# Investigating Semantic Clusters in the Verbal Fluency Task Across Native (L1) and Non-Native (L2) Hebrew Speakers

**Maya Zadok, Tomer Cohen**
Faculty of Data Sciences and Decisions, Technion, IIT

## Abstract

This study examines verbal fluency task performance among native (L1) and non-native (L2) Hebrew speakers, with a focus on predicting language nativeness and analyzing generated semantic content. We explored semantic topics, clustering, and connections between performance metrics to uncover potential cognitive patterns. Topic modeling was employed to identify underlying semantic themes in the verbal fluency data. Three machine learning models (Logistic Regression, SVM, and Random Forest) were used to classify nativeness. The SVM model achieved the highest AUC-ROC, indicating superior class separation with a highest score of 0.609, while the Random Forest model obtained the best F1 score, highlighting its predictive strength with a high score of 0.848. Results suggest that immersion in Hebrew by non-native speakers may diminish performance differences. This research is serving as a foundation for future research into the cognitive mechanisms influencing verbal fluency in multilingual contexts.

Git: `https://github.com/Specril/verbal-fluency-project`

## 1 Introduction

The verbal fluency task is a widely used tool for cognitive assessment. It is simple, short, and effective for diagnosis of various diseases, such as Alzheimer's disease (Mueller et al., 2015; Wright et al., 2023), Parkinson (Pettit et al., 2013), hyperactivity disorder (ADHD) (Andreou & Trott, 2013). In the task, participants are asked to name as many items under time constraint and meet certain rules, and avoid repetitions. In the semantic version of the task, also known as "categorical fluency", where the rule is that items have to be from a certain category (e.g., animals, fruits and vegetables, or vehicles). In the phonemic version, or the "letter fluency", the rule is that the word starts with a certain letter (e.g., F, A, or S). This task relies on numerous cognitive abilities, including semantic memory retrieval, executive function and working memory (Li et al., 2017; Miyake et al., 2000; Shao et al., 2014). However, this test also relies highly on verbal ability, specifically lexical access ability, which is described as "the ability to retrieve the grammatical representations and sounds forms of words from the mental lexicon" (Shao et al., 2014). Therefore, difficulties in word retrieval during the task may result from limited vocabulary, different lingual associations, or difficulties in retrieval of words in a second language.

Natural Language Processing (NLP) techniques have been employed to enable a more detailed analysis of word sequences produced during verbal fluency tasks. These methods include clustering words that are closely related to "semantic clusters": groups of semantically related words produced together. Effective cluster use boosts performance, while difficulties in switching between clusters reduces fluency (Troyer et al., 1997). Therefore, clustering enables and extracting measures that were proved to distinguish between clinical populations (Lindsay et al., 2021; Mueller et al., 2015; Paula et al., 2018; Troyer et al., 1998).

Learning about the topics that arise in the verbal fluency task may provide insights about the natural semantic associations of speakers in a certain language and the extent of which these associations are preserved in L2 speakers. By examining these associations, we can gain a deeper understanding of universal associations common for human beings and the unique relations between words in a particular language. This may have clinical applications, since the topics of people with cognitive decline and of healthy controls may be different. Thus, considering the use of non-native languages in assessments is crucial as it may affect the accuracy of diagnostic evaluations and the interpretation of cognitive abilities.

Moreover, the verbal fluency task offers insights into cognitive functions related to language processing and proficiency, and therefore, it has potential applications beyond medical diagnosis. For example, García-Castro et al. (2022) explored the impact of verbal fluency on vocabulary acquisition in both first language (L1) and second language (L2) contexts among university students. Their findings suggest that verbal fluency capacity can significantly enhance vocabulary learning (García-Castro, 2022). Luo et al. (2010) investigated the performance of English monolingual speakers and two groups of bilingual speakers with varying English vocabulary sizes on a fluency task. They measured the total number of responses, first response times (RT), subsequent mean RTs, and analyzed the time course of the retrieval processes. Though their findings revealed no significant differences among the three groups in the category fluency task, in the letter fluency task, bilingual speakers with a higher English vocabulary produced more responses compared to the other groups. Additionally, both bilingual groups exhibited longer subsequent RTs than the monolingual speakers. These findings suggest that bilingual speakers may exhibit better executive function (Luo et al., 2010).

