Exploring the Semantic Space of Second Language Learners

Trisha Godara[1, 4], Rui He[2], Wolfram Hinzen[2, 3], Yan Cong[4]
[1]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA
[2]Department of Translation & Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[3]Intitut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[4]School of Languages and Cultures, Purdue University, West Lafayette, Indiana, USA
tgodara@purdue.edu, rui.he@upf.edu, wolfram.hinzen@upf.edu, cong4@purdue.edu

Abstract

While the semantic space has been examined as a way to computationally represent language meaning-grammar interface, minimal research has been done comparing the semantic spaces of first and second language learners. We investigated the semantic space of university-level students learning French by extracting semantic features from narrative text over various time points from a 21-month period. After using machine learning models to classify native speakers' semantic features from second language learners', we used interpretability techniques to identify the most informative features per model. Through this, we discovered a variety of embedding similarity features to be decisive in language learning. We compared both groups to determine how the features differed per group and if there was any change over time. The findings demonstrated that the second language learners on average had higher semantic similarity scores than the native speakers at the token level. The similarity decreased over time but did not reach native-level values. Similarly, average surprisal was higher in the second language learner group, which steadily decreased over the course of the data collection period. These results provide insight into personalized education with more precise and effective computational indices tracking learners' progress.

1 Introduction

Distributional semantic models computationally capture language meaning through indices such as semantic similarity (Baroni and Lenci, 2010; Lenci et al., 2022), and the resulting semantic space can quantitatively and more precisely characterize and inform learners' interlanguage systems development (Bexte et al., 2022; Cong, 2024). Through this lens, this study aims to understand second language (L2) learners' language development by comparing their semantic spaces with native speakers' (L1). Using machine learning models, we analyzed L2 learners' semantic space features over time, utilizing the French dataset from the Languages and Social Networks Abroad Project (LANGSNAP) (Mitchell et al., 2017), which contains data from proficient French L2 learners over a 21-month period, including time living abroad in France.

Semantic space measures vary from study to study and have been widely used in many settings, such as clinical populations (He et al., 2024b). However, as far as our knowledge goes, there is no systematic investigation about these measures in the L2 population. Therefore, we used a comprehensive set of measures, focusing on implementing semantic similarity-related measures, to better understand L2 development trajectories. Our approach and findings might shed light on real-world practical applications, such as native language identification, teaching materials design, and personalized learning with these more precise and quantifiable measures provided by natural language processing (Bexte et al., 2022; Chen and Pan, 2022).

We asked two questions: Which features are the most important to determine native speakers from L2 speakers according to predictive machine language models? How do these features change over time for L2 learners? To approach these questions, we designed two experiments. After extracting 132 semantic space measures for each participant's data, we used predictive models to classify the semantic space data as either L1 or L2, and we performed SHapley Additive exPlanations (ShAP) analysis to determine which of the features contributed most to each model's classification results. We then conducted a time-based analysis on each of the top identified features, observing how they changed over time for the L2 learners on average. Both models identified a largely non-overlapping set of top contributing features. We found that overall token-level semantic similarity for L2 learners was higher than native speakers on average, and this similarity decreased as the participants spent more

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 332–346
March 26, 2026 ©2026 Association for Computational Linguistics

time learning, growing closer to native speakers' semantic similarity levels. However, we did not notice a true converging point, where the average L2 learner becomes indistinguishable from a native speaker, in terms of these features.

2 Related Work

2.1 Semantic Space Modeling in Second Language Acquisition

Semantic representations derived from distributional and neural language models provide a quantitative framework for modeling meaning in language use. High-dimensional embeddings allow semantic similarity, coherence, and dispersion to be measured across words, sentences, and larger discourse units, offering insights into how speakers organize and navigate semantic space (Ke et al., 2025; Goldstein et al., 2024). While these approaches have been extensively applied to native-speaker language and to clinical populations (Corcoran et al., 2018; Bedi et al., 2015; He et al., 2024a,b), their application to second language acquisition (SLA) remains comparatively limited.

In SLA, semantic development is closely tied to vocabulary growth, conceptual restructuring, and increasing efficiency in mapping form to meaning. Learner language is often characterized by reduced semantic specificity, increased redundancy, and less stable discourse coherence, particularly in oral production. Embedding-based semantic measures provide a way to operationalize these properties beyond surface-level metrics, such as lexical diversity or syntactic complexity, capturing how L2 learners structure meaning across linguistic scales (Bexte et al., 2022; Cong, 2024, 2025b).

Recent work has begun to explore semantic similarity and discourse-level coherence in learner language, suggesting that embedding-based features can distinguish proficiency levels and task conditions. However, most studies adopt semantic space measures originally developed and validated on native-speaker data, without directly assessing their behavior when applied to L2 learners. This raises important questions about how such measures reflect second language development rather than native-language distributions.

2.2 Cross-Linguistic Semantics and Learner Language

A substantial body of work in computational linguistics has investigated how semantic representations vary across languages and linguistic units (Beinborn and Choenni, 2020; Chersoni et al., 2019, 2021; Vulić et al., 2020; Lewis et al., 2023). These studies show that while semantic spaces often share global structural properties, local semantic relationships are sensitive to lexicalization patterns, typology, and model architecture. For L2 learners, this is particularly relevant, as their semantic representations are shaped by interactions between the target language and their first language (L1).

Research on bilingual and cross-lingual embeddings further highlights how semantic spaces reflect both shared conceptual structure and language-specific encoding (Gouws and Søgaard, 2015; Vivas et al., 2020). In learner language, these interactions may manifest as transfer effects, overgeneralization, or reliance on high-frequency, semantically broad lexical items (Cong, 2025a,b). Embedding-based analyses provide a principled way to examine these phenomena by quantifying similarity patterns at the token, sentence, and discourse levels.

Importantly, semantic space analyses allow distinctions between local semantic behavior, such as similarity between consecutive lexical items, and more global discourse-level organization. These distinctions align naturally with theories of L2 development, which propose that learners acquire lexical meanings earlier than they master discourse-level coherence and information structuring.

2.3 Semantic Measures and Cognitive Organization in L2 Development

Beyond descriptive modeling, semantic space measures have been linked to broader cognitive properties of language production. Prior work has used semantic similarity, surprisal, and perplexity to characterize differences in semantic organization across populations and tasks (Silva et al., 2021; He et al., 2024b; Palominos et al., 2024). Analyses of the geometry of semantic space, such as dispersion and clustering of sentence embeddings, have been shown to reflect how speakers navigate meaning over extended discourse.

Although much of this work has focused on neurotypical versus neuroatypical populations, the underlying methods are highly relevant to SLA. L2 development involves continuous reorganization of semantic networks as learners gain exposure and experience to the target language, particularly in immersive contexts such as study abroad. Longitudinal learner corpora therefore provide a valuable opportunity to examine how semantic space features evolve over time and how they differentiate native and non-native language use.

