An Exploratory Analysis on the Explanatory Potential of Embedding-Based Measures of Semantic Transparency for Malay Word Recognition

Mirrah Maziyah Mohamed*
University of Tübingen, Tübingen, Germany
maziyah.mohamed@uni-tuebingen.de

R. Harald Baayen†
University of Tübingen, Tübingen, Germany
harald.baayen@uni-tuebingen.de

Abstract

Studies of morphological processing have shown that semantic transparency is crucial for word recognition. Its computational operationalization is still under discussion. Our primary objectives are to explore embedding-based measures of semantic transparency, and assess their impact on reading. First, we explored the geometry of complex words in semantic space. To do so, we conducted a t-distributed Stochastic Neighbor Embedding clustering analysis on 4,226 Malay prefixed words. Several clusters were observed for complex words varied by their prefix class. Then, we derived five simple measures, and investigated whether they were significant predictors of lexical decision latencies. Two sets of Linear Discriminant Analyses were run in which the prefix of a word is predicted from either word embeddings or shift vectors (i.e., a vector subtraction of the base word from the derived word). The accuracy with which the model predicts the prefix of a word indicates the degree of transparency of the prefix. Three further measures were obtained by comparing embeddings between each word and all other words containing the same prefix (i.e., centroid), between each word and the shift from their base word, and between each word and the predicted word of the "Functional Representations of Affixes in Compositional Semantic Space" model. In a series of Generalized Additive Mixed Models, all measures predicted decision latencies after accounting for word frequency, word length, and morphological family size. The model that included the correlation between each

*Corresponding author. ORCID: 0000-0002-3164-3805
†ORCID: 0000-0003-3178-3944

word and their centroid as a predictor provided the best fit to the data.

Keywords: semantic transparency, embeddings, morphology, lexical decision, Malay

1 Declaration

The authors have no competing interests to declare that are relevant to the content of this article. This research was funded in part by the European Research Council, grant #101054902 (SUBLIMINAL) awarded to Harald Baayen. Data are available `https://osf.io/dhyzb/?view_only=e05e71b31cb54daf94a55f46f9cc82da`

2 Introduction

A remarkable phenomenon in language processing in skilled readers is the ability to rapidly decode and extract meaning from written words. A growing body of research on semantic transparency addresses the ease with which a word's meaning is understood, with greater degrees of transparency associated with easier word recognition (e.g., Chee and Yap, 2022; Diependaele et al., 2009; Feldman et al., 2002; Jared et al., 2017; Libben et al., 2003; Marelli and Baroni, 2015). Semantic transparency is typically defined in terms of compositionality, that is, the extent to which the meaning of a complex word can be predicted from the meaning of each of its constituents. From a decompositional perspective of morphological processing (Taft and Forster, 1975), *adapatable* is transparent because the morphemes *adapt* + *-able* describes something or someone that possesses the ability to *adapt*, whereas *moonshine* is fairly opaque as it refers to a type of liquor, rather than following straightfowardly from the meanings of *moon* and *shine*. Word and paradigm, or realizational morphology, offers an alternative explanation in which the word itself represents the most basic unit and that the relationship of words is governed by rules of analogy (e.g., Hockett, 1954; Blevins, 2016). In this case, *moonshine* is regarded as semantically opaque because its features are neither related to those of the whole words *moon* nor *shine*. A point of departure between the two main approaches is whether or not there is an explicit representation of morphemes. In more recent distributed accounts of morphological processing (e.g., Baayen et al., 2011; Baayen et al., 2019; Gonnerman et al., 2007; Plaut and Gonnerman, 2000; Rueckl and Seidenberg, 2011), typically implemented in the form of connectionist models, morphemes are not explicitly represented. Instead, the representations of a word's form and meaning are shaped by its distributional properties such as the statistical co-occurrences between spelling and meaning.

There is no clear consensus yet in the operational definition of semantic transparency because word meaning can be studied in various ways (for details on experimental discrepancies, see Auch et al., 2020). In many of such studies, researchers have relied on human participant ratings, a method that is relatively labor intensive. Here, we explore semantic transparency using multidimensional word embeddings. Westbury et al. (2024) observed that the initial idea of using high-dimensional matrices to represent word meaning traces back to Osgood et al. (1957), despite the lack of computing power at the time. To represent meaning numerically in a high-dimensional space is, therefore, not entirely a new concept, but rather, a technique that has been refined over time (e.g., Landauer and Dumais, 1997; Lund and Burgess, 1996). Moreover, evidence from Bruni et al. (2014) suggests that the semantic relatedness of words represented by embeddings and human ratings are comparable. The present study capitalizes on recent computational advances to explore the use of high-dimensional word embeddings that may capture word meaning more comprehensively and possibly offer a greater ecological validity compared to small-scale participant ratings.

A primary goal for the present study is to further facilitate studies of Malay word recognition. The Malay language, or *Bahasa Melayu*, is a relatively understudied Austronesian language spoken in many regions of Southeast Asia such as Singapore, Malaysia, Brunei, and Indonesia. Malay is rich in derivational morphology with minimal inflection. Derivational affixes are typically used to form words that are related in meanings (e.g., *'baik' good, 'kebaikan' a good action/well-being*). To accomplish the goal of the present study, our first objective is to further augment the Malay Lexicon Project 3, a morphological database, to include a variety of semantic properties for words in the database. At present, the Malay Lexicon Project 3 has estimates of orthographic-semantic consistency calculated for a large subset of simple words. In this study, we calculated several measures that estimate the degree of semantic transparency for a large subset of complex words which will be added to the database. The secondary objective is to evaluate whether, and how well, each measure predicts response times.