Lehtinen et al., (2023), (Lehtinen et al., 2023) learned about language attrition and second language acquisition. In language attrition, proficiency in a person's first language (L1) declines due to immersion in a second language (L2) environment, often leading to reduced fluency and increased errors in L1. We aim to explore the cognitive processes behind responses in the verbal fluency task, specifically, the conflict between the improved executive function abilities of L2 speakers and their limited proficiency in the language of assessment. Therefore, this study has two objectives:

- Predicting whether the verbal fluency task was taken in the participants' native language.
- Learning about different topics that arise in the verbal fluency task, and testing whether these topics are different among L1 vs L2 participants.

A classifier based on the verbal fluency task for distinguishing between native (L1) and non-native (L2) speakers could be beneficial, because unlike simply asking individuals about their language background, this method would be less susceptible to social desirability bias or inaccurate self-assessment. Furthermore, such a classifier could be developed to make more granular distinctions, not only identifying whether a person is native in the tested language but also potentially inferring their mother tongue. This approach could have applications in areas such as education, where it could inform tailored language instruction; in cognitive science, where it could contribute to our understanding of language acquisition and processing; and in professional settings, where accurate language proficiency assessment is crucial. The non-invasive nature of verbal fluency tasks also makes this method particularly attractive for large-scale studies or quick assessments.

# 2 Methods

## 2.1 Participants

Data collected for the Israeli Registry of Alzheimer's Prevention (IRAP) (Ravona-Springer et al., 2020), a longitudinal prospective study of asymptomatic middle-aged participants with a parental history of Alzheimer's disease. Inclusion criteria for participation are: members of Maccabi Health Services, 40-65 years old, and no signs of cognitive decline in the first visit. Participants visit the Joseph Sagol Neuroscience Center at the Sheba-Tel Hashomer Medical Center every 2-4 years for comprehensive assessment. This includes physical examinations, neurocognitive assessments, and lifestyle questionnaires. Additionally, participants are

| Characteristic | L1<br>N = 224 | L2<br>N = 57 | p-value |
|---|---|---|---|
| **Age** | 55.22 (6.67) | 57.68 (6.74) | 0.016 |
| **Gender** | | | 0.8 |
| female | 129 (58%) | 34 (60%) | |
| male | 95 (42%) | 23 (40%) | |
| **Education years** | 16.45 (3.04) | 16.86 (2.94) | 0.4 |
| **Age of learning Hebrew** | 0.13 (0.50) | 10.37 (8.08) | <0.001 |
| Unknown | 0 | 3 | |
| **Number of known languages** | | | <0.001 |
| 1 | 6 (2.7%) | 0 (0%) | |
| 2 | 113 (50%) | 9 (16%) | |
| 3 | 66 (29%) | 23 (40%) | |
| 4 | 30 (13%) | 18 (32%) | |
| 5 | 9 (4.0%) | 7 (12%) | |

Table 1: Description of the Sample.

For categorical variables (gender, number of known languages), number of participants and the corresponding percentage of the total are displayed. Categorical variables were compared using Chi-squared test. If cell number of samples in cell is under 5, Fisher Exact test was used. For continuous measures the table includes mean and SD, means are compared by a 2-sided T-test.

invited for one-time visits for medical imaging, including MRI, fMRI and PET-CT.

The dataset used for the current study includes answers for the phonemic fluency task, specifically the letter Bet, of 282 participants in their first visit in the IRAP research. All participants were assessed by a physician and neuropsychologist and were deemed healthy. All participants demonstrated sufficient proficiency in Hebrew for the assessments, reside in Israel where Hebrew is the national language, and use it in their daily lives. Participants also completed a languages questionnaire, where they reported the age at which they began learning Hebrew and the number of additional languages they know- defined as their ability to understand, read, write, or speak each language- and in what age each language was learned. The questionnaire also collected information about language proficiency levels and the frequency of usage. However, 265 participants completed the language questionnaire during a later visit, meaning the verbal fluency data and language information were collected separately, potentially years apart from the first visit. As a result, we only used data from questions that were unlikely to change significantly over time. This data includes which languages the participants know and at what age they learned each one. Based on these answers, nativeness was determined: a native Hebrew Speaker (L1) is defined as someone whose first language learned was Hebrew. In our sample, 225 are native Hebrew speakers (L1), and 57 are non-native speakers (L2). Descriptive statics for these groups are provided in Table 1.

