The 'aftermath' of compounds: Investigating Compounds and their Semantic Representations

Swarang Joshi
International Institute of Information Technology, Hyderabad, India
swarang.joshi@research.iiit.ac.in

Abstract

This study investigated how well computational embeddings aligned with human semantic judgments in the processing of English compound words. We compared static word vectors (GloVe) and contextualized embeddings (BERT) against human ratings of lexeme meaning dominance (LMD) and semantic transparency (ST) drawn from a psycholinguistic dataset. Using measures of association strength (Edinburgh Associative Thesaurus), frequency (BNC), and predictability (LaDEC), we computed embedding-derived LMD and ST metrics and assessed their relationships with human judgments via Spearman's correlation and regression analyses. Our study confirmed that contextualized embeddings (BERT) better mirror human semantic transparency judgments than static embeddings (GloVe)[1]. Specifically, BERT's ST values showed stronger correlation with human annotations (r=0.23 for frequency, r=0.10 for predictability) and ST predictions that more closely aligned with the expected range (BERT: 3.31-4.25 vs. human: 4.04-4.93), compared to GloVe's compressed range (1.62-3.16). BERT's LMD values also approximated the human midpoint (5.0) more closely than GloVe's representations. The results also showed that predictability ratings are strong predictors of semantic transparency in both human and model data. These findings advanced computational psycholinguistics by clarifying the factors that drove compound word processing and offered insights into embedding-based semantic modeling.

1 Introduction

Compound words, such as *teacup* or *bluebird*, pose a unique challenge for both psycholinguistic theory and computational semantics. They consist of two or more free morphemes whose combined meaning may be transparent, as in *teacup*, or less predictable, as in *butterfly*. Psycholinguistic research has long investigated how human readers decompose and interpret compounds, focusing on measures like lexeme meaning dominance (LMD) and semantic transparency (ST) to quantify how strongly constituents contribute to overall meaning (Juhasz et al., 2015). LMD quantifies which constituent (left or right) contributes more strongly to the compound's overall meaning, rated on a 1-9 scale where values <5 indicate left-constituent dominance, 5 represents equal contribution, and >5 indicates right-constituent dominance. ST measures how readily the compound's meaning can be inferred from its constituents, rated on a 1-7 scale where higher values indicate greater transparency.

With the advent of word embeddings, researchers have begun to probe whether static and contextualized vector representations capture such human semantic intuitions. Buijtelaar and Pezzelle (2023) pioneered an analysis using BERT embeddings, demonstrating that contextual models may better reflect psycholinguistic patterns than static models like GloVe. However, questions remain about which linguistic factors—frequency, predictability, and associative strength—most robustly predict human judgments and model-derived metrics across embedding types.

In this paper, we extended prior work by systematically comparing GloVe and BERT representations on a shared psycholinguistic dataset of 628 compounds annotated for LMD and ST. We integrated factor ratings from established resources—the Edinburgh Associative Thesaurus(Kazemi, 2015), the Large Database of English Compounds (LaDEC) (Gagné et al., 2019), and the British National Corpus (BNC)—and conducted correlation and regression analyses to evaluate the relative contributions of association, frequency, and predictability. Our contributions are threefold:

[1] Link to Code - https://github.com/jswarang12/aftermath-compounds

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 322–328
December 20-24, 2025 ©2025 Association for Computational Linguistics

1. We provide a comprehensive comparison of static versus contextual embeddings in modeling human compound processing.

2. We identify which linguistic factors most strongly drive embedding-based LMD and ST metrics and their alignment with human data.

3. We offer recommendations for embedding selection and feature integration in computational psycholinguistics.

2 Methodology

We used pre-trained versions of GloVe and BERT to obtain word embeddings. The Edinburgh Associative Thesaurus (Kazemi, 2015) and LaDEC: Large database of English compounds(Gagné et al., 2019) were used to get values of the factors - association strength, frequency, and predictability rating.

2.1 Embedding Extraction

We used the 300-dimensional GloVe vectors trained on 840B tokens. Each compound and constituent was extracted as its static vector representation. We used bert-base-uncased (Devlin et al., 2019) (12 layers, 768 dimensions) from Transformers (Wolf et al., 2020).

Contextualized and non-contextualized representations of compounds and their constituent lexemes were obtained. Cosine similarities between compounds and their constituent lexemes to model lexeme meaning dominance (LMD) and semantic transparency (ST) were computed using the formulae mentioned in (Buijtelaar and Pezzelle, 2023), and MAE and Spearman's correlation against human-annotated values were evaluated.

