4] Ceccagno, A., & Basciano, B. (2008). Classification of Chinese compounds. *Proceedings of the Sixth Mediterranean Meeting of Morphology*, 6, 71–83. [5] Song, Y., Shi, S., Li, J., and Zhang, H. (2018). Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 175–180.