## 2.2 Procedure

The verbal fluency test was taken by a neuropsychologist as part of the WAIS-R battery (Wechsler, 1981). Participants were asked to say as many words as they can that start with the letter Bet, the Hebrew equivalent of the letter “F” in English (Kavé & Knafo-Noam, 2015), during 60 seconds. Participant were instructed to avoid repetitions and to avoid names of people or places. The neuropsychologist wrote the responses on a paper and scored them according to 3 measures: number of correct words, number of incorrect words, and number of repetitions. Then, 5 undergraduates transcribed the words by the order they were produced along with these 3 measures.

| | Characteristic | L1<br>N = 224 | L2<br>N = 57 | t (df) | p-value |
|---|---|---|---|---|---|
| Classical | **Total words** | 12.38 (3.52) | 11.75 (3.97) | 1.16 (279) | 0.248 |
| | **Number of correct words** | 11.53 (3.67) | 11.02 (3.97) | 0.93 (279) | 0.355 |
| | **Number of repetition errors** | 0.26 (0.66) | 0.21 (0.45) | 0.53 (279) | 0.601 |
| | **Number of non-repetition errors** | 0.58 (0.84) | 0.53 (0.95) | 0.46 (279) | 0.647 |
| Non-Classical | **Frequency** | 4.20 (0.36) | 4.25 (0.34) | -0.86 (279) | 0.393 |
| | **Number of clusters** | 4.54 (2.39) | 4.28 (2.44) | 0.73 (279) | 0.466 |
| | **Number of switches** | 8.07 (3.98) | 7.05 (3.93) | 1.72 (279) | 0.086 |
| | **Mean cluster size** | 3.59 (1.92) | 3.54 (1.84) | 0.17 (279) | 0.865 |
| | **Inter Semantic Proximity** | 14.56 (2.51) | 14.38 (2.49) | 0.49 (279) | 0.623 |
| | **Intra Semantic Proximity** | 17.71 (1.15) | 17.75 (1.11) | -0.22 (279) | 0.824 |

Table 2: Means of Classical and Non-Classical Measures for the Native and Non-Native Groups.

The table includes mean and standard deviation for classical and non-classical measures. Means are compared by a 2-sided t-test.

Demographic information and languages questionnaire was taken and transcribed by a research coordinator.

Analyses were conducted using Python 3.10.12 and R4.4.1.

## 2.3 Data Preprocessing

Preprocessing steps included converting the verbal fluency answers to a numeric format known as “word embeddings”. Word embeddings are numerical representations of words in a high-dimensional vector space. These vectors capture semantic relationships between words, allowing similar words to have similar vector representations. We used the model HeBert (Chriqui & Yahav, 2022), a pretrained language

model in Hebrew, trained on 3 datasets: 9.8 GB data from the Hebrew version of OSCAR (Open Super-large Crawled Aggregated coRpus), 650 MB from Hebrew Wikipedia pages, and 150 MB of data from comments on news articles.

Then, we applied PCA (Principal Component Analysis) to reduce dimensions of the embeddings. This step is important because high-dimensional data can lead to issues such as the “curse of dimensionality” (Bellman, 1966), where data points become sparse, making it difficult to identify clusters. PCA is a technique that transforms high-dimensional data into a lower-dimensional space by converting the original variables into new variables called “principal components.” These components are linear combinations of the original variables and are designed to retain as much of the data's variance as possible. Clusters were created using the k-means algorithm. Number of clusters (k) was determined after examining possible silhouette scores. Silhouette score is the ratio between the average intra-cluster distance and the average nearest-cluster distance. Generally, it measures the quality of separation to clusters, depending on the similarity of words within the same cluster, and their discrimination from words of other clusters. The goal was to find a number of clusters that effectively captures the similarities among words while avoiding excessive fragmentation. We used this method to create clusters per participant and used them to extract non-classical measures (see Section 3.4). The same method to create clusters was used over the whole sample of words generated in the verbal fluency task for topic modeling.

## 2.4 Measures

We classify measures into two categories: classical and non-classical measures. Classical measures pertain to the quantity of words and errors produced during the verbal fluency task. In contrast, non-classical measures involve characteristics of the words generated in the task that require additional resources beyond the words themselves for computation. Each measure is extracted individually for each participant.

