WORD ASSOCIATIONS:

A COMPARISON OF AI AND HUMAN WORD ASSOCIATION PATTERNS

CEREN OKSAL

BOĞAZİÇİ UNIVERSITY

2025

WORD ASSOCIATIONS:

A COMPARISON OF AI AND HUMAN WORD ASSOCIATION PATTERNS

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Linguistics

by

Ceren Oksal

Boğaziçi University

2025

Word Associations:

A comparison of AI and Human Word Association Patterns

The thesis of Ceren Oksal

has been approved by:

Assist. Prof. Ümit Atlamaz _____
(Thesis Advisor)

Assoc. Prof. Kadir Gökgöz _____

Prof. Olcay Taner Yıldız _____
(External Member)

January 2025

DECLARATION OF ORIGINALITY

I, Ceren Oksal, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisors and thesis committee at Boğaziçi University, including final revisions required by them.

Signature: . . . . . .

Date: . . . . . .

ABSTRACT

Word Associations: A Comparison of AI and Human Word Association Patterns

This thesis explores the similarities and differences between human word associations and those generated by large language models (LLMs) through a systematic comparison of responses in a Turkish word association task. Using human participants and selected LLMs, this study examines key metrics, including response diversity, associative strength, sensitivity to linguistic features such as morphology and concreteness, and structural alignment within associative networks. The methodology involves a continuous word association experiment conducted with 381 human participants in addition to simulations using state-of-the-art LLMs, including GPT-4o mini, Trendyol LLM v1.8, and Cosmos LLaMA. Comparative analyses employ techniques such as cosine similarity, Jaccard similarity, and graph-based metrics to evaluate structural and associative parallels. Results indicate notable differences in the associative patterns of humans and LLMs, particularly in response diversity and sensitivity to morphological and semantic properties. While LLMs exhibit strengths in replicating certain associative structures, they face challenges in mimicking the nuanced cognitive processes observed in humans. The findings underscore the potential of word association tasks as a lens for understanding human cognition and evaluating artificial intelligence systems. This research bridges traditional psycholinguistic methodologies with contemporary computational approaches, advancing our understanding of semantic memory and highlighting the evolving potential and limitations of LLMs in approximating human cognition.

ÖZET

Kelime Çağrışımları: Yapay Zeka ve İnsan Kelime Çağrışım Yapılarının Karşılaştırılması

Bu tez, Türkçe kelime çağrışımı görevinde insan kelime çağrışımları ile büyük dil modelleri (LLM'ler) tarafından üretilen çağrışımlar arasındaki benzerlikleri ve farklılıkları sistematik bir karşılaştırma yoluyla incelemektedir. İnsan katılımcılar ve seçilen LLM'ler kullanılarak yapılan bu çalışma, yanıt çeşitliliği, çağrışım gücü, morfoloji ve somutluk gibi dilsel özelliklere duyarlılık ve çağrışım ağlarının yapısal uyumu gibi temel ölçütleri değerlendirmektedir. Metodoloji, 381 katılımcıyla yürütülen sürekli bir kelime çağrışım deneyini ve bu deneyin GPT-4o mini, Trendyol LLM v1.8 ve Cosmos LLaMA gibi güncel LLM'lerle simüle edilmesini içermektedir. Karşılaştırmalı analizler, yapısal ve çağrışım benzerliklerini değerlendirmek için kosinüs benzerliği, Jaccard benzerliği ve grafik tabanlı metrikler gibi teknikler kullanmaktadır. Sonuçlar, özellikle yanıt çeşitliliği ve morfolojik ve anlamsal özelliklere duyarlılık açısından insanlar ve LLM'ler arasındaki çağrışım yapılarında önemli farklılıklar olduğunu göstermektedir. LLM'ler belirli çağrışım yapılarını yeniden oluşturma konusunda güçlü yönler sergilerken, insanlarda gözlemlenen karmaşık bilişsel süreçleri taklit etmede zorluklarla karşılaşmaktadır. Bulgular, kelime çağrışım görevlerinin hem insan bilişini anlamak hem de yapay zeka sistemlerini değerlendirmek için bir araç olarak potansiyelini vurgulamaktadır. Bu araştırma, geleneksel psikolinguistik metodolojileri modern hesaplamalı yaklaşımlarla birleştirerek anlamsal belleği anlamamıza katkıda bulunmakta ve LLM'lerin insan bilişini taklit etme konusundaki gelişen potansiyel ve sınırlamalarını öne çıkarmaktadır.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Ümit Atlamaz. From the moment I sent him an email asking if I could work with him for my thesis, he has been an unwavering source of support and guidance. His kindness, wisdom, and encouragement have been instrumental throughout my journey. At the beginning of my studies, I often felt lost, but his thoughtful mentorship helped me navigate both linguistics and natural language processing with confidence. Dr. Atlamaz's supportive words and compassionate approach were a source of reassurance during stressful times, and his mentorship has been a defining aspect of my growth during this journey. I feel incredibly fortunate to have had him as my advisor, and I will always carry the lessons I learned from him into my future endeavors.