Following Buijtelaar and Pezzelle (2023), we computed LMD and ST using:

$$\text{LMD} = |\cos(\mathbf{v}_c, \mathbf{v}_l) - \cos(\mathbf{v}_c, \mathbf{v}_r)| \times 4 + 5$$

$$\text{ST} = \frac{\cos(\mathbf{v}_c, \mathbf{v}_l) + \cos(\mathbf{v}_c, \mathbf{v}_r)}{2} \times 3.5$$

where $\cos(\mathbf{v}_a, \mathbf{v}_b)$ computes cosine similarity between vectors $\mathbf{v}_a$ and $\mathbf{v}_b$, with subscripts c, l, r denoting compound, left constituent, and right constituent embeddings.

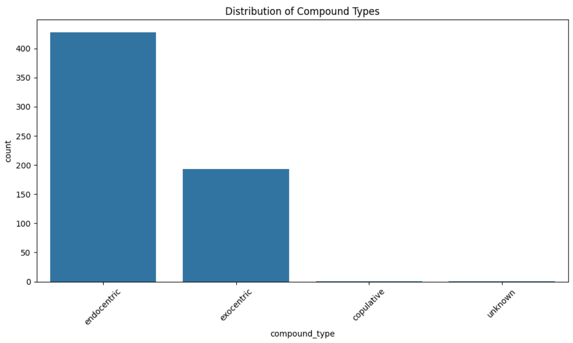

Figure 1: Compound type distribution in dataset (n=628): 68% endocentric, 31% exocentric, <1% copulative.

2.2 Metrics

Spearman's correlation and regression analysis were the primary statistical methods used to evaluate the relationship between the linguistic factors and LMD and ST values derived from human annotations, GloVe, and BERT embeddings. The association strength and frequency were measured only at the compound level, but the predictability rating for the lexemes (constituents) was also considered in the analysis.

Spearman's correlation was used to measure the strength and direction of the monotonic relationship between individual linguistic factors (association, frequency, and predictability) and our dependent variables (LMD and ST), identifying factors with significant standalone associations.

Regression analysis then assessed the predictive power of these factors. The resulting R^2 score from the regressors revealed the proportion of variance in LMD and ST that could be explained, offering deeper insight into a factor's explanatory utility beyond simple association.

3 Datasets

Psycholinguistic dataset (Juhasz et al., 2015) in processing containing 628 lexicalized English compounds annotated for LMD and ST.

We used Edinburgh Associative Thesaurus (EAT) (Kazemi, 2015) for word associations and LaDEC: Large database of English compounds(Gagné et al., 2019) for predictability and BNC word frequency.

4 Results

The MAE and Spearman correlation between the human judgments of LMD and ST and those de-