### 2.4.1 Classical Measures

**Total words**. The total number of words spoken by the participant, including both correct words and errors.

**Number of correct words.** The count of words correctly generated by the participant.

**Number of repetition errors**. The count of words repeated by the participant more than once.

**Number of non-repetition errors.** The count of errors where words do not adhere to the rules, such as words that do not start with the letter Bet.

### 2.4.2 Non-Classical Measures

**Frequency.** Word frequencies in the Hebrew language were extracted using “wordfreq” python library (version 3.1). To account for the differences is scales between frequencies and the other non-classical measures, frequency values were converted to a Zipf scale (Brysbaert et al., 2012). Zipf scale calculates the base-10 logarithm of the number of times a word appears per billion words. For instance, a word with a Zipf value of 6 occurs once per thousand words. The average frequency of a participant is calculated as the average of all the Zipf frequencies of the words they generated in the task.

**Number of Clusters**. The count of clusters generated by k-means algorithm.

**Mean cluster size.** The average number of words in each cluster.

**Number of switches.** The count of transitions between different clusters during the task. A switch occurs when a participant says a word from one cluster, followed by a word from another cluster.

**Inter-semantic similarity**. the averaged distance between centroids.

**A.** Demographic and Classical Measures

| | Precision | Recall | F1 | AUC-ROC |
|---|---|---|---|---|
| Logistic Regression | 0.820 | 0.562 | 0.665 | 0.515 |
| SVM | 0.848 | 0.544 | 0.661 | **0.609** |
| Random Forest | 0.789 | 0.835 | **0.811** | 0.491 |

**B.** Demographic and Non-Classical Measures

| | Precision | Recall | F1 | AUC-ROC |
|---|---|---|---|---|
| Logistic Regression | 0.835 | 0.612 | 0.705 | 0.559 |
| SVM | 0.849 | 0.549 | 0.664 | **0.594** |
| Random Forest | 0.788 | 0.911 | **0.844** | 0.417 |

**C.** Demographic, Classical and Non-Classical Measures

| | Precision | Recall | F1 | AUC-ROC |
|---|---|---|---|---|
| Logistic Regression | 0.839 | 0.630 | 0.716 | 0.567 |
| SVM | 0.855 | 0.548 | 0.666 | **0.589** |
| Random Forest | 0.794 | 0.910 | **0.848** | 0.447 |

Figure 2: Model Performance Metrics for Predicting Nativeness

Performance metrics (Precision, Recall, F1-Score, and AUC-ROC) of three predictive models: Logistic Regression, SVM, and Random Forest trained on different sets of predictors: (A) demographic and classical variables, (B) demographic and non-classical variables, (C) demographic, classical and non-classical variables. Best performance of the F1 and AUC-ROC metrics is bold.