I am also profoundly grateful to Dr. Pavel Logacev, who was my advisor during my undergraduate studies. His initial idea for my research, which evolved into this thesis, has been invaluable. Moreover, he introduced me to the world of data science, which is now my professional focus. His influence on both my academic and career paths has been immeasurable, and I will always cherish his guidance.

My sincere thanks go to the committee members for this thesis: Dr. Kadir Gökgöz and Dr. Olcay Taner Yıldız. During the second year of my undergraduate studies, I approached Dr. Kadir Gökgöz with a desire to engage more deeply with linguistics beyond my coursework. He took my aspirations seriously and welcomed me into his project, which introduced me to the fascinating world of TİD. This was an entirely new area for me, and I quickly grew to love it. His support continued throughout my MA years, and I am deeply appreciative of his encouragement and

belief in me. I would also like to thank Dr. Olcay Taner Yıldız for giving me the opportunity to work at Starlang. This experience contributed significantly to my professional development.

I also wish to acknowledge the institutions that provided financial support during my studies. I am grateful to Dr. Ümit Atlamaz for offering me the opportunity to work as a research assistant on his research project. Thanks to his support, I was able to be funded by the Boğaziçi University Research Fund Grant Number 19348. Working on this project was one of the highlights of my MA years, and I deeply appreciate the experience and guidance I gained under his mentorship. Additionally, I thank TÜBİTAK for their support through the 2210-A National Scholarship Programme for Master's Students, which contributed significantly to my academic journey.

Following the invaluable support I received from my mentors and the institutions that made my journey possible, I would like to express my gratitude to all the professors at the Linguistics department for fostering such a supportive environment. I am also deeply thankful to my cohort, whose brilliance and companionship have made the challenging times more manageable and the successes even more rewarding. Among them, I am especially grateful to my friend Onur Keleş, who began this MA journey alongside me. Thank you for being a constant source of support and fellowship. From answering my endless questions, no matter how trivial, to sharing the ups and downs of graduate life, your friendship has made this experience much brighter.

I am deeply thankful to my best friend, Helin Taşar, who has been a constant presence in my life since our high school days. From her perfectly timed jokes that never fail to make me laugh to our fascinating conversations that I always look

forward to, she has brought so much light and inspiration to my life. Being her friend is a joy in itself, and her presence has made my MA journey all the more memorable and meaningful.

I would also like to express my heartfelt gratitude to Rumet Solmaz, who entered my life as I was applying to this MA program and has been a constant source of light and comfort ever since. His kindness, humor, and encouragement have brought endless joy to these years. He has been by my side through every step of this journey, making even the most ordinary moments feel special. Our study dates remain some of the most cherished memories of my MA experience, and I am endlessly grateful for his presence in my life.

Finally, I want to thank my family −my mother, Nehir Oksal; my father, Muammer Oksal; and my brother, Ahmet Kaan Oksal− for their unconditional love and unwavering support. Whether I came home after long, exhausting days or with exciting news to share, they were always there to listen to me talk about even the smallest details of my day. Their belief in me has been my greatest strength, keeping me grounded and motivated throughout this journey.

I must also mention my dog, Oksi, the craziest yet most wonderful companion, whose boundless energy and love have brought so much joy to my life.

To all of you, I owe my deepest thanks. This thesis would not have been possible without your presence, guidance, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Word associations provide a unique lens through which the intricacies of human language can be investigated. This simple task, where individuals respond to a cue word with the first word that comes to mind, has served as a foundational tool in psycholinguistic and cognitive research over decades. Word associations, according to Nelson et al. (2004), are commonly accepted to represent our lexical knowledge gained via world experience through words and the connections between them. These structures are argued to capture significant facets of meaning or the semantic representation of words that go beyond lexical usage patterns (Prior & Bentin (2008); Szalay & Deese (1978); de Groot (1988); Nelson et al. (2004)). By bridging lexical knowledge and semantic representation, word association tasks have shaped our understanding of semantic memory, cognitive processes, and linguistic organization, making them a valuable tool for both theoretical and applied research.

As described by De Deyne et al. (2013), this research has followed three major traditions. The first focuses on direct associative strength, measuring the connections between words based on stimulus-response patterns, which has been instrumental in understanding memory processes and associative priming. The second tradition delves into second-order associations and distributional similarities, exploring the shared associations of word pairs to uncover indirect semantic relationships. Finally, the third tradition employs network-based approaches, constructing large-scale semantic networks to analyze the topology, centrality, and structural properties of the mental lexicon.

These methodologies collectively provide a foundation for understanding the complexities of human cognition and language. Recently, with the advent of large language models (LLMs), the landscape of linguistic research has expanded significantly. These models represent an important advancement in artificial intelligence, demonstrating remarkable capabilities in language translation, analysis, and contextual understanding (Vaswani et al., 2017; Brown et al., 2020; Radford et al., 2019). These models excel in producing fluent, coherent text. However, despite their impressive achievements, LLMs face notable limitations, particularly in their ability to generalize to novel contexts, update knowledge post-training, and capture common-sense semantic knowledge (Marcus & Davis, 2019; Mccoy, Min, & Linzen, 2020). These shortcomings underscore the challenges LLMs encounter in simulating the flexible and adaptive nature of human language use.

