Understanding written compounds: a computational model

Yu-Hsiang Tseng[1], Tian Shen, Maziyah Mohamed, Rolf Harald Baayen[1]

[1] Department of Linguistics, Eberhard-Karls-Universität Tübingen, Germany

Compound interpretation is a central topic in the literature on lexical processing. To better understand compound interpretation, we propose a computational model with two independent parallel procedures. The first procedure is based on subliminal low-level statistical learning. It takes as input a high-dimensional numeric representation of a word's form (e.g., based on English triphone features, encoded as letter n-grams), and maps this into a high-dimensional semantic space using word embeddings. This procedure is part of what Huettig and Hulstijn [1] call 'Basic Language Cognition', the cognition serving language use in informal settings.

The second procedure makes use of higher-level lexical form units that represent the orthographic words that readers know thanks to extensive training in literacy. Formally, these higher-level units function as visual access units for compounds and their immediate constituents. We assume that these visual access units are in a one-to-one correspondence with their associated semantic embeddings, that is, they effectively serve as pointers to their embeddings. This second procedure is meta-linguistic in nature and part of what Huettig and Hulstijn [1] call the 'Enhanced Literary Mind'. It incorporates an analogical computation which predicts a compound's meaning from two sources: first, the meanings of its left and right constituents; second, semantic shift vectors derived from the left and right constituent compound families, each computed as the centroid of offsets from a constituent itself to the constituent's compound family members. This analogical computation formalizes Libben's insight concerning morphological transcendence in compound interpretation, albeit in a way that differs from the CAOSS model [2].

These two procedures were computationally implemented using the Discriminative Lexicon modeling framework [3], and applied to three datasets: 4,817 compounds from the English LADEC database [4], 5,285 from the MELD-SCH [5] dataset in Simplified Chinese, and 8,607 compounds from the Chinese Lexicon Project [6, 7] in Traditional Chinese. The new analogy model takes the embeddings of the two constituents as input and predicts the embedding of the compound. We evaluated prediction accuracy by computing correlations between the predicted compound and all compound embeddings in the dataset, counting a prediction as correct when the target compound had the highest correlation. CAOSS received the same input and was evaluated in the same way. The analogy computational model provides a better account of compound interpretation than the CAOSS model, as reflected in the held-out prediction accuracies (Figure 1). Correlation-based predictors derived from the model were fitted to the response latencies using GAMs [8] and outperformed the predictors derived from the CAOSS model by a wide margin. For nominal compounds with clear conceptual relations, the analogical computation can be seen as a possible formalization of CARIN theory [9], but the analogical computation is more general and not restricted to conceptual relations.

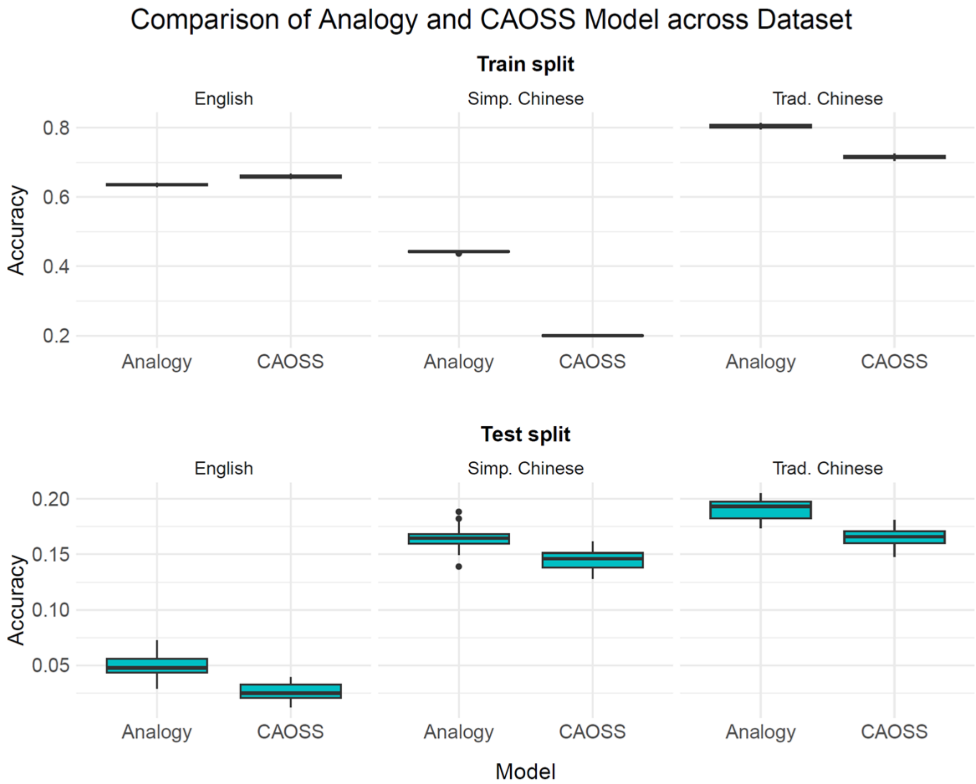

Figure 1: Prediction accuracies of compounds across three datasets. Analogy indicates predictions based on the proposed model in this study; CAOSS refers to the model of [2]. Each boxplot summarizes the scores obtained from 30 repeated random train-test splits. For the training split, the Analogy model has slightly lower performance than the CAOSS model in English, but better performance for both the simplified and traditional Chinese datasets. However, for the test split, the Analogy model consistently outperforms the CAOSS model across the three datasets.

References

[1] Huettig, F., & Hulstijn, J. (2025). The enhanced literate mind hypothesis. *Topics in Cognitive Science*, *17*(4), 909–918. https://doi.org/10.1111/tops.12731

[2] Günther, F., & Marelli, M. (2023). CAOSS and transcendence: Modeling role-dependent constituent meanings in compounds. *Morphology*, *33*, 409–432. https://doi.org/10.1007/s11525-021-09386-6

[3] Heitmeier, M., Chuang, Y.-Y., & Baayen, H. (2026). *The Discriminative Lexicon*. Cambridge University Press. https://doi.org/10.1017/9781009634564

[4] Gagné, C., Spalding, T., & Schmidtke, D. (2019). LADEC: The large database of english compounds. *Behavior Research Methods*, *51*(5), 2152–2179. https://doi.org/10.3758/s13428-019-01282-6

[5] Tsang, Y.-K., Huang, J., Lui, M., Xue, M., Chan, Y.-W. F., Wang, S., & Chen, H.-C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, *50*(5), 1763–1777. https://doi.org/10.3758/s13428-017-0944-0

[6] Tse, C.-S., Yap, M., Chan, Y.-L., Sze, W., Shaoul, C., & Lin, D. (2016). The chinese lexicon project: A megastudy of lexical decision performance for 25, 000+ traditional chinese two-character compound words. *Behavior Research Methods*, *49*(4), 1503–1519. https://doi.org/10.3758/s13428-016-0810-5

[7] Hsieh, C.-Y., Marelli, M., & Rastle, K. (2025). Compositional processing in the recognition of Chinese compounds: Behavioural and computational studies. *Psychonomic Bulletin & Review*, *32*(4), 1878–1889. https://doi.org/10.3758/s13423-025-02668-8

[8] Wood, S. (2017). *Generalized additive models: An introduction with r (2nd)*. CRC Press.

[9] Gagné, C., & Shoben, E. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 71–87. https://doi.org/10.1037/0278-7393.23.1.71