**Intra-semantic similarity**. The averaged distance between pairs of words of the same cluster

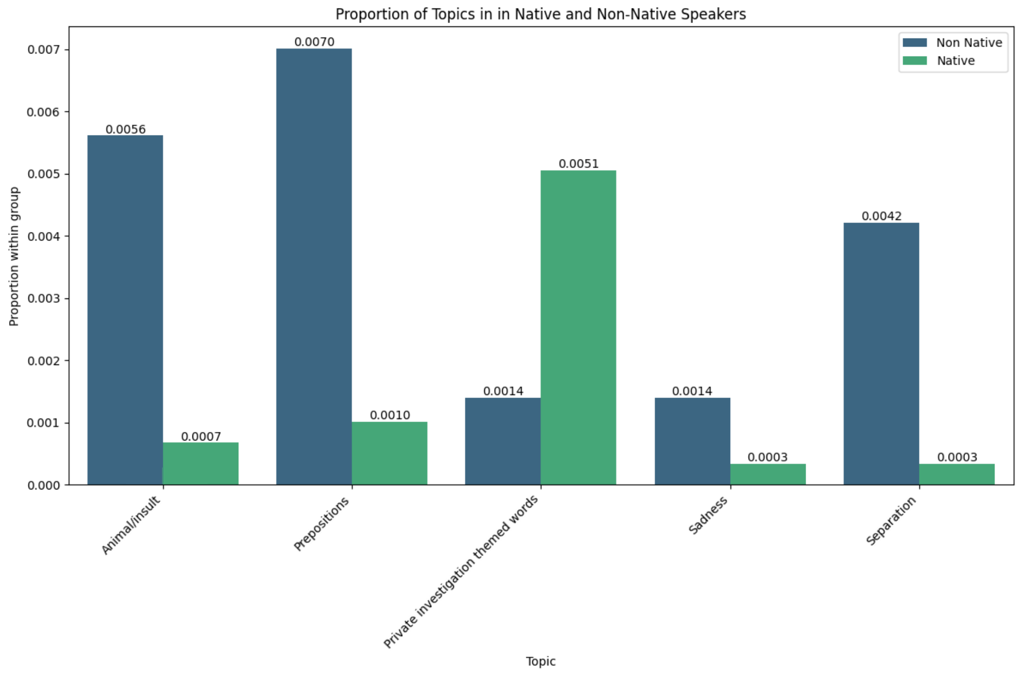


Figure 2: The 5 topics with largest difference of proportions

Proportions are calculated as the number of words from a certain topic relatively to the number of words generated by each group (Native vs Non-Native). Proportion values are presented at the top of each bar 2-sided t-tests were conducted to compare the proportions of each topic across groups. None of the comparisons is statistically significant.

# 3 Results

## 3.1 Models for predicting nativeness

**Training**. Three kinds of models were employed to predict whether a participant is a native Hebrew speaker utilizing different sets of predictors. The first model incorporated demographic and classical variables, the second included demographic and non-classical variables, and the third utilized all available variables. Demographic variables included in the models were: age, gender, and years of education. Classical and non-classical measures are specified in sections 3.4.1 and 3.4.2. Comparison of means of the classical and non-classical measures is presented in Table 2. None of the comparisons is statistically significant (α=0.05). We tried three different approaches for predictive models: Logistic Regression, Support Vector Machine (SVM), and Random Forest. For SVM, a radial basis function kernel was chosen.

Stratified K-Fold Cross-Validation with k = 5 was used to split the dataset, ensuring representative samples of each class in each of the five folds. To address class imbalance, the minority class was oversampled to match the size of the majority class in the training sets.

**Evaluation**. Precision, Recall, F1-Score, and Received Operating Characteristic Area Under the Curve (AUC-ROC), were recorded for each model. Results are presented in Figure 1.

The highest AUC-ROC score achieved with the demographic and classical measures (ROC-AUC = 0.609). SVM consistently obtained the highest AUC-ROC scores, suggesting it is the favorable model for distinguishing between native and non-native speakers. Random Forest consistently achieved the highest F1 scores. However, The F1 score should be interpreted cautiously because of our oversampling. This result may decline without sample balancing.

## 3.2 Topic modeling

Topic modeling was performed using the complete set of words generated in the verbal fluency task, comprising 3,681 words in total, of which 634 were unique. The word that appeared the largest number of times was "Bait" ("בית", meaning "home"), which appeared 253 times. It was followed by "Balon" ("בלון", meaning "balloon"), which appeared 115 times, and

"Beged" or "Bagad" (both written as "בגד", meaning "clothing item" or "betrayed", respectively), which appeared 93 times. These words were the most frequent across both the L1 and L2 groups separately.

Clusters were created using the k-means algorithm as outlined in Section 3.3, with number of clusters set to be $k = 233$. Refer to Appendix A for further details on the selection of number of clusters.

Then, we referred to Google's open API and inserted words of the same cluster to generate topic labels. The prompt is specified in Appendix B. For each topic, we calculated the number of people from each class (L1 vs. L2) that referred to this topic. To account for the imbalanced data (e.g., a larger number of native speakers), this value was divided by the total number of words that were generated by each class. The top-5 topics with the largest difference in proportions are presented in Figure 2. 2-sided t-tests were conducted to compare these proportions. None of the comparisons is statistically significant.

## 4 Discussion

This study leverages several classification algorithms to classify L1 and L2 speakers. Among the tested algorithms, best performance was achieved with SVM. Surprisingly, the non-classical measures decreased performance compared to the classical measures. We propose that the challenges in classification may arise from opposing factors that counteract each other: the high executive function abilities of L2 speakers and their limited semantic knowledge. However, to validate this theory, it is essential to address few challenges present in our study.

First, homographs present a challenge in our data as we cannot discern how the words were pronounced, making it impossible to distinguish between different meanings. For instance, the Hebrew word "בגד" could be pronounced as "Beged," meaning "clothing item," or as "Bagad," meaning "betrayed." Similarly, "בוקר" could refer to "morning" when articulated with penultimate stress or "cowboy" when pronounced with terminal stress. Since pronunciation is not documented, homographs are treated as a single entity in our analysis, potentially conflating distinct meanings.