LLMs are typically trained on massive datasets, relying on statistical patterns to generate predictions. While this approach allows them to process and produce language at scale, it also highlights their dependency on fixed training distributions and their difficulty in handling scenarios outside these distributions. This constraint contrasts with human linguistic capabilities, which are grounded in multimodal experiences, and enriched by continuous interaction with the environment.

These limitations provide a context for exploring how LLMs perform in tasks traditionally used to study human cognition, such as word associations. While LLMs have been employed to simulate various linguistic experiments (Aher et al., 2023; Shin and Song, 2023), their application to word association tasks remains relatively unexplored. Investigating this domain not only tests the boundaries of LLMs' linguistic abilities but also provides insights into the extent to which these models can approximate human-like cognitive processes.

This thesis aims to bridge this gap by addressing the research question: What are the similarities and differences between human word association patterns and those generated by LLMs? To answer this, I conduct a Turkish word association experiment with human participants and simulate this experiment using LLMs. This investigation explores areas such as response diversity, associative strength, sensitivity to linguistic features such as morphology and concreteness, and the structural alignment of their resulting associative networks.

Through a systematic comparison of word associations produced by human participants and those generated by selected LLMs, this study examines similarities and differences across key metrics. By exploring the influence of linguistic variables and leveraging established traditions in word association research, this work integrates computational techniques to address the unique challenges posed by LLMs. The findings contribute not only to our understanding of human cognition but also to the evolving role of artificial intelligence in modeling complex cognitive processes.

The structure of this thesis is as follows: Chapter 2 provides a literature review and discusses foundational and contemporary studies on word association tasks, semantic networks, and the role of LLMs in linguistic research, situating the current work within this broader context.

Chapter 3 describes the methodology, detailing the design of the word association experiment conducted with human participants, the simulation of this experiment using selected LLMs, and the analytical frameworks used for comparison.

Chapter 4 presents the findings of the analysis, highlighting key similarities and differences in descriptive results, response diversity, associative strength,

linguistic sensitivity, and network structures between human and LLM-generated associations.

Chapter 5 presents a discussion by interpreting the results in relation to the research question, considering their implications for understanding human cognition and evaluating LLMs, while addressing limitations and integrating perspectives on future research.

The thesis is concluded in Chapter 6, which provides a general summary of the study's contributions and key findings.

CHAPTER 2

LITERATURE REVIEW

Word association tasks are memory tasks where participants are prompted to respond to a given word, either written or spoken, with the first word that comes to mind. Since the late nineteenth century, word association studies have been a valuable addition to behavioral research (Boring, 1950; Deese, 1965). They have served as a conventional tool for assessing various lexical properties, such as association strength between words, which can be used as control variables in psycholinguistic experiments (Nelson et al., 2004). Over time, research into word associations expanded as modern cognitive psychologists and linguists have reintroduced these studies, focusing on language and memory. This expansion has allowed researchers to investigate and model the organization of semantic representations that are stored in memory (De Deyne et al., 2013, 2019; Dubossarsky et al., 2017; Elias Costa et al., 2009; Nelson et al., 2000; Steyvers et al., 2005; Steyvers & Tenenbaum, 2005)

2.1 Foundational studies on word association

According to Nelson et al. (2004) “free association taps into lexical knowledge acquired through world experience” and the generation of word associations can capture essential aspects of meaning or the semantic representation of words. To provide a comprehensive resource for researchers interested in understanding the structure of lexical knowledge through word associations, Nelson et al. (2004) conducted a pivotal study that involved the collection of a vast dataset, with more than 6,000 participants generating nearly three-quarters of a million responses to 5,019 stimulus words. This effort resulted in the largest free association database

available at the time, offering valuable insights into how word associations reflect the underlying structure of semantic memory.

In their methodology, participants were presented with a list of words and asked to respond with the first word that came to mind. This process, known as discrete free association, limits participants to producing only a single response per cue. The words selected for the study were predominantly nouns but also included adjectives, verbs, and other parts of speech. The selection of these words was driven by various research needs, including their potential utility in cuing and priming studies. The authors recorded each cue word and the frequencies of its responses, corrected spelling errors, and created rules to pool items such as plural answers or different tense and grammatical forms of the same word stems. They also calculated free association probabilities. In essence, these probabilities measure the likelihood that one word will activate another in memory by indexing the relative strength of the association between the cue word and the response. The authors calculated forward associations, which represent the likelihood that a particular word would cue another specific word in memory, and provided a n x n associative networks for more than 4,000 words. Their analysis revealed that word associations tend to form clusters, much like small-world networks, where certain words are more closely connected due to shared experiences and contextual usage. A small-world network is a type of graph where most nodes are not directly connected but can still be reached from one another through a small number of steps (Steyvers & Tenenbaum, 2005).