3 Present Study

First, we explore the geometry of semantic transparency of complex words in a high-dimensional semantic space and describe our calculations of several measures of semantic transparency that use word embeddings. Then, we evaluated each measure by determining whether they predict lexical decision latencies in a series of Generalized Additive Mixed Models (GAMM).

3.1 Semantic Geometry of Derived Words

A technique that has been gaining traction and typically used in areas of machine learning to visualize high-dimensional data is the t-distributed stochastic neighbor embedding clustering analysis (t-SNE; Van der Maaten and Hinton, 2008). A t-SNE is an unsupervised nonlinear dimensionality reduction technique. A key insight from Distributional Semantics is that semantically related words appear in similar contexts. As such, semantically related words have similar embeddings that appear closer in the t-SNE space than other words. t-SNE has been used successfully to explore productivity and semantic transparency of German (Stupak and Baayen, 2022) and Mandarin (Shen and Baayen, 2022). Of particular relevance, complex words in German show clustering by derivational suffix, but not by particles. Similarly, in Mandarin, clusters were detected for suffixes. Importantly, clusters in a t-SNE are driven by information that is most saliently encoded in the embeddings.

In this study, we adopted this technique to explore the morphological structure of Malay prefixed words. Recent computational (Denistia and Baayen, 2022) and corpus-based work (Denistia et al., 2022) on Indonesian prefixation, an Austronesian language closely related to Malay, have shown that embeddings are informative in discriminating the semantics of prefixes *pe-* and *peN-*. Pre-trained 300-dimensional word embeddings were first extracted from FastText (Bojanowski et al., 2017) for all words in the MLP database for which there are embeddings. Of those, 4,226 words containing at least one of 10 prefixes were analyzed using the *Rtsne* package (Krijthe and Van der Maaten, 2015) in R. The resulting output are the spatial coordinates for each word in two dimensions as depicted in Figure 1. Considerable clustering is revealed for a variety of words colour-coded by their derivational prefix, except for words containing *peri-* (n=3), *pra-* (n=4), *pe-* (n=35), and *se-* (n=31) for which there are too few of such words in our dataset for meaningful clustering, if any, to occur. *MeN-* words (cyan) appear largely in the middle and represent the majority of the data. *BeR-* words (yellow) cluster in two groups in the bottom right. *TeR-* words (purple) mostly appear on the outskirts of meN- words in the middle. *PeN-* words (red) appear on the edges forming an outer ring. *PeR-* words (teal) are sparsely scattered only on the left. *Ke-* words (navy blue) cluster in two groups in the bottom left. Importantly, across a variety of prefixes, we observe a strong form-meaning correspondence such that words containing the same prefix appear closer in semantic space to each other than words that contain different prefixes.

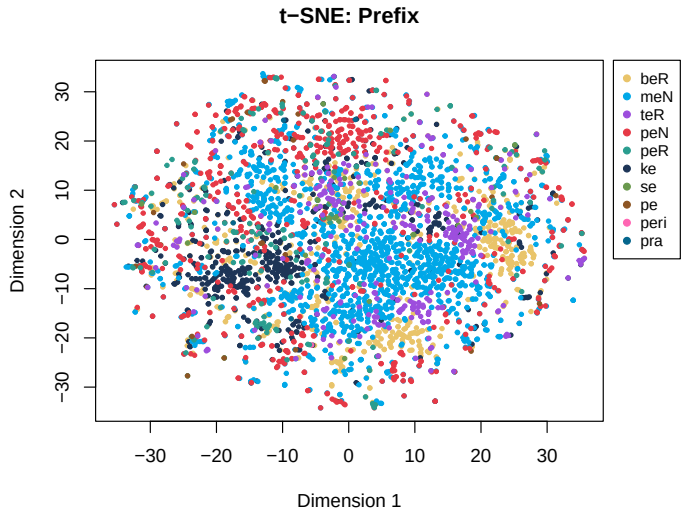

Figure 1: *Note.* Each coloured dot represents a derived word and each colour corresponds to a particular prefix a word contains. Words that have similar embeddings appear closer to each other in the t-SNE space than others.

3.2 Embedding-based Measures

3.2.1 Linear Discriminant Analysis

Following the t-SNE analysis, we computed two measures of semantic transparency of prefixes by conducting Linear Discriminant Analyses (LDA). LDA is an approach used in supervised machine learning to solve classification problems. Two sets of LDA models were run, each with a different input, and examined whether each word is linearly separable in the embedding space by their prefix. One dataset contained the vectors for each derived word. The second dataset contained shift vectors for each word, that is, the displacement in semantic space of the derived word from its base by subtracting their vectors. Importantly, in both cases, the accuracy with which the model successfully predicts a word's prefix is an index for the degree of correspondence between the form and meaning of a prefix. The predictions of the LDA model that are derived from a leave-one-out cross-validation approach are presented in Tables 1 and 2. Results from the LDA are in tandem with those of the t-SNE, that is, the LDA produced mostly correct classifications for words that contained prefixes that cluster successfully in the t-SNE. As an example, of 754 words that contained the prefix *beR-*, the LDA model accurately classified words containing the prefix *beR-* 713 times, yielding an accuracy of .946 for the prefix *beR-*. A cautious approach to the interpretation of the LDA results is to compare the proportion of correct classifications for each prefix against a baseline accuracy. The baseline accuracy in this case is .40, and can be calculated by taking the number of words that contain the prefix that occurs the most (i.e., *meN-*) divided by the total number of words. Overall, both LDA models predicted class membership accurately (.93 using derived word vectors, and .88 using shift vectors), providing evidence for the effectiveness of word embeddings and their shift vectors in discriminating words of different prefixes.