At the same time, prior work has highlighted that semantic space features can exhibit substantial variability across samples and tasks, especially when compared to more stable acoustic or prosodic features (Cokal et al., 2025). For L2 learners, whose linguistic systems are inherently dynamic, this variability indicates the need for careful feature selection and interpretation.

2.4 Current Work

Building on this line of research, the present study applies a comprehensive set of semantic space measures, previously used in native-speaker and clinical language analysis, to longitudinal L2 narrative data. By focusing on content-controlled oral narratives from the LANGSNAP dataset, we aim to evaluate whether semantic space features can reliably distinguish L1 and L2 speech and to identify which aspects of semantic organization are most informative for this distinction. We predicted that over time, measures should show L1-like patterns. In particular, similarity indices should be higher in L2 group than L1, due to more repetition, but decrease over time; predictability indices should be higher in L2 (more unpredictable) and decrease over time due to improvement in content flow and syntax. Results were mostly as predicted, with some inconsistencies. Overall, our findings and approach contribute to a growing effort to adapt and validate computational semantic measures for the study of second language development.

3 Dataset

The LANGSNAP corpus's data was collected between 2011 and 2013 to observe L2 French development over the course of 21 months, before, during, and after a 9-month study or work abroad (Mitchell et al., 2017). There were 39 total participants. Ten were native speakers, and the remaining 29 were L2 learners with English as their first or dominant language. One participant who was identified as an English and French bilingual speaker was determined to be an outlier and excluded from our study. Out of the native speakers, three were male, and seven were female. Out of the 28 L2 learners, 2 were male and 26 were female. All participants were in their third year of a four-year university program. The L2 learners were to spend nine months abroad in France, either attending college, teaching English, or doing an internship. All L2 participants had at least six years of experience learning French, with most having been learning for more (mean = 10.21 years, standard deviation = 2.45 years).

The LANGSNAP team collected data from the L2 participants at six time points: before the study abroad, three times during the study abroad (each about three months apart), and twice after the study abroad. They had the L2 participants complete a variety of tasks at each time point, including an oral interview, a writing task, and a picture-based narrative oral monologue.

For the purposes of our study, we chose to use the narrative task, as this would be the best oral task to control for content since the story was guided with images. There were three topics: the cat story (conducted during pretest and abroad visit 3), the sisters story (conducted during abroad visit 1 and post-test 1), and the brothers story (conducted during abroad visit 2 and post-test 2).

4 Experiments

4.1 Preprocessing and Feature Extraction

Each participant's utterances was extracted from their narrative interview and concatenated into a single paragraph. We passed this input to SpaCy's French model to segment it into sentences and to tag the parts of speech. We were only interested in the noun, adjective, and verb tokens to ensure the semantic feature calculations were based on the most meaningful tokens. Then we used three types of models to extract the embeddings. FastText was used for context-free, token-level embeddings; Bidirectional Encoder Representations from Transformers (BERT) was used for contextual, token-level embeddings; Sentence-BERT (SBERT) was used for contextual, sentence-level embeddings. These models were pretrained on French data. For perplexity and surprisal calculations, we used a generative text model, Mistral, as well as the Next Sentence Prediction (NSP) capability of BERT to calculate an additional surprisal metric. The BERT [1], SBERT [2], and Mistral [3] models we used are hosted on Huggingface, as part of the `transformers` library (Wolf et al., 2019).

We built our measurements on validated works

[1] dbmdz/bert-base-french-europeana-cased

[2] dangvantuan/sentence-camembert-base

[3] mistralai/Mistral-7B-Instruct-v0.1

(He et al., 2024a,b; Cokal et al., 2025; Palominos et al., 2024), focusing on semantic similarity measures, and we grouped these measures into three categories: statistical descriptors, dynamic descriptors, and graph measures. We also extracted probabilistic metrics, including perplexity and surprisal. In addition to calculating similarity of adjacent units, we also calculated the similarity of the units in reference to their static and cumulative centroids (Xu et al., 2021). The static centroid is the unchanging averaged embedding, capturing the overall text's topic, while the cumulative centroid incorporates the previous embeddings and captures the change in topic over time. As such, we calculated 132 measures in total per entry.

4.1.1 Feature Definitions

Statistical Descriptors The mean cosine similarity (MeanK) between a linguistic unit and the next unit located at an inter-word distance of k (Corcoran et al., 2018) is referred to as the k-order similarity. First-order (K1) similarity represents the average similarity between each embedding and its immediate successor, whereas second-order (K2) similarity represents the average similarity between each embedding and the embedding that follows it with exactly one intervening unit. We also computed the global-level mean similarity by taking the average cosine similarity between all unit pairs. Other statistical measures, like maximum and minimum, amplitude, variance, skewness, and excess kurtosis (tailedness in comparison to normal distribution), were computed to present a more comprehensive picture of the distribution of similarities.

Dynamic Descriptors These features were used to observe how semantic similarity behaves over time. We calculated mean crossing rate (MCR) as the number of times the semantic similarity crosses the mean value. Slope sign changes (SSC) measures how often the similarity time series changes signs (see Equation 1, where $\mathbb{I}$ is the indicator function). Wave length (WL) is defined as the average absolute change between successive semantic similarity scores (see Equation 2). Approximate entropy (ApEn) as defined in Pincus (1991) was used to represent the regularity of semantic similarity patterns in the time series (see Equation 3), computed via the `antropy` Python package. The autocorrelation function (ACF; see Equation 4), the correlation of a lagged similarity series with itself, was also calculated, along with its zero crossing rate (AcfZcr).

$$\text{SSC} = \frac{1}{N-2}\sum_{i=2}^{N-1}\mathbb{I}\left[(x_i - x_{i-1})(x_i - x_{i+1}) > 0\right] \tag{1}$$

$$\text{WL} = \frac{1}{N-1}\sum_{i=1}^{N-1}|x_{i+1} - x_i| \tag{2}$$

$$\text{ApEn}(m, r) = \Phi^m(r) - \Phi^{m+1}(r) \tag{3}$$

$$\text{ACF} = \frac{\sum_{i=1}^{N-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

Graph Measures These measures are extracted from the graph that is created when considering embeddings as points in space. Each embedding is a node, and an edge is created between two nodes if their cosine similarity is greater than the threshold value (in our case, 0.8). Closeness centrality quantifies a node's efficiency in spreading information by calculating the reciprocal of the average shortest path distance between it and all other reachable nodes in the network; this value was averaged across the entire network. The clustering coefficient, which was also averaged, refers to the fraction of a node's neighbors that are also connected to one another. This metric reflects the strength of semantic associations since words with deeper meaning-based links are more likely to form interconnected clusters (He et al., 2024a).

Probabilistic Measures Perplexity refers to a model's uncertainty about its predictions (Jurafsky and Martin, 2025) while surprisal measures the unexpectedness of a data point, given the previous inputs (Hale, 2001). Both of these measures were averaged over the whole input.