The 2004 study built on prior work by Nelson et al. (2000), which explored the role of free association probabilities in predicting memory performance and other cognitive behaviors. In the 2000 study, participants were asked to generate two responses to each word, allowing the researchers to explore the reliability of

associative strength and set size, showing that first responses tend to be more reliable indicators of strong associations, whereas second responses introduce weaker, less stable associations. These measures are:

- Associative Strength reflects the probability of a particular response being elicited by a specific cue word, serving as a measure of how strongly two words are connected in memory. This probability is calculated by dividing the number of participants who produced a given response by the total number of participants in the task. Higher associative strength indicates a robust and consistent lexical connection, while lower values suggest weaker or less stable associations. This metric captures the likelihood of a cue-response relationship and is essential for understanding the organization of the mental lexicon.
- Set Size refers to the number of unique responses (associates) elicited by a cue word across participants in a free association task. It provides a measure of the diversity of associations a cue can generate. Words with smaller set sizes typically evoke consistent and commonly shared responses, reflecting strong associative links. Conversely, words with larger set sizes are associated with greater variability in responses, indicating weaker or more diffuse associations. Set size offers insights into the breadth of a word's associative network and is often analyzed separately for different response orders, such as first and second responses, to examine the stability and variability of associations.

Nelson et al., (2000) found the free association probabilities to be useful for predicting memory performance and other behaviors. In this study, set size was

calculated separately for first and second responses. Words with smaller sets of associates were found to have higher reliability in both strength and set size. For example, they found that words with smaller associative sets and strong primary associations tended to produce more consistent responses across participants, meaning that the same words were often mentioned as associates. In contrast, words with larger associative sets produced more varied responses, indicating weaker associative strength. Their findings showed that free association can provide a robust measure of lexical strength, particularly for words with smaller sets of strong associative links. However, the reliability decreased when additional responses were elicited, suggesting that first responses are more predictive of memory performance. Both studies argue that these associative links are not arbitrary but instead reflect the organization of lexical knowledge in the mind. Nelson et al. (2000, 2004) argue that these associations are dynamic as "they are derived from the stream of everyday experience that changes gradually over time". They are also sensitive to cultural and regional influences, and capable of evolving over time. For example, certain associations may be stronger in specific populations or regions, reflecting the influence of local culture and experiences.

2.2 Advancements in Association Network Models

Building on the foundational work of Nelson et al. (2000, 2004), which established the significance of associative strength and set size in understanding lexical organization, De Deyne and Storms (2008) introduced a novel approach to word association tasks.

Authors conducted a study that expanded on traditional word association tasks by using a continuous association task, allowing participants to generate three

responses per cue word, rather than the typical one-response-per-cue approach used in Nelson et al. (2000, 2004). Authors argued that this design provides a more comprehensive dataset, capturing both strong and weak associations that a discrete task might miss. The study involved 10,292 participants, with an average age of 24 years, drawn from university students and high school students in Belgium. Participants were asked to provide associations for 1,424 Dutch words, selected from previous studies by Ruts et al. (2004) and Severens et al. (2005). These words spanned several semantic categories, including natural categories (e.g., fruits, animals), artifact categories (e.g., tools, musical instruments), and abstract concepts. Data collection took place in two phases. In Phase 1, participants were given paper questionnaires with stimulus words from specific semantic categories. They were instructed to write down the first three words that came to mind for each cue word. In Phase 2, the study expanded to include a web-based questionnaire, which followed the same format as the paper version. In both phases, the most common responses were iteratively added to the collection of cues, with each new set of cues generating additional responses that were then included in subsequent iterations, following a snowball sampling approach.

The study gathered an average of 268 responses per cue word. Responses were preprocessed to ensure consistency. This included standardizing spelling, converting responses to lowercase, and replacing numerical values with their text counterparts. A major innovation of this study was the use of a continuous task, which allowed participants to provide multiple responses per word. The authors were particularly interested in set size. By asking for more than one response, De Deyne and Storms (2008) was able to capture weaker associations that would not be detected in a discrete task. The set size increased substantially as second and third

responses were added, growing from an average of 11 associations for the first response, 13 and 14 for second and third responses respectively, to 31 associations when all responses were combined.

To avoid biasing responses due to chaining (when participants base their responses on their own previous responses rather than the original cue), participants were given clear instructions to focus only on the original cue word when providing subsequent responses. The authors calculated split-half reliability by dividing the dataset into two random halves and calculating the correlation between the sets for each cue–response pair. This method evaluates the consistency of the data by assessing whether the associations generated in one half of the dataset are similar to those generated in the other half. The logic behind split-half reliability is that if the associations are stable and reflect true underlying relationships rather than random noise, the two halves should produce highly correlated results. This approach is comparable to the reliability procedures used by Nelson et al. (2000). The results showed high reliability for the first response ($r = 0.82$), which increased to 0.89 when both first and second responses were included, and 0.91 when all three responses were combined. These findings demonstrate that collecting additional responses improves reliability by capturing a broader and more representative sample of the word's associative structure.

The study also measured correlations between the set sizes of Dutch words and English words from Nelson et al. (2004). Although there was some overlap between the Dutch and English sets (1,022 cues), the correlation was only moderate ($r = 0.28$ for first responses), suggesting that word associations vary significantly between languages, especially for weaker associations captured in the continuous task. The use of multiple responses also enabled a richer exploration of semantic

proximity –the closeness between words in meaning. The study found that later responses contributed significantly to the diversity of associations, indicating that continuous tasks provide a denser representation of semantic memory than single-response tasks.