Table 1: Predictions of Class Membership using Embeddings

	beR	ke	meN	pe	peN	peR	peri	pra	se	teR	Total	Accuracy
beR	**713**	4	21	0	1	1	2	0	0	9	754	.946
ke	5	**456**	2	0	16	25	0	0	3	2	509	.896
meN	15	5	**1649**	0	0	2	0	0	4	5	1680	.982
pe	0	7	0	**14**	10	4	0	0	0	0	35	.400
peN	1	20	0	7	**618**	31	0	0	1	0	678	.912
peR	2	22	1	2	15	**155**	0	1	1	1	200	.775
peri	0	2	0	0	0	1	**0**	0	0	0	3	.000
pra	0	0	0	0	1	1	0	**2**	0	0	4	.500
se	5	1	3	0	0	0	0	0	**20**	2	31	.645
teR	14	1	3	0	1	1	0	0	1	**311**	332	.937

*Note.*The prefix in each row represents the prefix of a word and the prefix in each column represents the predicted prefix of a word. The bolded values represent the number of correct classifications of the LDA model. The overall accuracy is .93

Table 2: Predictions of Class Membership using Shift Vectors

	beR	ke	meN	pe	peN	peR	peri	pra	se	teR	Total	Accuracy
beR	**595**	6	23	0	4	4	0	0	3	11	646	.921
ke	15	**396**	4	3	22	26	0	0	1	4	471	.841
meN	32	4	**1290**	0	4	2	0	0	5	6	1343	.961
pe	2	6	0	**11**	10	1	0	0	0	0	30	.367
peN	5	24	0	5	**503**	33	0	0	1	1	572	.879
peR	8	29	1	2	30	**100**	0	0	0	3	173	.578
peri	1	1	0	0	0	1	**0**	0	0	0	3	.000
pra	1	1	0	0	1	0	0	**1**	0	0	4	.250
se	7	6	0	0	2	2	0	0	**8**	2	27	.296
teR	15	3	5	0	1	3	0	0	2	**214**	243	.881

*Note.*The prefix in each row represents the prefix of a word and the prefix in each column represents the predicted prefix of a word. The bolded values represent the number of correct classifications of the LDA model. The overall accuracy is .88

3.2.2 Correlation Measures

Three measures of semantic transparency were further calculated for each word. These are the correlation between each word and its prefix centroid (i.e., the mean vector of all words containing a particular prefix), between the vector of a word and its predicted vector derived from the Functional Representations of Affixes in Compositional Semantic Space (FRACSS; Marelli and Baroni, 2015) model, and between the derived and shift vectors for each word.

In distributional semantics, the more two words occur in similar con-

texts, the smaller the cosine of the angle between their vectors in semantic space, and the greater their overlap in meaning. Alternatively, a very similar measure that tends to be highly correlated with cosine similarity is the Pearson correlation between two vectors. A greater cosine similarity corresponds to a stronger correlation between two vectors. As such, the correlation between each derived word and its centroid estimates the similarity in meaning between each word and all other words containing the same prefix.

The correlation estimates derived from the FRACSS model (Marelli and Baroni, 2015) represent the similarity in meaning of the derived word and the predicted word. The FRACSS model proposed a linear mapping between the vectors of a derived word (e.g., revisit) and those of its base (e.g., visit). To implement FRACSS, the first step is to calculate the linear transformation that maps the vectors of a base word onto their corresponding derived words. This can be done by multiplying the matrix of the word vectors and the inverse of that of their base words. The next step is to calculate the predicted vectors of each word by multiplying the linear transformation with the vectors of the corresponding base words. For a step-by-step code on the implementation of the FRACSS model and a detailed discussion of FRACSS, see a JudiLing tutorial by Heitmeier et al. (2024). To compute the correlation estimates derived from FRACSS, the predicted vectors of each word are correlated with the vectors of the same word extracted from FastText. In simpler terms, the FRACSS correlation estimates represent the accuracy of the model in predicting a word, such that larger correlation coefficients indicate a more precise mapping for a derived word that is informed by its base.

Recall that the shift vector represents the displacement of a derived word from its base. To illustrate, on the top panel of Figure 2, a larger angle between the derived word and the shift vectors of a particular word corresponds to a smaller angle between the derived and its base vectors, thereby suggesting a smaller displacement of the derived form from its base as they appear closer to each other in semantic space. In contrast, on the bottom panel of Figure 2, a smaller angle (or stronger correlation) between the derived and shift vectors of a particular word corresponds to a larger angle (or weaker correlation) between the derived and its base vectors. In such a case, there is greater displacement of the derived word from its base, and are thus, semantically dissimilar as they appear far apart in semantic space. In our dataset, the two pairs of vectors (i.e., derived-shift, and derived-base) are strongly correlated, r = -.7.