4.2 Experiment 1: L1 vs. L2 Classification and Feature Importance

4.2.1 Classification

We decided to use two different types of models to see how they performed in comparison to each other: support vector classifier (SVC) and decision tree classifier (DTC). Our machine learning models selection strategies were inspired by Cawley and Talbot (2010). Each model would determine whether the sample was L2 (i.e. positive class) or not (i.e. L1, negative class). We used a 70% training and 30% testing split.

After running the classifiers once, it became apparent that the imbalanced dataset was causing overfitting issues. As mentioned earlier, there were 28 participants for the L2 class, but only 10 for the L1 class. To combat this, we looked into a variety of methods to balance the dataset. One method was upweighting the minority class using the `class_weight='balanced'` hyperparameter. We also tried using synthetic sample generation techniques, like Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), to upsample, but these results were only marginally better, and it was difficult to verify the accuracy of the generated data. Therefore, we decided to continue with just using the `class_weight` hyperparameter.

Before moving onto the feature importance step, we decided to reduce the number of features to both make the final results more interpretable but also to reduce possible redundancy in the data. To do so, we first calculated the Variance Inflation Factor (VIF) of each feature to detect multicollinearity (Thompson et al., 2017). The resulting values were significantly larger than expected, so instead of using solely VIF, we used those values in tangent with the correlation coefficients of each pair of features. For each highly correlated pair (i.e. coefficient > 0.8), we dropped the feature with the higher VIF. This led us to dropping 41 of the features, leaving us with 91 features.[4] With this reduced dataset, DTC performance improved, suggesting the extra data was acting as clutter to this model rather than being helpful. On the other hand, SVC performance worsened.

We did also do further hyperparameter tuning on the full feature dataset, using five-fold cross validation with the `roc_auc` scoring method. Running the cross validation on the SVC model was very slow, meaning we were unable to use the more exhaustive grid search and opted to use randomized search instead. For DTC, we were able to use grid search to find the optimized set of hyperparameters, though the values that the grid search found did not end up producing results that were better than the previous configuration with the reduced feature dataset, so we proceeded with our manually selected hyperparameters.[5]

4.2.2 Feature Importance

With our selection of 91 features, we were able to move to SHAP analysis for both the SVC and DTC models. SHAP analysis, a machine learning model interpretability method, was originally adapted from the game theory idea of Shapley values (Ponce-Bobadilla et al., 2024). These values indicate how much a feature contributed to the model's prediction.

The DTC model only had six features that contributed, so to balance the total set of features we were looking at between both models, we took only the top six SHAP-identified features of the SVC model as well. To determine whether the differences between the L1 and L2 data were significant, we ran ANOVA tests on the total top 12 features, resulting in a total of 6 significant features.

4.3 Experiment 2: Time-Based Analysis

This second experiment built off of the first one, using the six significant features found in the last step. We wanted to compare both the overall average differences between the two groups, L1 and L2, and the progression of L2 learners for each feature. For the overall averages, we simply took the average of all the native speakers' data for that feature and the average of all the L2 learners across all the time points to compare the two.

For the time-based comparison, we used the six timepoints used by the LANGSNAP team as our points of reference. For each feature, we took all the L2 learners' data and averaged it at each time point and plotted it on a line graph. To compare the trajectory with the native speakers' we plotted the L1 averaged data at the six timepoints as well. The L1 data was not collected over the same period of time as the L2 data, so instead we aligned the L1 data with the L2 data points by narrative content to ensure valid comparisons. As noted previously, there were three total narrative topics, which were repeated for the second half of the L2 learners' time, giving us six tasks' data to work with. On the other hand, L1 speakers were given the three narrative tasks as well, but the tasks were conducted only once each, so we matched by content: the pretest and abroad 3 timepoints equated to the cat story, abroad 1 and post-test 1 to the sisters story, and abroad 2 and post-test 2 to the brothers story.

[4] This feature list can be found in Appendix A.

[5] We provide the main classification pipeline script with model hyperparameters in this GitHub repository: https://github.com/trishagodara/l2-semantic-space

Positive Class (L2)		
	SVC	**DTC**
Precision	0.94	0.98
Recall	0.65	0.92
F1-Score	0.77	0.92
Negative Class (L1)		
	SVC	**DTC**
Precision	0.22	0.60
Recall	0.71	0.86
F1-Score	0.33	0.71

Table 1: Classification report for both models - positive testing class size: 52, negative testing class size: 7

5 Results

5.1 Model Performance

Between the support vector classifier and the decision tree classifier, the decision tree model was better at classifying the difference between L1 and L2. Here we will discuss the final results, which were found using the reduced feature dataset with each model's best-performing hyperparameters. SVC's overall accuracy was 66% with an area under curve (AUC) of 0.72. DTC's overall accuracy was 92% with an AUC of 0.89. The final classification report is available in Table 1.

5.2 Top SHAP Features

Figures 1 and 2 show the ranked feature importance for each model. As mentioned earlier, six features from the SVC model and six features from the DTC model were considered when compiling the list of top contributing features, out of which six total ended up being significant. These features are as follows: average surprisal, sentence-level skewness (from SBERT, in relation to the static centroid), meanK1 (from BERT embeddings), average clustering coefficient (from BERT), mean crossing rate (from BERT, in relation to cumulative centroid), and variance (from fastText, in relation to static centroid).

5.3 Time-Based Analysis

We looked into the specifics of each of the significant features, comparing the L2 data points with the L1 data points and also observing the trajectory of the feature values over time for the L2 learners. Our findings are summarized in Table 2. The graphs for each feature can be found in Appendix B, and some example narrative excerpts are provided in Appendix C for comparison.

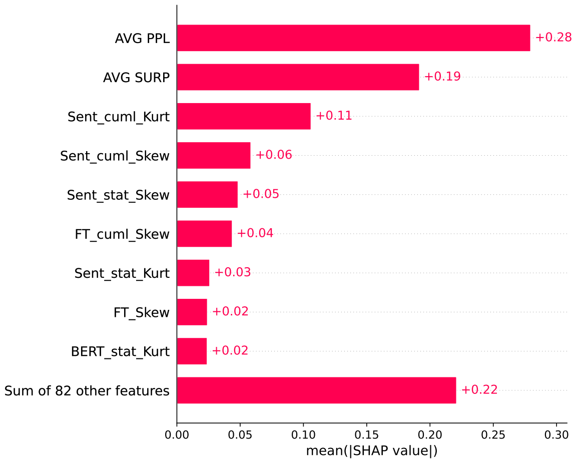

Figure 1: SVC's SHAP values per feature, ranked. Notation: PPL = perplexity, SURP = surprisal, Sent = from SBERT embeddings, cuml = cumulative centroid, stat = static centroid, Kurt = excess kurtosis, FT = fastText

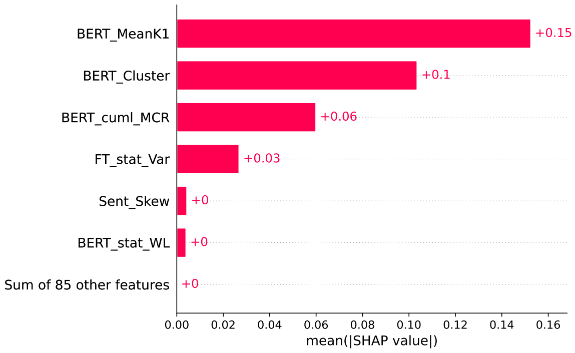

Figure 2: DTC's SHAP values per feature, ranked. (Although the plot shows them as "+0," the BERT static WL's SHAP value is 0.005, and sentence skewness's SHAP value is 0.0045). Notation: Cluster = clustering coefficient, cuml = cumulative centroid, FT = fastText, stat = static centroid, Var = variance, Sent = from SBERT embeddings, WL = wave length

Considering the relatedness of some of the features, we created broad categories to encompass the six measures of interest in a way that similar findings were grouped together (Table 2). Surprisal is in a category of its own, as the sole predictability-related feature. Vocabulary includes variance and mean while content flow includes mean crossing rate and sentence-level skewness. Semantic structure includes the average clustering coefficients.