By plotting the growth of set size over time, De Deyne and Storms demonstrated that the set size increases in a nonlinear fashion as additional responses are gathered, stabilizing after the third response (Figure 1). De Deyne and Storms also investigated the relationship between association frequency and other psycholinguistic variables such as word frequency, age of acquisition (AoA), and imageability. They found that associations generated frequently were more likely to represent early acquired words that were easy to visualize. The correlations were stronger for age of acquisition (AoA) than for word frequency, particularly for nouns and adjectives, indicating that early-learned words tend to dominate associations.

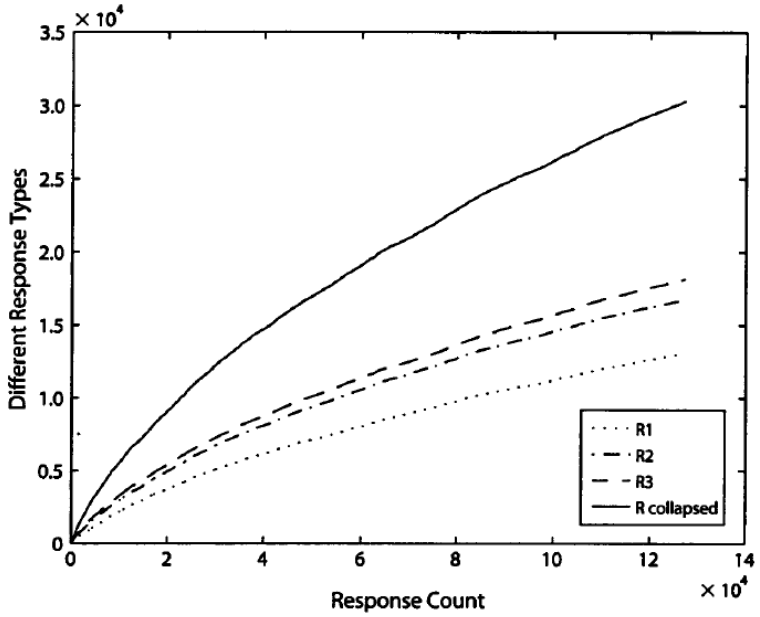


Figure 1. Growth of set size (i.e., different response types) for the first (R1), second (R2), third (R3), and collapsed (R collapsed) responses from De Deyne and Storms (2008)

The authors noted that this richer dataset could be useful for high-dimensional vector models of semantic memory, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and Hyperspace Analogue to Language (HAL; Burgess & Lund, 1997), which rely on text colocation data but could struggle to capture the nuanced associative links between words if word associations does not "simply stem from the learner's sensitivity to the statistical properties of language" (De Deyne & Storms, 2008). The continuous association data provides a more accurate yardstick for evaluating such models.

Together, these studies underscore the utility of free association tasks in capturing both the personal and collective dimensions of lexical knowledge. De Deyne and Storms (2008) emphasized how the continuous task leads to richer, more dynamic associations, while the earlier studies by Nelson et al. (2000, 2004) highlighted the role of associative strength and set size in revealing the structure of lexical memory.

De Deyne et al. (2013) continued their data collection from their previous study (De Deyne & Storms, 2008) and further investigated the use of extensive word association data to create a more realistic approximation to the human lexicon. The authors argue again that the use of continued associations and their richer dataset enables a better understanding of lexical access and semantic cognition, as it captures weaker and more diverse associations that are often overlooked in single-response tasks.

The study draws on data collected from over 70,000 participants, who provided associations for more than 12,000 cue words. The participants were drawn from a diverse pool, including students from Belgian universities and volunteers from a broader online audience. The researchers used the collected data to construct a

weighted, directed network where each node represents a word, and the links between them represent the associative strength between a cue and its response as "networks are considered the natural representation of word associations" and they "correspond to an idealized localist representation of our mental lexical network" (De Deyne et al. 2013, p.1).

A methodological innovation in this study was the construction of three separate networks based on the number of responses: one that included only the first response (G1), one that incorporated the first and second responses (G2), and one that included all three responses (G3). By comparing these networks, the authors were able to demonstrate that the multiple-response approach resulted in a more heterogeneous and informative dataset. In particular, the third network, which included all three responses, provided better predictions for lexical decision tasks and measures of semantic relatedness. This finding supports the idea that richer data from multiple responses allows for more accurate modeling of the mental lexicon, as it captures a broader range of associations. The researchers examined how this richer network structure affected key network properties, such as centrality, clustering, and betweenness, which are important for understanding word retrieval processes. In-degree centrality reflects the number of connections a node (word) has, indicating its importance in the network. Betweenness is another centrality measure that quantifies how often a particular node acts as a bridge along the shortest path between two other nodes. Clustering coefficient assesses how densely connected the nodes in the network are. They found that using multiple responses led to networks with higher density and lower path lengths between nodes, suggesting that the mental lexicon is more interconnected than previously thought. These networks were then used to predict performance in various cognitive tasks by calculating the correlations

between datasets collected from lexical decision tasks (which measure how quickly participants can recognize words) and semantic similarity judgments (where participants rate the similarity between pairs of words) with word association networks. For example, they calculated the semantic relatedness using the cosine overlap between the pairs of words in each set, by calculating all pairwise cosine indices, from which a full similarity matrix was obtained consisting of 12,428 × 12,428 similarity values. And then, calculated the correlations between semantic relatedness scores in Miller and Charles (1991), new data they collected for different semantic categories and the model inferred semantic relatedness for directed networks G1, G2, and G3 and undirected network G'3.