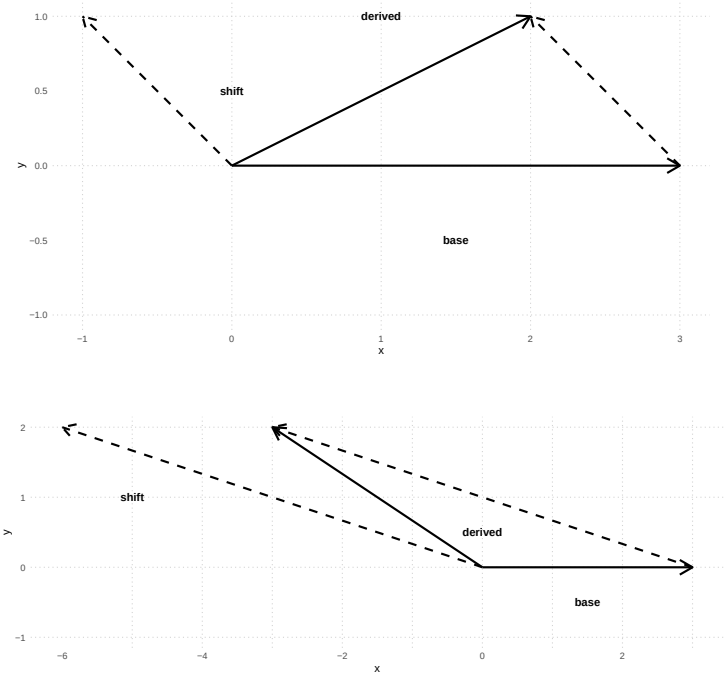

Figure 2: *Note.* Vector illustration. Solid lines indicate vectors for the base and derived word, and the dotted line represent the shift vectors. Of interest is the angle between the base and derived vectors, and the derived and shift vectors from the point of origin at (0, 0).

3.3 Evaluation of Measures on Behavioural Data

Each measure of interest was entered as a predictor, one at a time, in a series of Generalized Additive Mixed Models (GAMM) in R (R Core Team, 2025) using the *bam* function from the *mgcv* package (Wood, 2017) with the directive discrete set to TRUE. Both *bam* and *discrete=TRUE* make fitting a GAM model to a dataset much faster. Lexical decision latencies for 1,719 words from 280 participants were extracted from previous experiments of Malay visual word recognition (Maziyah Mohamed et al., 2023; Maziyah Mohamed and Jared, 2023; Maziyah Mohamed and Jared, 2025). Only correct responses and RTs between 350ms and 3000ms that were within 2.5 SDs from the overall mean RTs were analyzed, yielding a total of 42,934 observations. Below we report whether each proposed measure was a significant predictor of RT and compared the effectiveness of each measure as predictors of Malay word recognition, with careful

considerations of model residuals and concurvity statistics for model interpretability (see Supplementary Materials).

For comparison, we first ran a baseline model without the predictors of interest (see Table 3 and Figure 3). Whole-word frequency, word length, morphological root family size and its interaction with word frequency were entered as predictors. These predictors are shown to be crucial for Malay word recognition in previous studies of the Malay Lexicon Project (Maziyah Mohamed et al., 2023; Maziyah Mohamed and Jared, 2023; Maziyah Mohamed and Jared, 2025). Word frequency and root family size were log-transformed, and a *te* tensor product smooth was used to account for the main effect each of frequency and root family size, and their interaction. Random effects included trial number (centered and scaled) and subjects, using a factor smooth interaction *bs* = "*fs*" and a shrinkage directive *m=1*. This structure of random effects, adopted from Chuang et al. (2021) and Baayen et al. (2022), is analogous to a random effects structure in a linear mixed model that has by-subject random intercepts and random slopes for trials. Results revealed a significant interaction between whole-word frequency and root family size, such that a facilitative effect of root family size on RT was most evident for lower frequency words. A facilitative effect of word frequency was observed across a large range of root family sizes. In addition, an inhibitory effect of word length was observed, particularly for words that were very long (>11 letters). These effects of whole-word frequency, root family size, and word length on RT were consistently observed in subsequent models.

Table 3: GAMM - Baseline

Parametric coefficients				
Variable	Estimate	Std. Error	*t*	*p*
Intercept	-1.18	.03	-44.90	<.0001
ExpNo2	-0.21	.03	-7.18	<.0001
ExpNo3	-0.14	.03	-4.62	<.0001
Smooth terms				
Variable	edf	Ref.df	F	*p*
Frequency*Family Size	13.84	16.57	146.01	<.0001
Word Length	6.47	7.49	154.52	<.0001
TrialNo, Subjects	863.38	2518.00	7.95	<.0001
R^2 = .417				
AIC = 12415.81				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ~ te(frequency*family size) + s(word length) + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data.