6 Discussion

While the first experiment informed the identification of the features that would be used in the second experiment, the interpretation of those features only became apparent in the second experiment with the

Category	Meaning	Group Prediction	As Predicted?	Time Prediction	As Predicted?
Surprisal	Describes how unexpected the input is; irregular speech will result in larger values	L2 values would be greater than L1 values	Yes	Values would decrease over time	Yes
Vocabulary	Describes repetitive usage of (limited) vocabulary (higher semantic similarity)	L2 values would be greater than L1 values	Yes	Values would decrease over time	No*
Content Flow	Describes how on-topic, logically relevant the input is	L2 values would be less than L1 values	Yes	Values would increase over time	Yes
Semantic Structure	Describes strength of semantic connections	L2 values would be less than L1 values	No	Values would increase over time	No

Table 2: Overview of variables design, predictions, and results. Notation: *: overall there was no downward trend but when comparing each content-matched pair, there was a decrease.

comparisons between the two groups. Therefore, we found the time-based analysis more insightful in terms of understanding L2 learners' development trajectory.

Surprisal As previous studies have also found (Cong, 2025a), non-native speakers' speech resulted in higher surprisal scores overall. This may be an indication that the non-native speakers did not have the same grasp over syntactic structure as the native speakers did, resulting in a higher unpredictability rating from the large language model. We did notice a steady decrease in the average L2 surprisal as time went on, with the post-test 2 data point almost reaching the L1 level. This suggests that the L2 learners' speech patterns, on average, gradually began to evolve into a more native speaker-like pattern the more time passed.

Vocabulary This category contains the statistical semantic similarity-related features at token level. Since semantic similarity often corresponds to words having similar meanings (Kolb, 2009), we interpreted this group to represent speakers' vocabulary: a diverse vocabulary would have lower similarity scores at the token level, while a more limited vocabulary would have higher values due to repeated or similar words.

In regards to mean semantic similarity, the L2 group exhibited higher scores on average than the L1 group. Previous studies have shown similar results where L2 learners typically produce speech with higher inter-word mean similarity (Cong, 2024). Looking at the variance of the two groups, we found all the values to be rather small (values between 0 and 0.014). The L2 average variances were greater than the L1 averages, suggesting that there was a larger spread in similarity scores across the board for the L2 learners. This could possibly be related to the variability in vocabulary levels between learners, but there may be other interpretations for this metric, especially considering how close to 0 all the values were. We leave systematic investigation of alternative factors and explanations for future research. Future study could implement the proposed pipeline to a larger datasets, validating the role of vocabulary.

Conceptually, we predicted a decrease of this similarity over time, as the L2 learners grew more proficient. However, this is not exactly what we observed. We speculate that this category is more related to the content itself rather than time-based proficiency because overall there was no consistent downward trend, like we saw with the surprisal scores. When comparing each content-matched timepoint, however, the second time the L2 learners were presented with the same topic they had done months ago, their values decreased somewhat, moving in the direction of the L1 group's value for that topic.

Content Flow While skewness could have been relevant in the vocabulary category as well, considering it was reflecting the similarity distribution at the sentence level, and specifically in comparison to the static centroid, we thought it would be more appropriate in this section where we are discussing the overall content flow – how on-topic is the input (Bexte et al., 2022; Cong, 2024)? We found that both groups' distributions were slightly skewed left (values between 0 and -0.35) with the L1 group's values being more negative (i.e. more skewed). This indicates that while both L1 and L2 had fairly balanced semantic similarity across the entirety of the input, on average, the L1 group's data was considered slightly more similar. However, unlike the vocabulary category, here we are looking at the *sentence* level of similarity, which is why we interpreted this as the overall cohesion of the text, rather than word usage. Over time, the L2 group had limited change in the distribution's skewness, staying relatively the same from start to end, though at the abroad 2 timepoint, there was a dip where the L2 group average nearly met the L1 group average. This may have had something to do with the content of that particular narrative, but we did not observe such behavior at the post-test 2 timepoint when that same narrative prompt was administered again. As for the MCR, by our predictions, a native speaker would have better flow and transitions in their speech, and therefore their similarity score would cross the mean more often – in a similar vein as our reasoning for the higher sentence-level similarity scores. This is what we observed, with the L1 group's MCR consistently being higher than the L2 group. We did see an overall increase in the L2 learners' MCR over the course of the program, though there again seemed to be a connection with content as well, for it was not a totally steady upward trajectory. This could mean the L2 learners' content flow improved over time, thereby increasing their MCR.

Semantic Structure Lastly we have the semantic structure category, which consists of the clustering feature. Contrary to what we had expected, the L2 average clustering coefficients were higher than the L1 averages. As clustering coefficients usually signify the strength of semantic connections, we had thought the native speakers' data would be higher in this regard, as they have more of a mastery over the language. Therefore, these results were rather confounding. Additionally, over time, the L2 average clustering coefficients *decreased* for the most part. While this meant the L2 group moved toward the L1 group values, from the meaning of semantic connection strength, it would indicate that these connections grew weaker the more time passed. However, these results are similar to our findings about the overall semantic similarity of the two groups. It could be interpreted that the clustering coefficient is related to the inter-token similarity of the text, hence the higher number of clusters for the L2 group. In this case, it would not be that the connections grew weaker but that as the L2 learners' semantic similarity decreased and more lexical and syntactic variability was introduced, the denseness of the clusters also decreased.

There may be more to what semantic connections really mean, and the clinical population might inform this direction (Cokal et al., 2025). The precise interpretations for the L2 population are outside of our current scope; more experiments are needed to address this in L2 development contexts.

Real-World Implications and Future work In educational contexts, our studies can support and enhance automatic assessment (Ramesh and Sanampudi, 2022; Chen and Pan, 2022) in the future. We have outlined which metrics are most interpretable and effective to track learners' proficiency stages. We intend the proposed pipeline to serve as an initial framework for using semantic measures in benchmarking L2 development and conducting general language assessment, thereby illustrating a promising route for integrating computational linguistics with language learning.