Second, the word embeddings did not ideally capture the semantic relationships between words in Hebrew. The embeddings seem to represent orthographic structure of words more than the semantic relationships. For example, the word "Balsami" ("בלסמי", "Balsamic"), which refers to a an Italian sauce, is closer to the word "Balam" ("בלם", "Breaked") and "Baldar" ("בלדר", "Courier") more than it is closer to "Bashlan" ("בשלן", "cook"). This could be due to the relatively low performance of non-English language models. We also tried with `Norod78/hebrew-gpt_neo-tiny` and impressed that embeddings are not highly improved. However, follow-up experiment should test additional models.

Moreover, the clustering process did achieve high performance. Although we selected the number of clusters (k) based on the silhouette score, the score was quite low (0.059), indicating suboptimal clustering performance. Though this issue may stem from inadequate word embeddings, exploring alternative clustering methods could be beneficial. Also the labeling of clusters was suboptimal, potentially due to the limitations of the Gemini model in processing Hebrew language. This resulted in issues such as repetitive labels across clusters, including categories like foods, emotions, biblical concepts, plants, blessings, and family. Additionally, some clusters produced artificial or topics that combined unrelated items, such as "plants and body parts", "animals and products", or "religion and cooking". These suggest that the clustering approach did not capture the semantic relationships accurately, highlighting the need for improved models and methods tailored to the nuances of the Hebrew language. A possible alternative could be using BertTopic (Grootendorst et al., 2022), a well-known library for clustering and topic modeling, we initially avoided it because it is primarily designed for extracting topics from documents. Since our dataset consists of individual words rather than complete texts, we were skeptical of BertTopic's effectiveness in this context. However, given the unsatisfactory results from simpler algorithms, we believe it's worth reconsidering BertTopic as a potential solution.

We conclude that our results could stem from

## Acknowledgments

We express our sincere gratitude to Yevgeni Berzak, Omer Shubi, and Yoav Meir for their continuous and unwavering support.

## References

Andreou, G., & Trott, K. (2013). Verbal fluency in adults diagnosed with attention-deficit hyperactivity disorder (ADHD) in childhood. *Attention Deficit and Hyperactivity Disorders*, 5(4), 343–351. https://doi.org/10.1007/S12402-013-0112-Z

Bellman, R. (1966). *Dynamic programming. Science*, 153(3731), 34–37. https://doi.org/10.1126/SCIENCE.153.3731.34

Brysbaert, M., New, B., & Keuleers, E. (2012). *Adding part-of-speech information to the SUBTLEX-US word frequencies. Behavior Research Methods*, 44(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Chriqui, A., & Yahav, I. (2022). HeBERT and HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition. *INFORMS Journal on Data Science*, 1(1), 81–95. https://doi.org/10.1287/ijds.2022.0016

García-Castro, V. (2022). Exploring the role of verbal fluency in L2 Vocabulary Learning: Evidence from university students in the United Kingdom. *Actualidades Investigativas En Educación*, 22(2), 1–24. https://doi.org/10.15517/aie.v22i2.48887

Kavé, G., & Knafo-Noam, A. (2015). Lifespan development of phonemic and semantic fluency: Universal increase, differential decrease. *Journal of Clinical and Experimental Neuropsychology*, 37(7), 751–763. https://doi.org/10.1080/13803395.2015.1065958

Lehtinen, N., Kautto, A., & Renvall, K. (2023). Frequent native language use supports phonemic and semantic verbal fluency in L1 and L2: An extended analysis of verbal fluency task performance in an L1 language attrition population. *International Journal of Bilingualism*. https://doi.org/10.1177/13670069231193727

Li, Y., Li, P., Yang, Q. X., Eslinger, P. J., Sica, C. T., & Karunanayaka, P. (2017). *Lexical-semantic search under different covert verbal fluency tasks: An fMRI study. Frontiers in Behavioral Neuroscience*, 11. https://doi.org/10.3389/fnbeh.2017.00131

Lindsay, H., Mueller, P., Linz, N., Mina, M., Zeghari, R., König, A., & Tröger, J. (2021). Dissociating Semantic and Phonemic Search Strategies in the Phonemic Verbal Fluency Task in early Dementia.