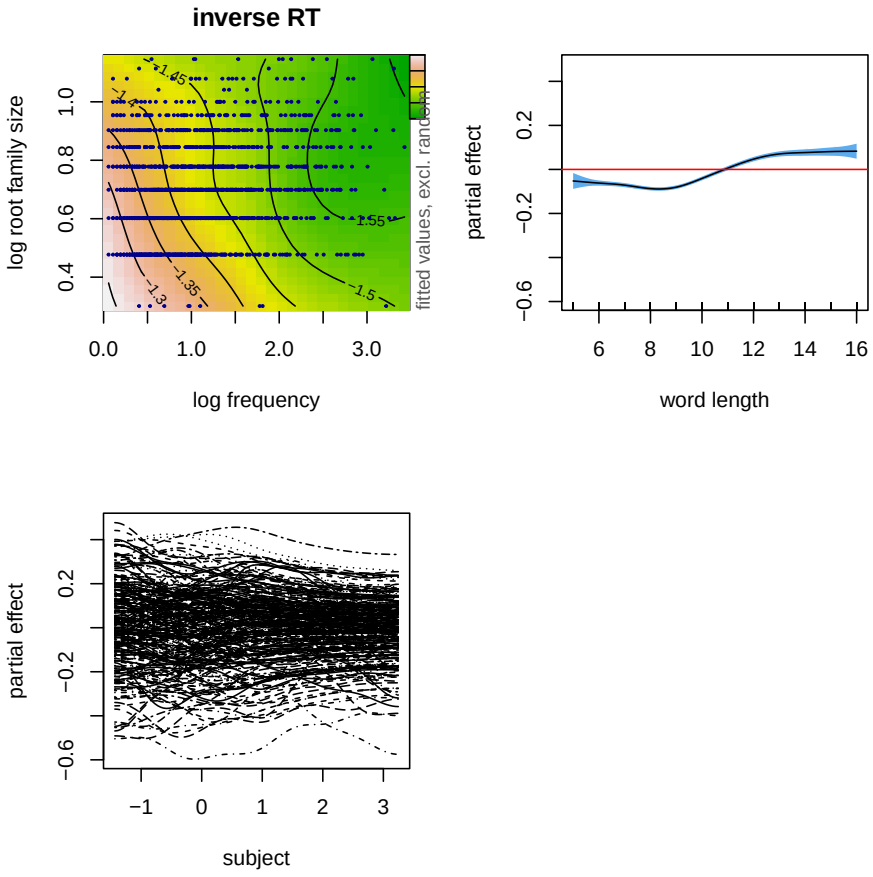

Figure 3: *Note.* Top row: Interaction between frequency and root family size (left). Data represented by dark blue points. Warmer colours (e.g., pink, orange) on the left-hand side denote longer RTs and cooler colours (e.g., green) on the right-hand side denote shorter RTs. Numbers on contour lines represent fitted inverse RT values. Partial effect of word length (right), with rugged lines on the x-axis representing the distribution of the data. Bottom row: Partial effects of trial number (centered and scaled) by subjects (left)

An additional five GAM models were run, each with all terms in the baseline model and one measure of interest at a time. The two LDA classification scores were each entered as a linear predictor because there were only 10 unique values, one for each prefix. All three correlation estimates were entered as predictors using a thin plate regression spline smooth. All five measures were significant predictors of RT. Most crucially, each model that included the predictors of interest provided a better fit to the

data than the baseline model (AIC = 12415.81).

A facilitative effect of the LDA accuracy was observed on RT. The greater the proportion of correct classifications for a particular prefix, the faster the responses. The accuracy of the LDA model in which shift vectors were used as input was a better predictor of RT (AIC = 12398.76; see Table 5) than the accuracy of the LDA that used the embeddings (AIC = 12406.96; see Table 4).

The effect of the correlation between each derived word and their centroid (AIC = 12325.08; see Table 6) on RT was facilitative, except for a relatively small number of words that were very strongly correlated with their centroids (r >.7 and above; see left panel of Figure 4). For a large majority of the data, the closer a word is to all other words that share the same prefix, the more easily it is recognized, as observed by shorter RTs. If, however, a word is too close in semantic space to all other words that share the same prefix, then word recognition appears more effortful, as observed by longer RTs. It is possible that processing is more effortful because such a word appears more confusable with its morphologically related words.

Additionally, we observed a facilitative effect of the FRACSS correlation estimates on RT (AIC = 12340.77; see Table 7). Higher estimates indicate a more precise prediction of the derived word. Faster responses were observed for words that were predicted more accurately by FRACSS (see middle panel of Figure 4). In contrast, we observed an inhibitory effect of the correlation between the derived and shift vectors of each word (AIC = 12381.06; see Table 8). As mentioned earlier, a stronger correlation between the derived and shift vectors of a word indicate a greater displacement of the derived form from its base. In such cases, slower responses are elicited (see right panel of Figure 4). Across all five measures, the model that included the correlation between the vectors of the derived word and its corresponding centroid provided the best fit to the data (see Table 9).

Table 4: GAMM - LDA (Embeddings)

Parametric coefficients				
Variable	Estimate	Std. Error	t	p
Intercept	-1.14	.03	-38.02	<.0001
LDA-Embedding	-0.05	.02	-3.40	.0007
ExpNo2	-0.21	.03	-7.18	<.0001
ExpNo3	-0.14	.03	-4.58	<.0001
Smooth terms				
Variable	edf	Ref.df	F	p
Frequency*Family Size	13.66	16.38	148.40	<.0001
Word Length	6.45	7.47	156.46	<.0001
TrialNo, Subjects	863.22	2518.00	7.95	<.0001
R^2 = .418				
AIC = 12406.96				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ∼ te(frequency*family size) + s(word length) + LDA-Embedding + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data

Table 5: GAMM - LDA (Shift)

Parametric coefficients				
Variable	Estimate	Std. Error	*t*	*p*
Intercept	-1.14	.03	-40.66	<.0001
LDA-Shift	-0.05	.01	-4.49	<.0001
ExpNo2	-0.21	.03	-7.17	<.0001
ExpNo3	-0.14	.03	-4.55	<.0001
Smooth terms				
Variable	edf	Ref.df	F	*p*
Frequency*Family Size	13.52	16.24	150.19	<.0001
Word Length	6.47	7.49	157.36	<.0001
TrialNo, Subjects	863.21	2518.00	7.95	<.0001
R^2 = .418				
AIC = 12398.76				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ∼ te(frequency*family size) + s(word length) + LDA-Shift + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data

Table 6: GAMM - Correlation (Derived Word and Centroid)

Parametric coefficients				
Variable	Estimate	Std. Error	*t*	*p*
Intercept	-1.18	.03	-44.79	<.0001
ExpNo2	-0.21	.03	-7.20	<.0001
ExpNo3	-0.14	.03	-4.61	<.0001
Smooth terms				
Variable	edf	Ref.df	F	*p*
Frequency*Family Size	13.41	16.06	146.12	<.0001
Word Length	6.47	7.49	158.03	<.0001
Correlation-dev.centroid	6.13	7.30	13.24	<.0001
TrialNo, Subjects	864.68	2518.00	7.97	<.0001
R^2 = .419				
AIC = 12325.08				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ∼ te(frequency*family size) + word length + s(correlation-dev.centroid) + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data