Also, future research could incorporate a broader range of models to further investigate and leverage these semantic features. Given that the interpretation of such features may be model-dependent, architectures employing distinct embedding paradigms may yield differing instantiations or realizations of these semantic properties. Nevertheless, the current pipeline has been evaluated on multiple representative model families, suggesting that the approach should be generalizable to future models that share comparable architectures and embedding paradigms.

7 Conclusion

In this study, we used defined semantic space features to classify native speakers and second language learners using machine learning models.

From this we were able to extract key features that, upon further analysis, provided some insight into the differences between the two groups. Semantic similarity has been known to be a distinguishing factor between the two, and this study corroborates that, focusing on specific measurements within that. From our longitudinal analysis, we found that L2 learners' abilities did develop over time, which was reflected by lower surprisal scores, a moderate decrease in semantic similarity measures (hence more diverse and expanded vocabulary as L2 interlanguage systems evolve), and improved content flow. Not all of the analyzed features showed this same level of improvement, but those remain open points of discussion regarding second language learners' capabilities and development trajectories.

Limitations

The dataset was inherently imbalanced, as there was limited native speakers' data due to fewer L1 participants and only one set of narratives to extract features from. We considered the imbalance and addressed it with widely used techniques. We acknowledge that it is possible there was still some overfitting by the models, and a larger scale, full validation is out of the scope. In a similar vein, since the dataset was so small, it is possible that minor changes in the data, including choosing different threshold values or a different training/testing split, could have affected the features outcome. To ensure the soundness of the filtering pipeline, further testing can be done in the future.

As mentioned in the Experiments section, the SVC model was very computationally expensive and time-consuming, which limited the amount of hyperparameters that could be tuned. While we attempted to run the cross validation with the various kernel types that the SVC library provides, with the large number of features, the only one that was able to run in the alloted time was the 'linear' kernel. Therefore, it is possible there was a better set of hyperparameters that could have been used for the SVC, but we were unable to explore all of them in this study. We leave it for future research.

It should be noted that the perplexity and surprisal values were calculated using the `Mistral 7B Instruct-v0.1` model, which was primarily pretrained on English data. We acknowledge that more accurate values may have been achieved with a better suited multilingual model or even French-specific model.

8 Acknowledgments

We thank the anonymous reviewers for the constructive reviews, which helped improve the paper. This project is supported by the Computation and Linguistic Meaning Lab at Purdue University.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1).

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123, Seattle, Washington. Association for Computational Linguistics.

Gavin C. Cawley and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Huimei Chen and Jie Pan. 2022. Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on chinese college students' english writing. *Asian-Pacific Journal of Second and Foreign Language Education*, 7(1).

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Emmanuele Chersoni, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2019. A structured distributional model of sentence meaning and processing.

Derya Cokal, Martin Villalba, Rui He, Claudio Flores Palominos, Annkathrin Böke, Philipp Homan, Klaus von Heusinger, Joseph Kambeitz, and Wolfram Hinzen. 2025. What is the retest reliability of computationally extractable speech and language markers?

Yan Cong. 2024. AI language models: An opportunity to enhance language learning. *Informatics*, 11(3).

Yan Cong. 2025a. Demystifying large language models in second language development research. *Computer Speech & Language*, 89:101700.

Yan Cong. 2025b. Second language learning of degree expressions: A computational approach. *Natural Language Processing*, 31(5):1187–1209.

Cheryl M. Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C. Javitt, Carrie E. Bearden, and Guillermo A. Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, and 3 others. 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1).

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Rui He, Maria Francisca Alonso-Sánchez, Jorge Sepulcre, Lena Palaniyappan, and Wolfram Hinzen. 2024a. Changes in the structure of spontaneous speech predict the disruption of hierarchical brain organization in first-episode psychosis. *Human Brain Mapping*, 45(14):e70030.

Rui He, Claudio Palominos, Han Zhang, Maria Francisca Alonso-Sánchez, Lena Palaniyappan, and Wolfram Hinzen. 2024b. Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, 333:115752.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2025. Exploring the frontiers of llms in psychological applications: A comprehensive review. *Artificial Intelligence Review*, 58(10):305.

Peter Kolb. 2009. Experiments on the difference between semantic similarity and relatedness. In *NODALIDA 2009 Conference Proceedings*, pages 81–88.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, 56(4):1269–1313.

Molly Lewis, Aoife Cahill, Nitin Madnani, and James Evans. 2023. Local similarity and global variability characterize the semantic space of human languages. *Proceedings of the National Academy of Sciences*, 120(51):e2300986120.

Rosamond Mitchell, Nicole Tracy-Ventura, and Kevin McManus. 2017. *Anglophone students abroad: Identity, social relationships, and language learning*. Routledge.

Claudio Palominos, Rui He, Karla Fröhlich, Rieke Roxanne Mülfarth, Svenja Seuffert, Iris E. Sommer, Philipp Homan, Tilo Kircher, Frederike Stein, and Wolfram Hinzen. 2024. Approximating the semantic space: Word embedding techniques in psychiatric speech analysis. *Schizophrenia*, 10(1).

Steven M. Pincus. 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301.

Ana Victoria Ponce-Bobadilla, Vanessa Schmitt, Corinna S. Maier, Sven Mensing, and Sven Stodtmann. 2024. Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17(11).

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Angelica Silva, Roberto Limongi, Michael MacKinley, and Lena Palaniyappan. 2021. Small words that matter: Linguistic style and conceptual disorganization in untreated first-episode schizophrenia. *Schizophrenia Bulletin Open*, 2(1):sgab010.

Christopher Glen Thompson, Rae Seon Kim, Ariel M. Aloe, and Betsy Jane Becker. 2017. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, 39(2):81–90.

Leticia Vivas, Maria Montefinese, Marianna Bolognesi, and Jorge Vivas. 2020. Core features: Measures and characterization for different languages. *Cognitive Processing*, 21(4):651–667.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Weizhe Xu, Jake Portanova, Ayesha Chander, Dror Ben-Zeev, and Trevor Cohen. 2021. The centroid cannot hold: Comparing sequential and global estimates of coherence as indicators of formal thought disorder. *AMIA Annual Symposium Proceedings*, 2020:1315–1324.

A Features Used in Classification

Table 3 lists the features that remained after reducing the feature set.

B Significant Features Over Time

Figures 3 to 8 show the feature trajectories of the L1 and L2 groups over the six time points for easy visual comparison.

C Example Narrative Excerpts

Provided below are brief excerpts[6] from various participants' narrative interviews through the lens of a few different features for illustration purposes.

C.1 Average Surprisal Comparisons

Overall highest average surprisal score: 10.85, L2 participant (#109) at the pretest time point.

- Tous les matins étaient pareils. Natalie -euh Natalie lit -euh et dans Natalie lit un livre dans son lit.
  "Every morning was the same. Natalie -uh Natalie reads -uh and in Natalie reads a book in her bed." [7]

Lowest average surprisal score for the same story: 6.69, L1 participant (#135).

- Tous les matins étaient pareils pour la petite fille Nathalie. Nathalie tous les matins se réveillait.
  "Every morning was the same for little Nathalie. Every morning, Nathalie woke up."