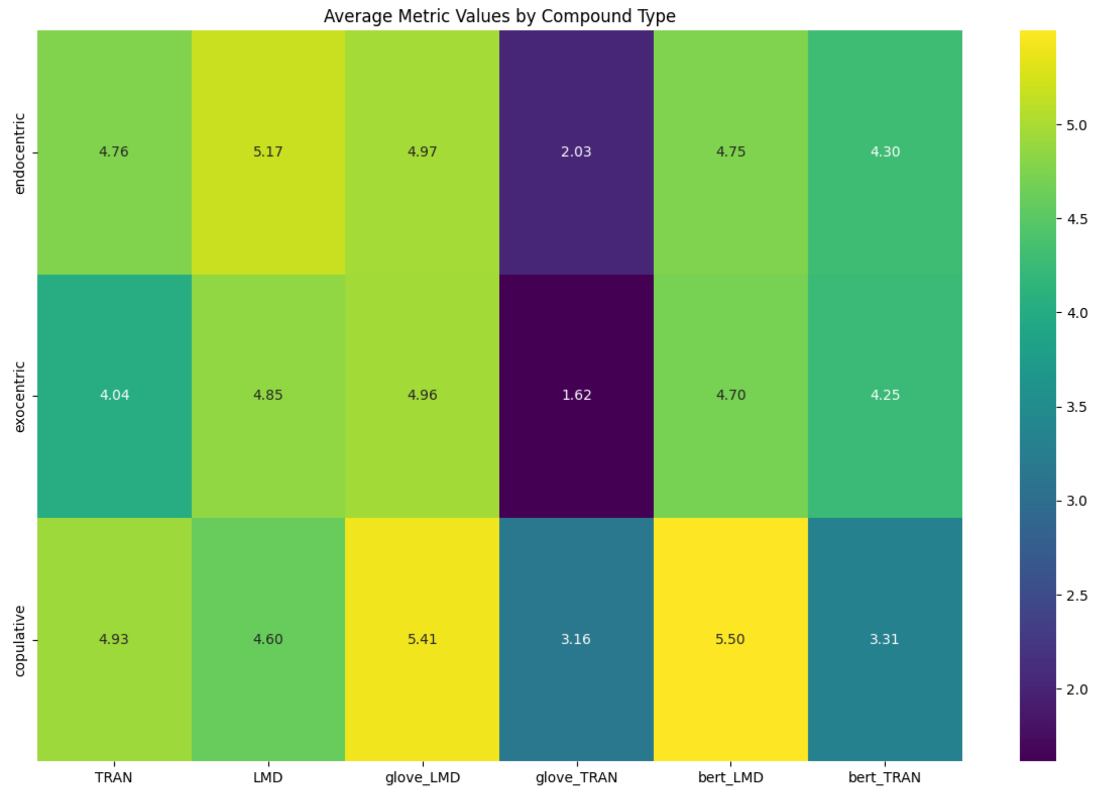

Figure 2: Compound Metrics Heatmap. TRAN refers to ST

Factor	Humans	Glove	BERT
Association	-0.0719	-0.2415	-0.0536
Frequency	-0.1395	-0.0172	0.0636
Frequency (R-L)	**-0.1714**	**-0.4345**	**-0.2303**
Frequency (R+L)	-0.1585	-0.0410	0.1023
Predictability	-0.1575	-0.0657	-0.0458

Table 1: Spearman correlation between **LMD** values and the factors

Factor	Humans	Glove	BERT
Association	0.2365	0.2300	0.0281
Frequency	-0.0588	**0.4410**	0.2319
Frequency (R-L)	-0.0351	0.0091	0.0636
Frequency (R+L)	0.0351	-0.0306	**0.2478**
Predictability	**0.7326**	0.3096	0.1033

Table 2: Spearman correlation between **ST** values and the factors

rived from Glove and BERT embeddings matched the values mentioned in the main reference paper (Buijtelaar and Pezzelle, 2023).

4.1 Correlation

From the Table 1 we can see that LMD had a negative correlation with all the factors. Among human-annotated values, predictability rating and frequency had a significant correlation. Only association are significantly correlated with Glove's values of LMD. In contrast, none of the linguistic factors we examined showed a significant correlation with the LMD values derived from BERT embeddings. Frequency (R-L) had the strongest correlation across all the representations.

From the Table 2 we can see that all the significant correlation between ST values and the factors are positive. Among human-annotated values, association strength was strongly correlated, followed by predictability strength. All three factors were significantly correlated with Glove's values of ST. Only the frequency and predictability ratings showed a significant correlation with the ST values of the BERT embeddings.

4.2 Regressors to Predict LMD and ST

The graphs in Figure 3 show the results of the regressors trained on the factors to predict the LMD and ST values. We can see that association strength is a poor predictor for both LMD and ST values. Frequency is only able to predict the LMD values from Glove embeddings. Predictability rating is a good predictor of only the ST values from human annotations.

5 Discussion

5.1 Compound Type Distribution and Embedding Model Performance

Our analysis revealed significant insights into both the distribution of compound types in English and how different embedding models capture their semantic properties. Figure 1 shows the overwhelming predominance of endocentric compounds in our dataset (approximately 68% endocentric vs. 31% exocentric and <1% copulative) confirms previous linguistic analyses of English compound formation preferences. Our dataset's composition, 68% endocentric vs. 31% exocentric— is consistent with patterns observed in previous compound studies (Libben et al., 1998), though we note this reflects the sampling strategy of Juhasz et al. (2015) rather than a representative survey of English compounding. This distribution reflected English's tendency toward transparent, compositional word formation strategies, where the semantic head is explicitly represented within the compound.

5.2 Semantic Transparency Across Compound Types

The transparency (ST) metrics revealed patterns that largely align with theoretical predictions from morphological theory. Figure 2 shows that Endocentric compounds demonstrated higher transparency values (4.76) than exocentric compounds (4.04), confirming that head-modifier relationships contributed to semantic predictability. This finding supported Libben et al. (1998) transparency hierarchy and Gagné and Spalding (2009) relational framework theories, which posit that compounds with clear internal semantic structures are more easily processed and interpreted. The surprisingly high transparency value for copulative compounds (4.93) suggested that coordinate relationships may be particularly accessible to speakers, despite their relative rarity in English. This might indicate that the balanced semantic contribution from both constituents created a unique form of transparency that differs from the asymmetrical relationship in endocentric compounds.