Table 7: GAMM - Correlation (Target Word and Predicted Word; FRACSS)

Parametric coefficients				
Variable	Estimate	Std. Error	*t*	*p*
Intercept	-1.18	.03	-45.02	<.0001
ExpNo2	-0.22	.03	-7.29	<.0001
ExpNo3	-0.14	.03	-4.69	<.0001
Smooth terms				
Variable	edf	Ref.df	F	*p*
Frequency*Family Size	14.00	16.73	147.08	<.0001
Word Length	6.38	7.40	129.71	<.0001
Correlation-FRACSS	5.52	6.63	12.35	<.0001
TrialNo, Subjects	862.79	2518.00	7.97	<.0001
R^2 = .419				
AIC = 12340.77				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ~ te(frequency*family size) + word length + s(correlation-FRACSS) + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data

Table 8: GAMM - Correlation (Derived Word and Shift)

Parametric coefficients				
Variable	Estimate	Std. Error	*t*	*p*
Intercept	-1.19	.03	-45.18	<.0001
ExpNo2	-0.21	.03	-7.08	<.0001
ExpNo3	-0.14	.03	-4.48	<.0001
Smooth terms				
Variable	edf	Ref.df	F	*p*
Frequency*Family Size	13.57	16.73	134.90	<.0001
Word Length	6.49	7.40	153.90	<.0001
Correlation-dev.shift	2.88	6.63	10.47	<.0001
TrialNo, Subjects	863.84	2518.00	7.96	<.0001
R^2 = .419				
AIC = 12381.06				

Note. Word frequency and root family size were log transformed. The model syntax is inverse RT ~ te(frequency*family size) + word length + s(correlation-dev.shift) + experiment + s(trial number, subjects, bs= 'fs', m=1). Inverse RT = -1000/RT; a negative sign is used to make the interpretability more like traditional RT data

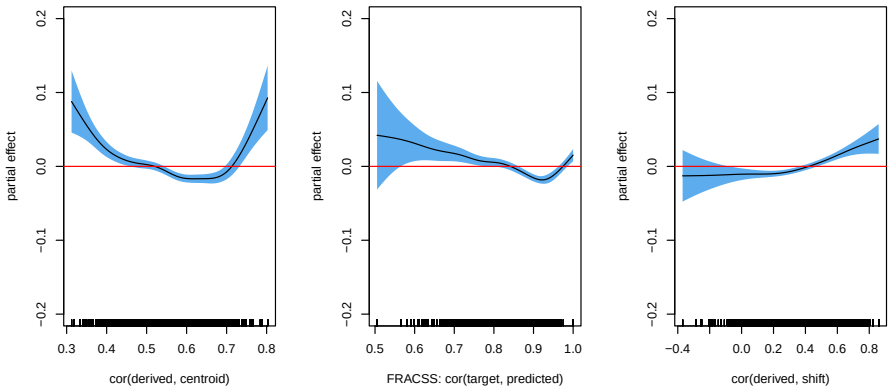

Figure 4: *Note.* Partial effect of each correlation measure on RT. Black lines on the x-axis of each plot represent the data

Table 9: AIC Comparison

Model	Δ AIC
LDA-Derived	8.85
LDA-Shift	17.05
Correlation-Dev.centroid	90.73
Correlation-FRACSS	75.04
Correlation-Shift.centroid	34.75

Note. Difference in AIC scores are calculated by subtracting the AIC of each model that included a predictor of interest from the AIC of the baseline model. Greater values in the change of AIC score indicate a better fit to the data.

3.4 General Discussion

In the present study, we first sought to explore whether the embeddings of Malay complex words cluster in semantic space by prefix, using t-SNE. To make any meaningful interpretation, we focus on words containing a prefix of a sizeable count. Namely, these are words that contained the prefix *beR-, meN-, teR-, peN-, peR-,* or *ke-*. A key take-away from the t-SNE analysis is the observation that there is considerable variation between complex words that arise from prefixes in Malay, a stark contrast to polysemous particles in German that are semantically ambiguous (Stupak and Baayen, 2022). More broadly, we have demonstrated that inquiries into the semantic transparency of words can be meaningfully explored using *t*-SNE.

A detailed linguistic analysis of how these clusters emerge as depicted in the t-SNE plot is beyond the scope of the present study. As an aside, however, we inspected whether clustering occurs by word category. Across languages, many previous studies have suggested that nouns and verbs differ in their semantics and distributional properties (for a review, see Vigliocco et al., 2011). We extracted word category information from the msTenTen corpus, a Malay web corpus, on SketchEngine for each word in our dataset. A large majority of the words in our dataset were assigned as nouns. Some clustering was observed for verbs, although they appear largely nested in the cluster of nouns (see Figure S1 in Supplementary Materials). No obvious clusters were observed for adjectives. Relatedly, the LDA model correctly classified nouns and verbs to a large extent, but not for adjectives. The overall accuracy with which the LDA model predicts word category is much lower (.77) than the LDA models that predict a word's prefix (.93 using word embeddings and .88 using shift vectors). Most crucially, the clustering reported for prefixed words in the present study is not confounded with word category.

A next step would be to identify a set of features in which these clusters embody. Such a study will shed light on the kind of semantics that could be extracted from word embeddings. We leave more fine-grained analyses for future work. For an initial exploratory study on prefixed words, it makes most sense to first examine whether or not complex words cluster meaningfully by their prefixes. The results of the present study make it sufficiently clear that word embeddings capture a rich knowledge of semantic information that could be used to discriminate between complex words.