C.2 MeanK1 Comparisons

Highest mean similarity value: 0.4678, L2 participant (#104) at abroad 2 time point.

- Les frères en deux mille le frère aîné de Jacques était est allé étudier à l'étranger.
  "The brothers in 2000, Jacques' older brother had gone to study abroad."

Lowest mean similarity value for the same story: 0.4037, L1 participant (#130).

- Euh c'est l'histoire des frères. En deux mille le frère aîné de Jacques est allé étudier à l'étranger.
  "Uh, it's the story of the brothers. In 2000, Jacques' older brother went to study abroad."

C.3 Clustering Coefficient Comparisons

Highest average clustering coefficient value: 0.5719, L2 participant (#121) at pretest time point.

- Euh tous les matins matins étaient pareils pour -euh Nathalie et Pompon -euh. Nathalie regardait un lit.
  "Well, every morning morning was the same for -uh Nathalie and Pompon -uh. Nathalie was looking at a bed."

Lowest average clustering coefficient value for the same story: 0.3578, L1 participant (#135)

- Tous les matins étaient pareils pour la petite fille Nathalie. Nathalie tous les matins se réveillait.
  "Every morning was the same for little Nathalie. Every morning, Nathalie woke up."

[6] For the full passages, please see the LANGSNAP database: `https://talkbank.org/slabank/access/French/LANGSNAP.html`.

[7] All translations in this appendix were done using Google Translate, for demonstration purposes.

Feature	**Model(s)**
MeanK1	fastText, BERT, SBERT
MeanK2	fastText, SBERT
Mean Crossing Rate (MCR)	fastText, BERT, SBERT
Slope Sign Changes (SSC)	fastText, BERT, SBERT
Wave Length (WL)	BERT, SBERT
Variance	fastText, BERT, SBERT
Peak	fastText, BERT, SBERT
Valley	fastText, BERT, SBERT
Skewness	fastText, BERT, SBERT
Excess Kurtosis	fastText, BERT
Approximate Entropy (ApEn)	fastText, SBERT
Autocorrelation Function (Acf)	fastText, SBERT
Acf Zero Crossing Rate (AcfZcr)	fastText, BERT, SBERT
MCR (static centroid)	fastText, SBERT
SSC (static centroid)	fastText, SBERT
Variance (static centroid)	fastText, BERT
Peak (static centroid)	fastText
Valley (static centroid)	fastText, SBERT
Skewness (static centroid)	fastText, SBERT
Excess Kurtosis (static centroid)	fastText, BERT, SBERT
ApEn (static centroid)	fastText, SBERT
Acf (static centroid)	fastText, SBERT
AcfZcr (static centroid)	fastText, SBERT
WL (static centroid)	BERT
Amplitude (static centroid)	BERT, SBERT
MeanK1 (cumulative centroid)	fastText
MCR (cumulative centroid)	fastText, BERT, SBERT
SSC (cumulative centroid)	fastText, BERT, SBERT
Variance (cumulative centroid)	SBERT
Peak (cumulative centroid)	fastText, BERT, SBERT
Valley (cumulative centroid)	fastText, SBERT
Amplitude (cumulative centroid)	fastText, SBERT
Skewness (cumulative centroid)	fastText, BERT, SBERT
Excess Kurtosis (cumulative centroid)	fastText, SBERT
Acf (cumulative centroid)	fastText
AcfZcr (cumulative centroid)	fastText, BERT, SBERT
ApEn (cumulative centroid)	SBERT
Closeness Centrality	fastText, BERT, SBERT
Clustering Coefficient	fastText, BERT, SBERT
Next Sentence Prediction-Based Perplexity	BERT
Average Perplexity	Mistral
Average Surprisal	Mistral