Table 1. Correlations (ρ) Between Semantic Relatedness (Miller and Charles (1991) or Similarity and The Model Inferred Semantic Relatedness for G1, G2, G3 and G'3 from De Deyne et al. (2013)

Set	Category	n	G_1	G_2	G_3	G'_3
MC		30	.88	.91	.91	.86
Abstract	Virtues	105	.70	.76	.77	.56
	Emotions	91	.52	.55	.57	.45
	Art forms	91	.55	.59	.65	.48
	Media	105	.55	.60	.66	.50
	Crimes	91	.49	.54	.59	.43
	Sciences	105	.36	.41	.47	.32
	Diseases	105	.24	.43	.61	.05†
	M		.49	.55	.62	.40
Animals	Birds	406	.39	.49	.49	.26
	Insects	300	.56	.62	.66	.45
	Fish	231	.67	.76	.75	.66
	Mammals	435	.44	.50	.54	.26
	M		.51	.59	.61	.41
Artifacts	Clothing	378	.63	.65	.65	.55
	Kitchen utensils	496	.41	.47	.51	.33
	Music instruments	325	.50	.51	.50	.21
	Tools	325	.48	.55	.59	.45
	Vehicles	435	.63	.68	.69	.57
	Weapons	171	.65	.65	.66	.65
	M		.55	.59	.60	.46
Animals	set A	253	.59	.67	.71	.58
	set B	253	.65	.70	.73	.56
Artifacts	set A	435	.61	.65	.63	.59
	set B	435	.44	.56	.61	.41

The results showed that the network-based measures, particularly those derived from the G3 network, outperformed traditional single-response networks in predicting participants' performance on these tasks. Their findings suggest that continued word associations, rather than single responses, offer a more realistic representation of the mental lexicon, allowing researchers to make more accurate predictions about cognitive tasks involving language processing and comprehension.

Another comprehensive dataset of English word associations is presented in De Deyne et al. (2019), collected from more than 90,000 participants for over 12,000 cue words as part of the “Small World of Words” project, making it the largest such dataset in English. As in their previous works, a continuous free association task was implemented. Participants were recruited online using various platforms, including social media and university websites, with the only requirement being fluency in English. Data were collected from participants of different demographics, including native English speakers from various regions. The majority were native speakers from the U.S., U.K., Canada, and Australia. The cue words themselves were chosen using the snowball sampling method again, ensuring that both frequent and less frequent words were represented. The set included words from previous studies, such as the University of South Florida norms (Nelson et al., 2004) and semantic feature production norms (McRae et al, 2005). Participants were asked to avoid typing full sentences and to respond directly to the cue word presented, without chaining responses based on their previous ones. Each stimulus was shown in random order, with participants providing up to three responses for each cue word.

The data went through preprocessing, where responses were normalized to remove inconsistencies such as punctuation and multiple spaces. Spellings were standardized to American English, and responses were manually checked for proper

capitalization and corrections. The dataset includes filtering criteria that excluded certain participants, such as those who used non-English words, produced sentences rather than single-word responses, or gave overly repetitive responses. Based on the maximum highly connected component, which guarantees that only words with both incoming and outgoing edges are kept and that there is a path linking every conceivable pair of words in the graph, they created two distinct graphs. While the second graph, GR123, was built using all the replies that were generated (R123), the first graph, GR1, was only built using the initial response data (R1).

They evaluated response chaining, which is a participant “using their own previous response as a cue or prime for their next response to the same word”, by comparing how likely a specific second response (R2) was to occur depending on the participant's first response. They tested the conditional probability of a specific second response occurring, given that a specific first response was made. For instance, in the example of the cue word "sun" they analyzed whether participants were more likely to give "star" as their second response if they gave "moon" as their first response. They then constructed 2x2 contingency tables for R1 and R2 pairs to examine how frequently a second response occurred based on whether a specific first response was provided. For the example of "sun", they compared the frequency of “star” as the second response when “moon” was the first response against cases where "moon" was not the first response. The authors calculated Bayes factors to quantify the evidence for or against the presence of chaining. A high Bayes factor indicated strong evidence for chaining between specific R1 and R2 pairs. For instance, a Bayes factor greater than 10 was considered strong evidence for chaining between the first and second responses. They found strong evidence for chaining in around 1% of the cases, where certain cue–R1–R2 triples had a high likelihood of

chaining. Examples include the cue word "siblings," where the first response "brothers" was likely to be followed by the second response "sisters." Other examples included “salt” followed by “pepper,” and “arm” followed by “leg.” This analysis revealed that while there was some evidence of response chaining, it was not widespread. Since strong evidence for chaining was present in only about 1% of all cases, and moderate evidence in 19% of cases, the authors concluded that while response chaining occurs to some extent, especially for highly related word pairs, the majority of participants did not rely on chaining. Most responses reflected genuine associations with the original cue word.

