# A Semantic-pragmatic Embedding Approach to Investigate the Evolution of Basic-level Categories in Ancient Chinese

Anonymous Author 1[*1,2], Anonymous Author 2[2,3], and Anonymous Author 3[3]

[*]Corresponding Author: name@domain.com

Extending a weighted-entropy embedding model recently developed (Ji, 2022), this paper presents a novel computational approach to investigate the evolution of basic-level categories in ancient Chinese in a unified semantic-pragmatic 2D embedding space. The embedding space is formed by a basic semantic dimension (semantic typicality) and a basic pragmatic dimension (pragmatic salience). In particular, this study will focus on examine ancient Chinese basic-level categories in three domains that draw significant research interests in cognitive science, linguistics, and linguistic anthropology: color, smell, and classifiers.

Recent advancements in NLP, particularly distributional algorithms like word embedding, have revolutionized diachronic linguistics and research on language evolution, enabling nuanced analyses of linguistic features and sociocultural phenomena over time (Garg et al., 2016; Hamilton et al., 2016; Koslowski et al., 2018). But distributional embedding methods are often domain-generic and poorly address domain-specific questions on language variations and evolution. On the other hand, though studies in cognitive science and anthropology on domain-specific basic-level categories have a long tradition of addressing diachronic questions (Berlin & Kay, 1969; Regier et al., 2015, 2017; Zaslavsky et al., 2018), there is a lack of effective computational methodology to study domain-specific language evolution based on large-scale textual data.

The weighted-entropy embedding model (Authors, 2022) arguably provides an effective modeling approach to bridge the gap. In this work, we extend the model and argue that the model includes two interpretable dimensions that can co-embed both semantic hierarchy and pragmatic similarity. Applying it to diachronic changes of basic-level categories can provide a new computational methodology to investigate domain-specific language evolution.

Empirically, we extend the model to evaluate and map the evolution of basic-level categories of color, smell, and classifiers in ancient Chinese, benefiting from the region's extensive and relatively continuous textual history. In cognitive science and anthropology, these three domains are among the most studied and debated for questions on cross-linguistic domain-specific categorization (besides color, for smell: Majid, 2021; for classifiers: Lucy & Gaskins, 2001; Hopkins, 2012). Specifically, many studies in cognitive science emphasize the semantic and denotational nature of basic-level categories and its fundamental role in driving their evolution (Berlin & Kay, 1969; Regier et al., 2015, 2017; Zaslavsky et al., 2018). But critiques in anthropology instead argue that basic-level categories are predominantly structural and pragmatic (Sahlins, 1976; Lucy, 1997).

Results from the semantic-pragmatic embedding modeling on ancient Chinese in this paper demonstrate the follows: (1). While basic-level categories in color in contemporary simplified Chinese predominantly align along the semantic typicality dimension, linguistic color categories in ancient Chinese predominantly aligned along the dimension of pragmatic salience for more than 1000 years (e.g., Figure 1). (2). Basic-level smell terms, on the other hand, are shown to co-evolve with several major basic-level color terms. (3). The more structurally significant classifiers are shown to mainly align along the semantic typicality dimension in the medieval period (prior to 1000 AD), but some of major classifiers shift towards aligning along the pragmatic salience dimension throughout the second millennium.

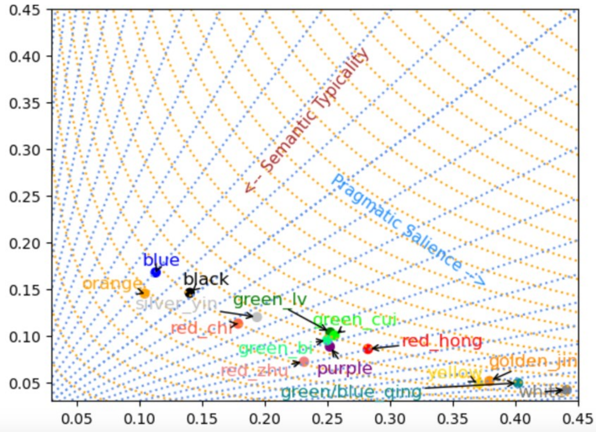

Figure 1. A semantic-pragmatic mapping of candidate basic-level color categories in ancient Chinese in Ming dynasty (1368 -1644). Unlike basic-level color terms in modern English and Chinese, which are predominantly aligned along the semantic typicality dimension (which would be distributed diagonally, not shown in the figure), colors terms in Ming dynasty are mainly aligned along the pragmatic salience dimension.

Together, this study provides a new distributional embedding method for modeling evolution of domain-specific lexical categories. The results can help address several major related debates between cognitive science and anthropology. Empirical results from the evolution of ancient Chinese demonstrate that the evolution of domain-specific linguistic categories can maintain on either semantic or pragmatic dimensions, or drift between the two, depending on the domain and the historical period.

## References

Berlin, B., & Kay, P. (1969). Basic Color Terms: Their Universality and Evolution. University of California Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender & Ethnic Stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644. https://doi.org/10.1073/pnas.1720347115

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2116–2121). Austin, TX: Association for Computational Linguistics.

Hopkins, N. A. (2012). The Noun Classifiers of Cuchumatán Mayan Languages: A Case of Diffusion From Otomanguean. International Journal of American Linguistics, 78(3), 411–427. https://doi.org/10.1086/665654

Ji, E. Y. (2022). A cognitively driven weighted-entropy model for embedding semantic categories in hyperbolic geometry (Preprint). https://arxiv.org/pdf/2112.06876.pdf

Lucy, J. A. (1997). The Linguistics of 'Color'. In C. L. Hardin & L. Maffi (Eds.), Color Categories in Thought and Language (pp. 320-346). Cambridge University Press. https://doi.org/10.1017/CBO9780511519819.015

Lucy, J. A., & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: A comparative approach. In M. Bowerman & S. C. Levinson (Eds.), Language acquisition and conceptual development (pp. 257-283). Cambridge University Press.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2018). The Geometry of Culture: Analyzing Meaning through Word Embeddings. American Sociological Review, 84(2). https://doi.org/10.1177/0003122419877135

Majid, A. (2021). Human olfaction at the intersection of language, culture, and biology. Trends in Cognitive Sciences, 25(2), 111–123. https://doi.org/10.1016/j.tics.2020.11.005

Regier, T., Kemp, C., & Kay, P. (2015). Word Meanings across Languages Support Efficient Communication. In B. MacWhinney & W. O'Grady (Eds.), The Handbook of Language Emergence (pp. 237-263). Wiley-Blackwell.

Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. Wiley Interdisciplinary Reviews: Cognitive Science, 8, e1440. https://doi.org/10.1002/wcs.1440

Sahlins, M. (1976). "Colors and cultures." Semiotica 16, no. 1: 1–22. https://doi.org/10.1515/semi.1976.16.1.1

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In Proceedings of the 41st Annual Meeting of the Cognitive Science Society.