Table 3: The 91 measures used for classification

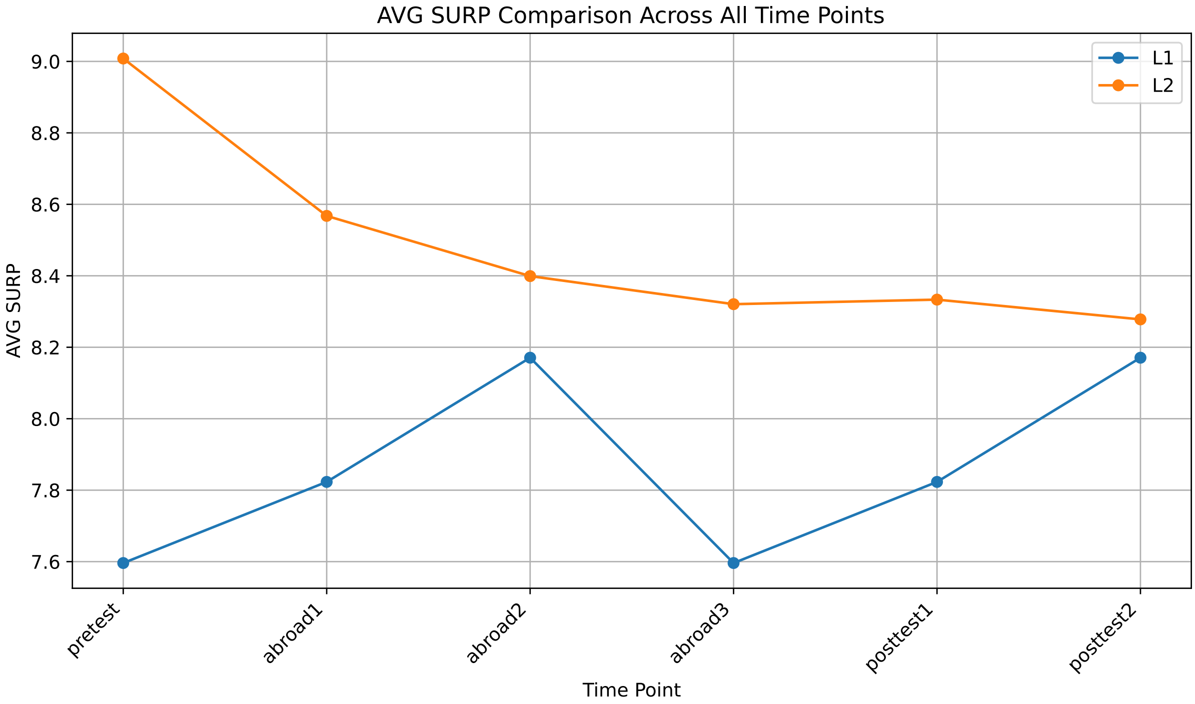

Figure 3: Longitudinal analysis of average surprisal

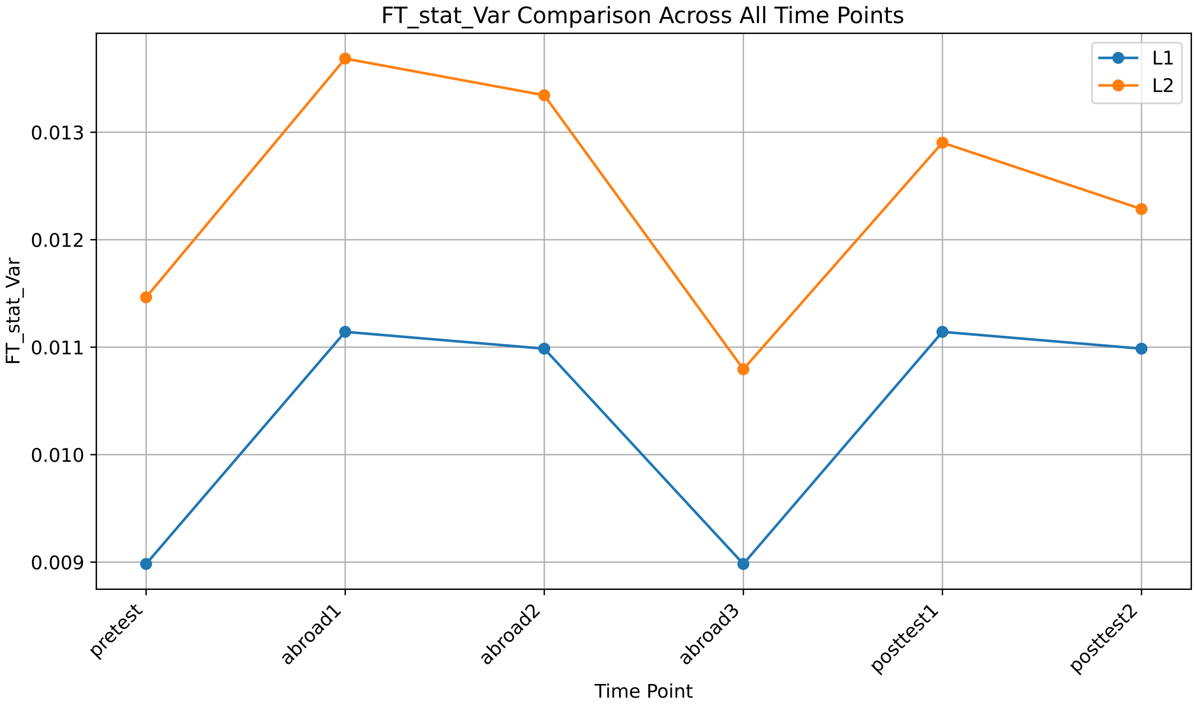

Figure 4: Longitudinal analysis of variation in reference to static centroid (from fastText)

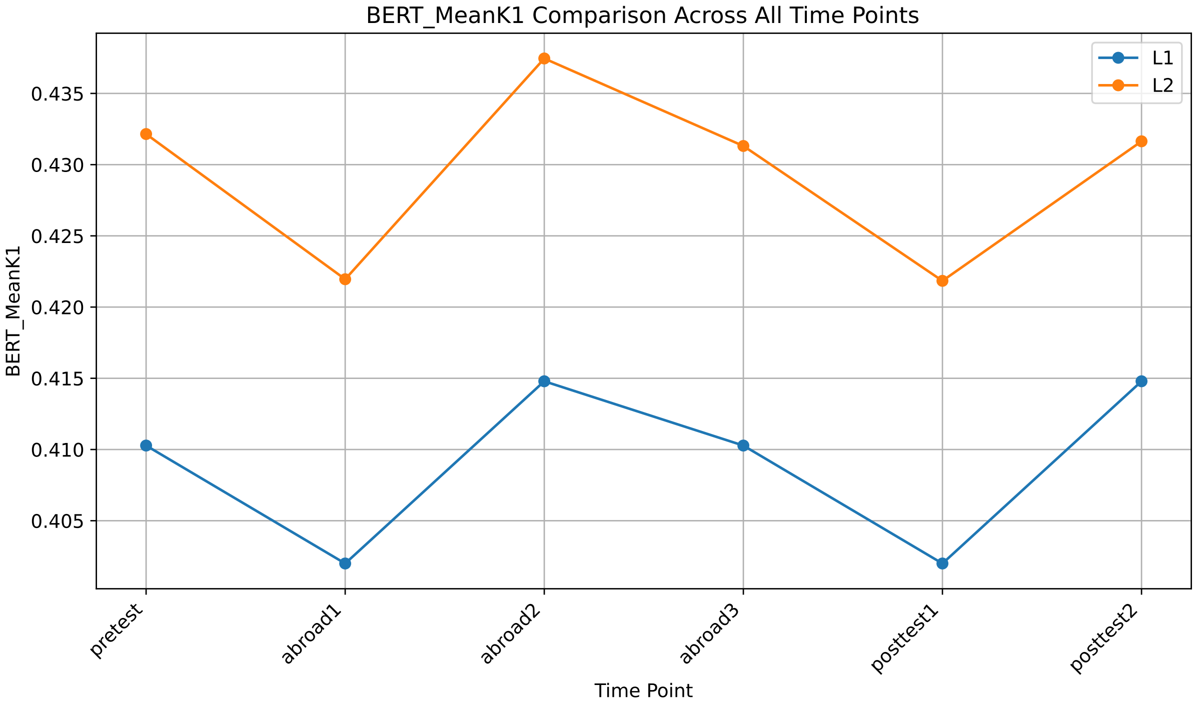

Figure 5: Longitudinal analysis of meanK1 similarity (from BERT)