A secondary objective of the present study was to calculate several measures of semantic transparency and evaluated their impact on lexical decision latencies in a series of GAM models. All five measures significantly predicted decision latencies above and beyond classical predictors of lexical processing and root family size. This finding complements prior work in German (Stupak and Baayen, 2022) and Mandarin (Shen and Baayen, 2022) in which distinct clusters in semantic space were formed for words that share a derivational affix and show a strong semantic association with their base words. In the present study, the correlation between a word and its corresponding centroid emerged as the best predictor of decision latencies.

Unlike word embeddings, the vector of a centroid does not represent a real word, but rather, an average embedding that could be understood as the prototypical meaning of the prefix. In our case, words are considered related if they share a prefix. We showed that the speed with which a word is processed is predictable in part from the strength of the semantic relationship between a particular word and all other related words, such that faster responses were elicited for a word that shares a strong semantic relationship with its centroid. Such evidence provide empirical support for

the potential of realizational morphology as a theory of lexical processing, even for Malay, a language that is morphologically rich in derivation and contains minimal inflection. Although subword embeddings are considered, morphemes are not explicitly represented. Realizational morphology is typically discussed in studies concerning inflected forms. Only derived forms were analyzed in this study. These results further support a previous exploratory study on Indonesian morphology (Denistia and Baayen, 2022) that used the Discriminative Lexicon Model (DLM; Baayen et al., 2019. The DLM, grounded in word and paradigm morphology, is a computational theory of the mental lexicon in which the whole word is taken to be the most basic unit. The DLM consists of simple linear mappings between high-dimensional representations of form and meaning (for details on the implementation of the DLM, see JudiLing tutorial; Heitmeier et al., 2024). In that study of Indonesian morphology, the DLM accurately discriminated between words containing prefixes *pe-* and *pen-*, often associated with similar meanings, even though the model was not explicitly informed about exponents and stems. On our to-do list is a closer inspection of the potential role of centroids in the DLM. Preliminary evidence from the DLM trained on words in the present study's dataset revealed that there is indeed a strong correspondence between the centroid and the linear mappings of a word's form and its meaning (see Figure 5), further justifying a promising role of centroids in word recognition.

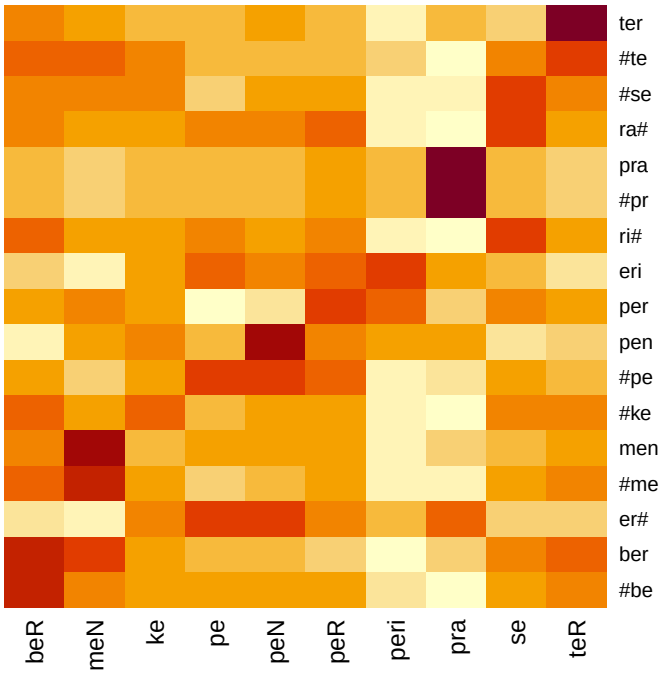

Figure 5: *Note.* Correlation heatmap of prefix centroid embeddings and a subset of the linear comprehension mapping F that maps the embeddings of a word form (trigrams) to its meaning (FastText) in the DLM. The set of trigrams presented correspond to at least one of the prefixes. Darker shades, compared to lighter shades, indicate a stronger correlation between the embeddings of the centroid and comprehension mapping F. The strongest correlations are present for the trigrams that correspond to the prefix, indicating that it is the prefixal trigrams that contribute most to realizing the meaning of the centroid

Furthermore, each of the three models that included a correlation measure provided a better fit to the data than either of the two models that included a measure derived from the LDA, providing support for analogical patterns in contrast to strict classification. These findings resonate with the results of several studies by Westbury and colleagues. In Westbury (2023), both human animacy judgments and word embedding models are shown to produce good approximations of animacy ratings on the basis of family resemblance rather than distinct category memberships. For instance, words related to human beings such as *professors* received an animacy rating of .60 by human participants and .55 by word embedding models, providing more support for the idea of similarity compared to binary classifications even for a concept as basic as animacy. Westbury and

Hollis (2019a) and Westbury and Hollis (2019b) successfully demonstrated that the centroid, in their case, the mean vector of words in a particular word category can be used to assess category membership such as nouns, verbs, and adjectives. Words of a particular category are highly correlated with their centroid. The authors noted that their findings regarding centroids extend to many semantic properties. In the present study, we establish that the centroid can be used to assess semantic transparency of a prefix in Malay.

Between the two LDA-based measures, the model that included the classification scores derived from the shift vectors provided a better fit to the data than the model that included the classification scores derived from just the word embeddings. The shift vectors represent a snapshot of the transformation in meaning of the derived form from its base. Recent work on English has shown that the semantics of pluralization varies by semantic class, even though such differences are not marked morphologically (Shafaei-Bajestan et al., 2024). For instance, the nature of the change in semantic space from singular to plural differs between words that describe a person or an animal. In that study, using shift vectors, distinct clusters in a t-SNE were observed for a large set of WordNet supersenses that include broad semantic categories for nouns. Our findings lend support to prior work in that such movements through semantic space, up to the point that the meaning of a derived form is realized, account for additional information that is meaningful for word recognition.