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, 114(1), 29–41. https://doi.org/10.1016/j.cognition.2009.08.014

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, 41(1), 49–100. https://doi.org/10.1006/COGP.1999.0734

Mueller, K. D., Koscik, R. L., LaRue, A., Clark, L. R., Hermann, B., Johnson, S. C., & Sager, M. A. (2015). Verbal fluency and early memory decline: Results from the Wisconsin registry for Alzheimer's prevention. *Archives of Clinical Neuropsychology*, 30(5), 448–457. https://doi.org/10.1093/arclin/acv030

Paula, F. S. F., Wilkens, R., Idiart, M. A. P., & Villavicencio, A. (2018). Similarity Measures for the Detection of Clinical Conditions with Verbal Fluency Tasks.

Pettit, L., McCarthy, M., Davenport, R., & Abrahams, S. (2013). Heterogeneity of letter fluency impairment and executive dysfunction in Parkinson's disease. *Journal of the International Neuropsychological Society : JINS*, 19(9), 986–994. https://doi.org/10.1017/S1355617713000829

Ravona-Springer, R., Sharvit-Ginon, I., Ganmore, I., Greenbaum, L., Bendlin, B. B., Sternberg, S. A., Livny, A., Domachevsky, L., Sandler, I., Ben Haim, S., Golan, S., Ben-Ami, L., Lesman-Segev, O., Manzali, S., Heymann, A., & Beeri, M. S. (2020). The Israel Registry for Alzheimer's Prevention (IRAP) Study: Design and Baseline Characteristics. *Journal of Alzheimer's Disease : JAD*, 78(2), 777–788. https://doi.org/10.3233/JAD-200623

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5(JUL). https://doi.org/10.3389/fpsyg.2014.00772

Troyer, A. K., Moscovitch, M., Gordon, W., Alexander, M. P., & Stusss, D. (1998). Clustering and switching on verbal fluency: the effects of focal frontal- and temporal- lobe lesions (Vol. 25).

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and Switching as Two Components of Verbal Fluency: Evidence from Younger and Older Healthy Adults (Vol. 11, Issue 1).

Wechsler, D. (1981). Wechsler adult intelligence scale-revised. *New York*, N.Y. :Psychological Corporation, 1896–1981.

Wright, L. M., De Marco, M., & Venneri, A. (2023). Current Understanding of Verbal Fluency in Alzheimer's Disease: Evidence to Date. *Psychology Research and Behavior Management*, 16, 1691–1705. https://doi.org/10.2147/PRBM.S284645

# Appendix

## A Number of Clusters for the K-Means Algorithm.

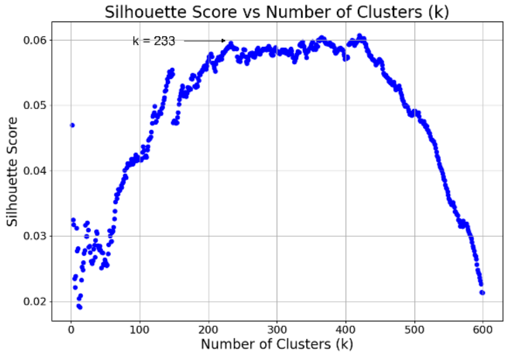


Figure A: Silhouette score vs Number of Clusters

A wide range of possible clusters from k=2 to k=600 by checking their silhouette score. A silhouette score of 0.059 was obtained for $k = 233$ clusters. This number was selected because it achieves a relatively high silhouette score, just before scores reach a plateau as k increases.

## B Prompt for topic modeling.

You are getting a list of words that start with the letter "Bet" in Hebrew. The words were generated in the 'verbal fluency' task, a cognitive test where participants have 60 seconds to say as many words that start with the letter bet as they can.
Our research question is: are the topics that emerge from native Hebrew speakers' answers different from the topics that emerge from non-native speakers' answers?
To answer, please find the common thread between words in the same cluster, and create a short label. The label cannot be 'words that start with the letter Bet'. Make sure you only return the label and nothing more.
Your words are: [words]

Figure B: Prompt for Topic Modeling

This prompt was used in Google's free API to gather topics for each of the 233 clusters. In Figure 2 we compare the number of words that were generated by L1 vs. L2 participants from each topic.