The final SWOW-EN dataset was evaluated in a variety of contexts, including its ability to predict lexical decision times, naming times, and semantic similarity judgments. The authors found that measures derived from the dataset were highly predictive of human performance in these tasks, validating the utility of the dataset for psychological and linguistic research. For example, the response frequencies from the dataset were used to predict lexical decision times (how quickly participants could determine whether a string of letters was a valid word) and naming times (how quickly participants could name a word when presented with its visual form). The authors calculated correlations between the association frequency data from SWOW-EN and the reaction times in the lexical decision and word naming tasks. Association frequency is the number of times a word is given as a response to a cue word. The frequency of a response, according to the authors, can indicate which words are central in the lexicon and “might determine how efficiently lexical information can be retrieved”. The authors hypothesized that higher association frequencies –representing stronger, more frequently accessed associations– could be predictive of faster lexical processing times because strongly associated words

should be retrieved more easily. To validate this hypothesis, the authors tested how well association frequencies from their dataset could predict performance in three key lexical processing tasks:

- Lexical Decision Task (LDT): In this task, participants decide as quickly as possible whether a string of letters is a valid word or a non-word.
- Word Naming: Participants are asked to read a word aloud as quickly as possible after seeing it.
- Semantic Decision Task: Participants categorize words as abstract or concrete, requiring them to access the meaning of the word.

The data for the lexical decision and word naming tasks came from the E-Lexicon Project (Balota et al., 2007), which contains reaction times (RTs) for over 40,000 English words. The data for the semantic decision task came from the Calgary Semantic Decision Project (Pexman et al., 2017), which includes judgments for about 10,000 words.

The authors calculated correlations between the association frequency data from SWOW-EN and the reaction times in the lexical decision and word naming tasks. For comparison, they also used word frequency from the SUBTLEX-US database, which provides word frequency counts based on large corpora of natural language use (Brysbaert & New, 2009). Word frequency is known to be one of the most powerful predictors of lexical processing, as words that occur more frequently in language are generally processed more quickly (Brysbaert & New, 2009). The association frequency data from SWOW-EN showed a moderate correlation with both lexical decision times and naming times. Words that had higher association frequencies (i.e., words that were common responses to many different cues) tended to have shorter reaction times. This suggests that words that are more central in the

associative network are easier to access in the mental lexicon, leading to faster processing times. However, word frequency (as measured by SUBTLEX-US) was still a stronger predictor of reaction times than association frequency. This is consistent with previous research, which has found that word frequency is a dominant factor in determining how quickly words are processed. In contrast to the lexical decision and naming tasks, association frequency from SWOW-EN outperformed word frequency when predicting reaction times for the semantic decision task. This is likely because the semantic decision task requires participants to access word meanings, which is more directly related to word associations. Words with higher association frequencies tend to have richer associative networks, meaning they are connected to more words and concepts, which facilitates faster access to their meanings.

To further isolate the unique contribution of association frequency in predicting lexical processing, the authors conducted partial correlation analyses. This allowed them to examine the predictive power of association frequency after controlling the effects of word frequency from SUBTLEX-US. After controlling for word frequency, association frequency from SWOW-EN still contributed significantly to the prediction of reaction times in all three tasks. This suggests that association frequency captures additional variance in lexical processing that word frequency alone does not account for, particularly in tasks that involve accessing the meanings of words (like the semantic decision task). Interestingly, the reverse was not always true: when the authors controlled for association frequency, word frequency still explained most of the variance in the lexical decision and naming tasks, but not in the semantic decision task. This further highlights the importance of association frequency for tasks that require deeper semantic processing.

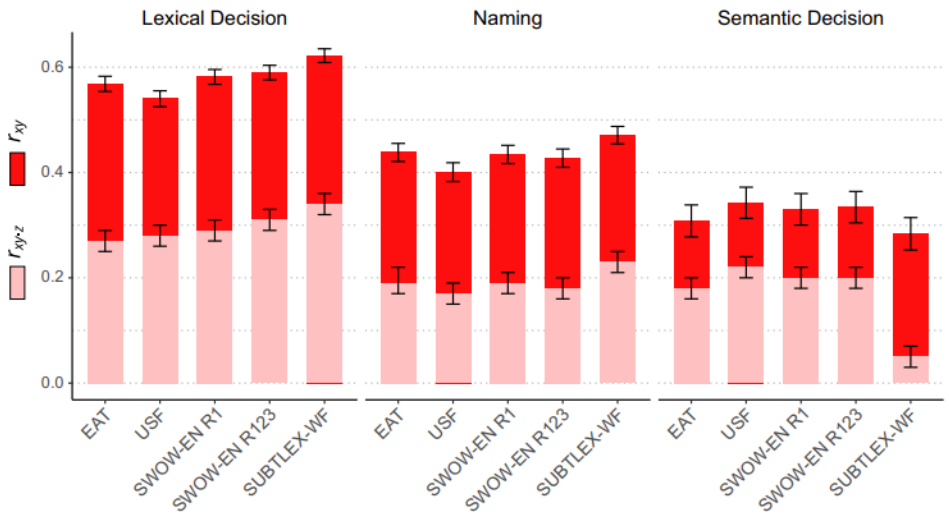


Figure 2. "Pearson correlations r_{xy} and 95% confidence intervals for naming and LDT data from the E-lexicon project and semantic decision latencies from the Calgary Semantic Decision (CSD) project (correlations multiplied by -1 for readability). Three different word association datasets (EAT, USF, and SWOW-EN) and one language-based measure of frequency derived from SUBTLEX are included. For the word association datasets, the partial correlations, indicated as $r_{xy \cdot z}$, are calculated given word frequency based on z = SUBTLEX-WF; for SUBTLEX-WF, the partial correlation $r_{xy \cdot z}$ removes the effect of the word-association datasets" (De Deyne et al., 2019)