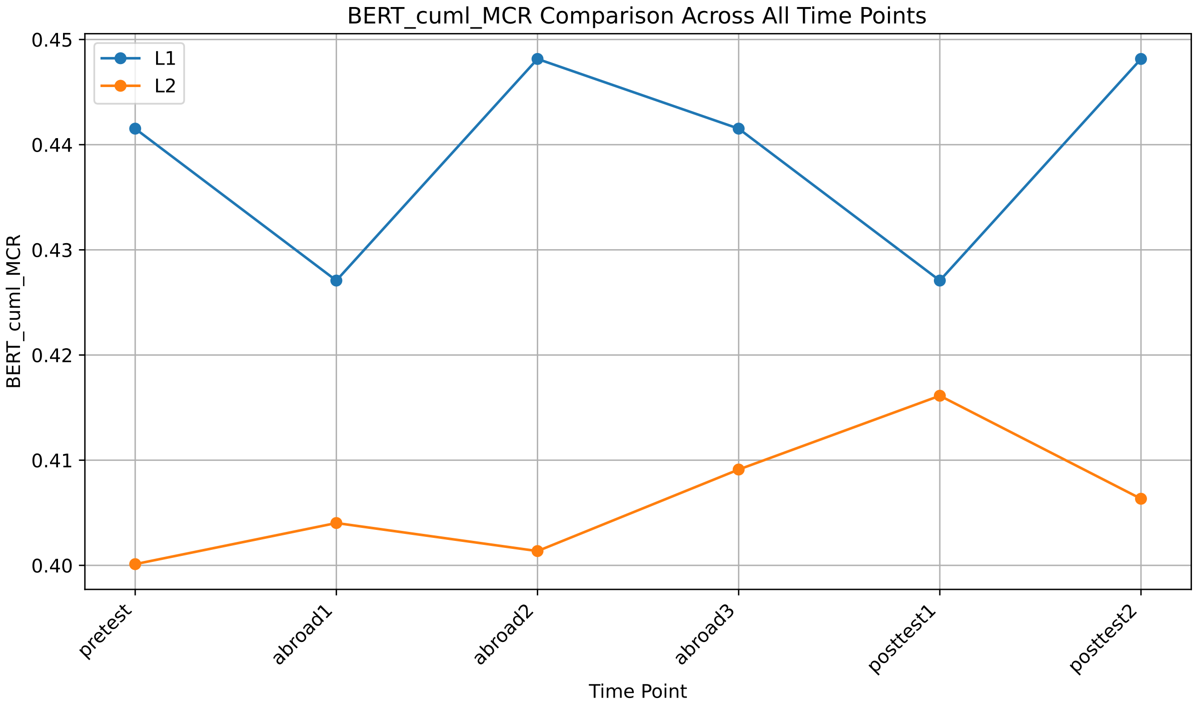

Figure 6: Longitudinal analysis of mean crossing rate in reference to cumulative centroid (from BERT)

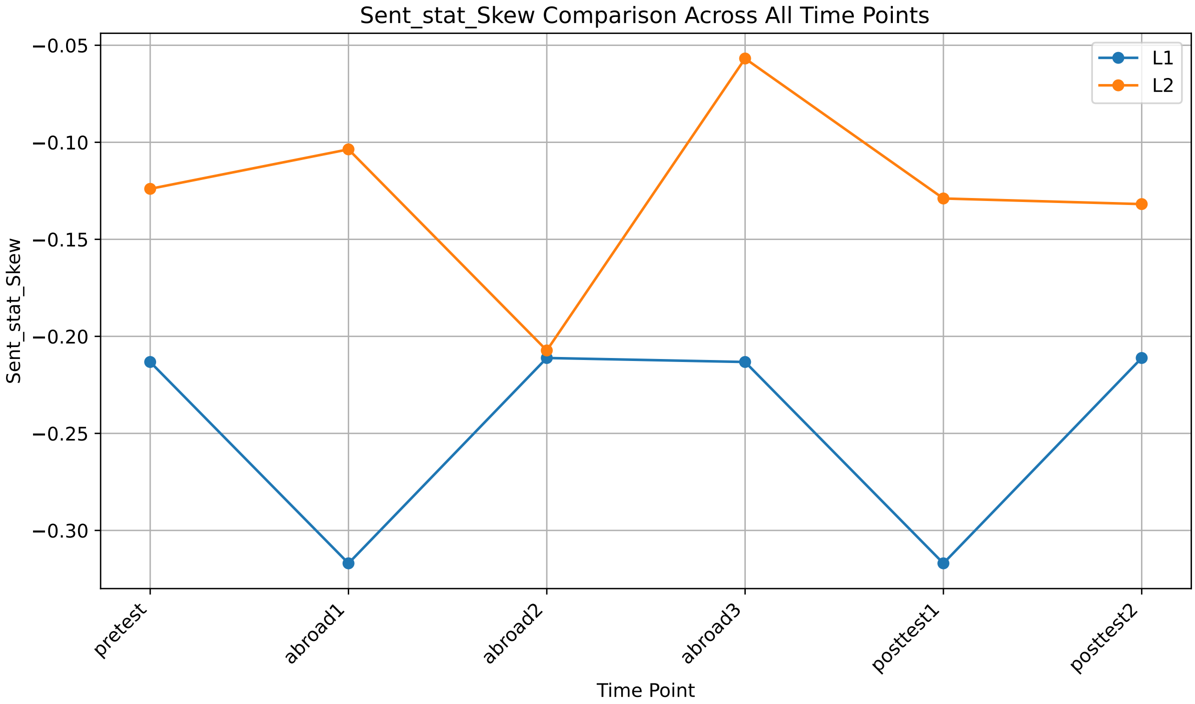

Figure 7: Longitudinal analysis of skewness in reference to static centroid (from SBERT)

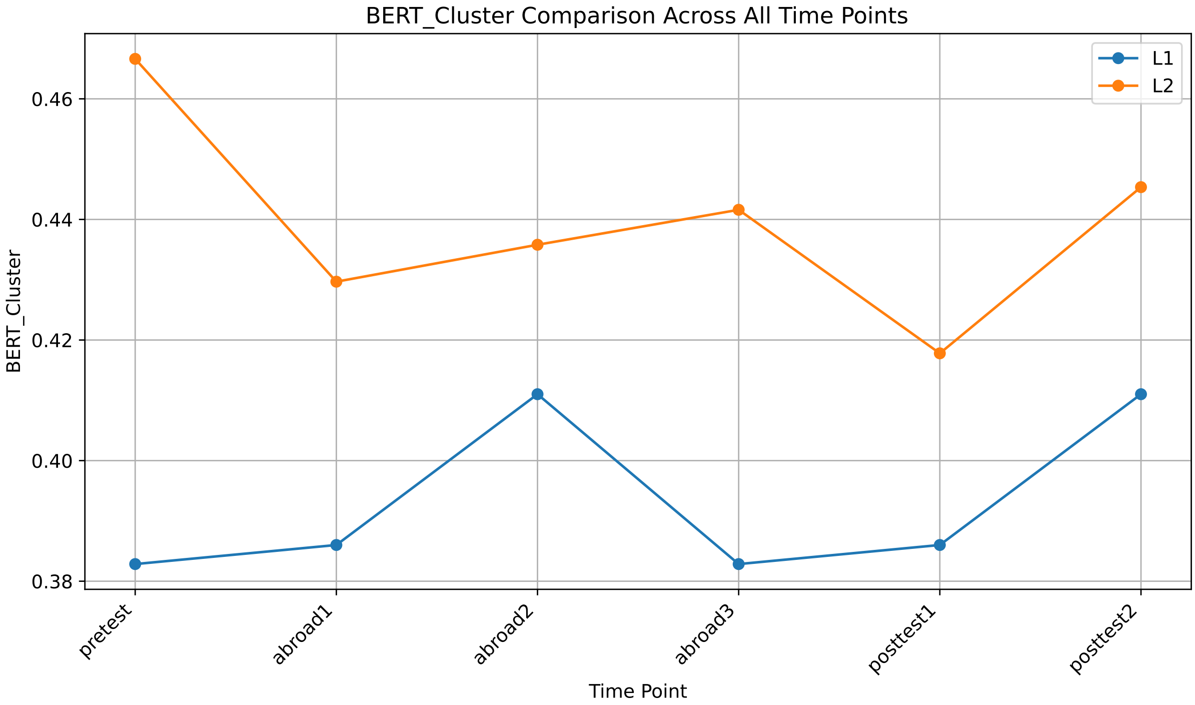

Figure 8: Longitudinal analysis of average clustering coefficients (from BERT)