5.3 Model-Specific Representations of Compound Semantics

5.3.1 Divergence Between BERT and GloVe

The stark contrast between how BERT and GloVe represented compound transparency is one of our most striking findings. GloVe's transparency values were dramatically lower across all compound types (endocentric: 2.03; exocentric: 1.62; copulative: 3.16) compared to BERT's values, which more closely aligned with the original ST ratings. This suggested that contextual embeddings (BERT) may better capture the compositional nature of compounds than static embeddings (GloVe). The divergence can be attributed to fundamental architectural differences: BERT's bidirectional, contextual nature allowed it to better represent how compound meanings emerge from the interaction between constituents, while GloVe's context-independent vectors may struggle to capture these compositional semantics.

5.3.2 Lexical-Morphological Distance Patterns

The LMD metrics revealed a more complex picture than anticipated by straightforward compositional theories. Endocentric compounds showed higher LMD values than expected (5.17), suggesting that even semantically transparent compounds maintained distinct representations from their constituents in embedding space. This supported dual-route theories of compound processing (Kuperman et al., 2009), which proposed that compounds are accessed both as whole units and through individual units.

6 Conclusion

Our study confirmed that contextualized embeddings (BERT) better mirrored human semantic transparency judgments than static embeddings (GloVe), likely due to their capacity to model contextual interactions between morphemes. Predictability emerged as the most robust factor driving transparency, highlighting the role of semantic expectation in compound processing. These insights contributed to dual-route theories of morphological processing and informed the choice of embedding models for downstream applications.

Limitations

While our study shed light on how static (GloVe) and contextualized (BERT) embeddings captured human semantic intuitions for English compounds, there remain several limitations:

- **Language and Genre Coverage.** We focused exclusively on lexicalized English compounds drawn from a psycholinguistic dataset of 628 items. Our findings may not generalize to other languages (e.g., German, where compounding is more productive) or to less-frequent, novel compounds encountered in large-scale corpora.
- **Embedding Variants.** Only one static embedding (GloVe) and one contextualized model ($\text{BERT}_{\text{base}}$) were evaluated. Future work should explore additional architectures (e.g., RoBERTa, ALBERT, or contextualized static hybrids) and compare multilingual or specialized domain embeddings.
- **Psycholinguistic Measures.** We relied on pre-existing human ratings for lexeme meaning dominance (LMD) and semantic transparency (ST). These measures came from a single study and may embed annotation biases or inter-rater variability that could have influenced our correlation and regression results.
- **Downstream Task Validation.** Our evaluation metric is correlation with human judgments. We did not assess the impact of compound representation quality on downstream tasks (e.g., machine translation, lexical semantic annotation), which is an important avenue for future validation.

Acknowledgements

I would like to thank Pranav Agrawal and Prof. Rajakrishnan Rajkumar for their insightful inputs and guidance. I am also grateful to LTRC, IIITH for support. I also thank all reviewers for their insightful feedback, and the organizers of AACL-IJCNLP 2025 and the Student Research Workshop for their dedicated efforts.

References

Lars Buijtelaar and Sandro Pezzelle. 2023. A psycholinguistic analysis of BERT‘s representations of compounds. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2230–2241, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. Ladec: The large database of english compounds. *Behavior Research Methods*, 51(5):2152–2179.

Christina L. Gagné and Thomas L. Spalding. 2009. Constituent integration during the processing of compound words: The role of relational structures. In Brian H. Ross, editor, *The Psychology of Learning and Motivation*, volume 51, pages 97–130. Elsevier.

Barbara Juhasz, Brian Lai, and Ian Woodcock. 2015. Semantic transparency and constituent frequency effects in compound word processing. *Journal of Memory and Language*, 83:1–17.

Darius Kazemi. 2015. The edinburgh associative thesaurus (eat).

Victor Kuperman, Melvin J. Traxler, Ken McFalls, and Charles Cairns. 2009. Effects of morphological structure in compound word processing. *Journal of Memory and Language*, 61(1):24–44.