4 Conclusion

The present study reports an exploratory analysis of the semantic geometry of Malay word embeddings in high dimensional space. Techniques used in machine learning were employed in the visualization of word embeddings for ease of interpretability and in the computation of embedding-based measures of semantic transparency. We observed distinct clusters of complex words varied by their prefix class. In addition, we provide evidence that each embedding-based measure significantly predicts lexical decision latencies. In particular, the model that included the correlation between each derived word and their centroid appears to be the best fit to the data. That is, the similarity of each word to the prototypical meaning of its prefix appears to be the best way to characterize semantic transparency. These measures are available for a large set of Malay words and can be downloaded in the latest version of the Malay Lexicon Project 3: `https://osf.io/dhyzb/?view_only=e05e71b31cb54daf94a55f46f9cc82da`

References

Auch, L., Gagné, C. L., and Spalding, T. L. (2020). Conceptualizing semantic transparency: A systematic analysis of semantic transparency measures in english compound words. *Methods in Psychology*, 3:100030.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019(1):4895891.

Baayen, R. H., Fasiolo, M., Wood, S., and Chuang, Y.-Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon*, 17(2):178–212.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Chee, Q. W. and Yap, M. J. (2022). Are there task-specific effects in morphological processing? examining semantic transparency effects in semantic categorisation and lexical decision. *Quarterly Journal of Experimental Psychology*, 75(11):2073–2086.

Chuang, Y.-Y., Fon, J., Papakyritsis, I., and Baayen, H. (2021). Analyzing phonetic data with generalized additive mixed models. In *Manual of clinical phonetics*, pages 108–138. Routledge.

Denistia, K. and Baayen, R. H. (2022). The morphology of indonesian: Data and quantitative modeling. In *The Routledge handbook of Asian linguistics*, pages 605–634. Routledge.

Denistia, K., Shafaei-Bajestan, E., and Baayen, R. H. (2022). Exploring semantic differences between the indonesian prefixes pe-and pen-using a vector space model. *Corpus Linguistics and Linguistic Theory*, 18(3):573–598.

Diependaele, K., Sandra, D., and Grainger, J. (2009). Semantic transparency and masked morphological priming: The case of prefixed words. *Memory & cognition*, 37(6):895–908.

Feldman, L. B., Barac-Cikoja, D., and Kostić, A. (2002). Semantic aspects of morphological processing: Transparency effects in serbian. *Memory & Cognition*, 30(4):629–636.

Gonnerman, L. M., Seidenberg, M. S., and Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: evidence for a distributed connectionist approach to morphology. *Journal of experimental psychology: General*, 136(2):323.

Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2024). The discriminative lexicon: Theory and implementation in the julia package judiling.

Hockett, C. F. (1954). Two models of grammatical description. *Word*, 10(2-3):210–234.

Jared, D., Jouravlev, O., and Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(3):422.

Krijthe, J. H. and Van der Maaten, L. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation. r package version 0.13. *Computer Software*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Libben, G., Gibson, M., Yoon, Y. B., and Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and language*, 84(1):50–64.

Lund, K. and Burgess, C. (1996). Hyperspace analogue to language (hal): A general model semantic representation. *Brain and cognition*, 30(3):5–5.

Marelli, M. and Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485.

Maziyah Mohamed, M. and Jared, D. (2023). The distributional properties of prefixes influence lexical decision latencies: Evidence from malay. *The Mental Lexicon*, 18(2):218–264.

Maziyah Mohamed, M. and Jared, D. (2025). Malay lexicon project 3: The impact of orthographic–semantic consistency on lexical decision latencies. *Quarterly Journal of Experimental Psychology*, 78(1):22–47.

Maziyah Mohamed, M., Yap, M. J., Chee, Q. W., and Jared, D. (2023). Malay lexicon project 2: Morphology in malay word recognition. *Memory & Cognition*, 51(3):647–665.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. Number 47. University of Illinois press.

Plaut, D. C. and Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4-5):445–485.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rueckl, J. G. and Seidenberg, M. S. (2011). Computational modeling and the neural bases of reading and reading disorders. In *How Children Learn to Read*, pages 101–133. Psychology press.

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R. H. (2024). The pluralization palette: unveiling semantic clusters in english nominal pluralization through distributional semantics. *Morphology*, 34(4):369–413.

Shen, T. and Baayen, R. H. (2022). Adjective–noun compounds in mandarin: a study on productivity. *Corpus Linguistics and Linguistic Theory*, 18(3):543–572.

Stupak, I. V. and Baayen, R. H. (2022). An inquiry into the semantic transparency and productivity of german particle verbs and derivational affixation. *The Mental Lexicon*, 17(3):422–457.

Taft, M. and Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6):638–647.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., and Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, 35(3):407–426.

Westbury, C. (2023). Why are human animacy judgments continuous rather than categorical? a computational modeling approach. *Frontiers in Psychology*, 14:1145289.

Westbury, C. and Hollis, G. (2019a). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, 51:1371–1398.

Westbury, C. and Hollis, G. (2019b). Wriggly, squiffy, lummox, and boobs: What makes some words funny? *Journal of Experimental Psychology: General*, 148(1):97.

Westbury, C., Yang, M., and Anderson, K. (2024). The principal components of meaning, revisited. *Psychonomic Bulletin & Review*, pages 1–23.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman & Hall/CRC, New York.