Gillian Libben, Angela Y. Weber, Michael Jarema, and Michael J. Pollatsek. 1998. Semantic transparency in native and second language compound processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5):1256–1273.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface’s transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

A Appendix

A.1 Graphs

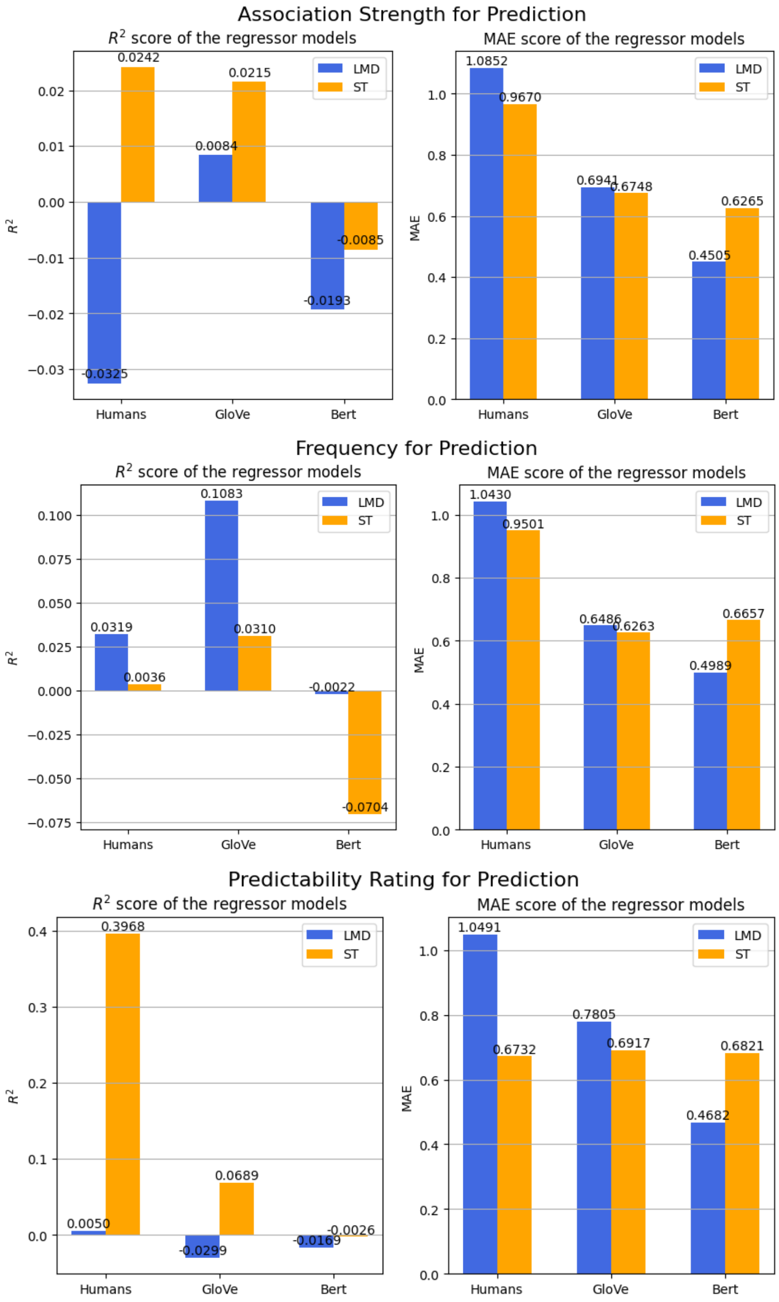

Figure 3: Performance of Regressors

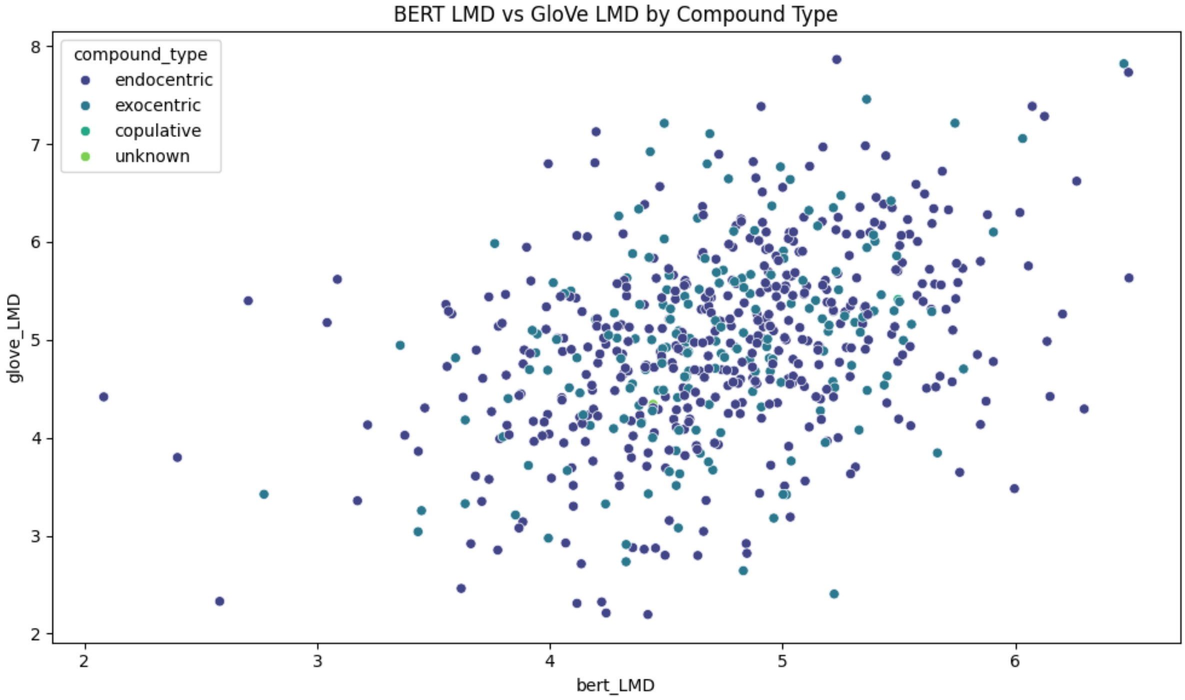

Figure 4: Bert vs GloVe LMD distribution

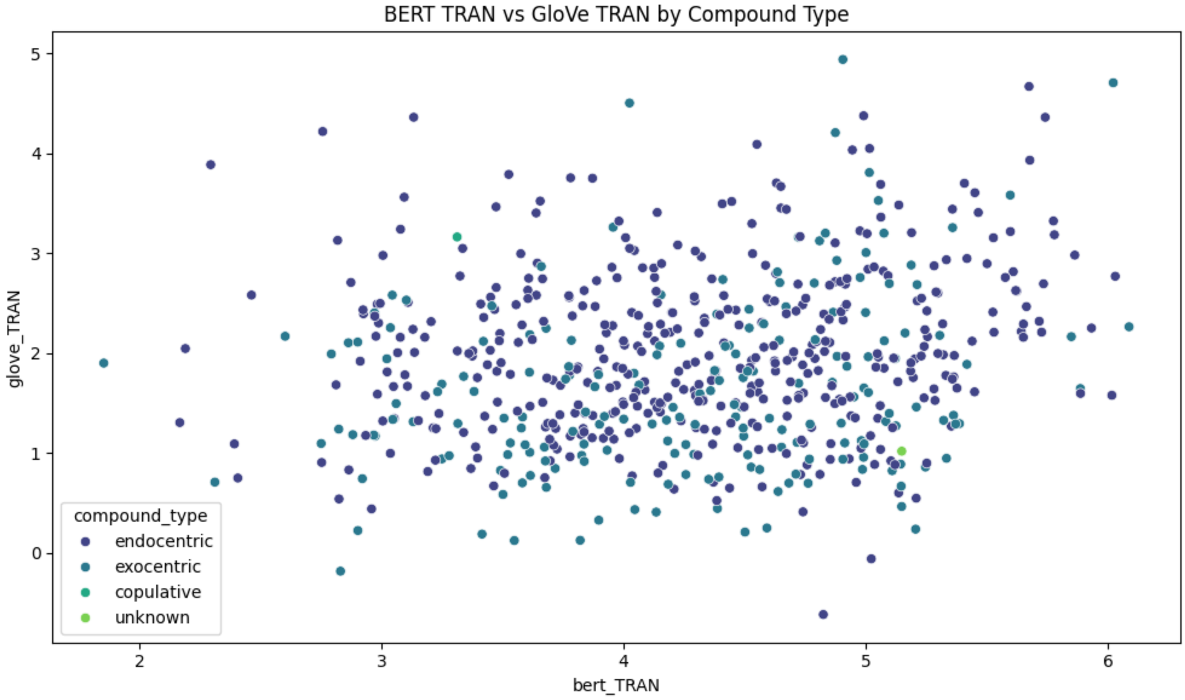

Figure 5: Bert vs GloVe LMD distribution