Regarding similarity judgements, as traditional methods for estimating semantic similarity have relied heavily on corpus-based approaches, which use patterns of word co-occurrence in large text datasets, the authors argue that these methods may fail to capture deeper, conceptual connections between words that are not directly reflected in their co-occurrence patterns. Word associations, by contrast, can offer a more grounded measure of semantic similarity, as they reflect direct connections made by participants in response to cues. These associations are not constrained by linguistic context or frequency of co-occurrence but rather reflect meaning-based relationships between words in the human mental lexicon. The authors propose and explore three different ways to estimate semantic similarity based on their word association data. These measures vary in the amount of information they use, ranging from direct associations to more complex network-

based approaches that take into account both direct and indirect paths between words.

The most straightforward way to estimate semantic similarity between two words is to use their associative strength. This is the probability that a participant will respond with a particular word r when presented with a cue word *c* (denoted as *p(r | c)*). To compute semantic similarity using associative strength, the similarity is calculated by examining how many common responses there are for a pair of words in the dataset. In other words, two words are considered similar if they elicit similar sets of responses from participants. Authors used the cosine similarity measure to compute the similarity score. This is done by representing the association strengths between the two words and all possible responses as vectors and calculating the cosine of the angle between these vectors.

(1) $$S(c_a, c_b) = \frac{\sum_{i=1}^{N} p(r_i|c_a)p(r_i|c_b)}{\sqrt{\sum_{i=1}^{N} p(r_i|c_a)^2}\sqrt{\sum_{i=1}^{N} p(r_i|c_b)^2}}$$

The authors improve upon the basic associative strength measure by leveraging Pointwise Mutual Information (PMI), an information-theoretic measure that considers not just the frequency of responses but also how informative each response is. PMI takes into account the global distribution of responses, meaning it considers how often a response is given across all cue words, not just for a particular pair. PMI normalizes the associative strength by penalizing responses that are frequently given for many different cues, as these responses are less informative about the unique relationship between two words. This approach helps to emphasize more meaningful associations that go beyond mere frequency of co-occurrence, capturing the semantic specificity of associations.

The third and most sophisticated approach the authors use to estimate semantic similarity involves random walk models. These models go beyond direct associations and look at how words are connected through indirect paths in a broader associative network. In a random walk model, a word is considered similar to another word not only if they share direct associations but also if they are connected through a series of intermediary words. For example, if "apple" is associated with "fruit" and "fruit" is associated with "banana," then "apple" and "banana" are considered similar, even though they may not be directly associated. The random walk similarity is calculated by simulating a spreading activation process in the word association network (Collins & Loftus, 1975). Starting from a cue word, activation spreads to neighboring words (direct associations) and then to neighbors of neighbors (indirect associations). This process is akin to how information might spread through the brain's semantic network. To quantify this, the authors use a decaying random walk process, where the weight of indirect paths decreases as the number of intermediary words increases (Abott, Austerweil, & Griffiths, 2015). This is formalized in a recursive formula, where each step in the walk adds another layer of indirect associations.

Finally, the similarity between two words is calculated as the cosine similarity between the rows of the random walk matrix, corresponding to the two words in question. To evaluate the effectiveness of these measures, the authors compared the semantic similarity scores derived from their word association data to human judgments of similarity from various benchmark datasets. These datasets include well-known similarity and relatedness tasks such as SimLex-999 (Hill et al., 2015), WordSim-353 (Agirre et al., 2009), MEN (Bruni et al., 2012) and other smaller datasets such as RG1965 (Rubenstein & Goodenough, 1965), MTURK-771

(Halawi et al., 2012), and Radinsky2011 (Radinsky et al., 2011). The authors found that PMI outperformed basic associative strength in predicting human similarity judgments. This suggests that PMI's ability to account for global distributional information (by down-weighting frequent responses) provides a better measure of semantic similarity. Random walk models produced the best performance across all tasks. By considering both direct and indirect associations, these models capture a richer, more comprehensive view of semantic similarity than either associative strength or PMI alone. The authors' findings suggest that word association data can predict human judgments of similarity with a high degree of accuracy, outperforming simpler measures and text-based models, especially for tasks that rely on conceptual meaning rather than linguistic co-occurrence.

Figure 3. Pearson correlations and confidence intervals for judged similarity and relatedness across seven different benchmark tasks from De Deyne et al. (2019)

Studies like those by Nelson et al. (2000, 2004) De Deyne and Storms (2008), and De Deyne et al. (2013, 2019) have shown that word associations not only reflect associative strength but also provide a window into the organization of lexical

knowledge in memory. These network models demonstrate how words are interconnected in small-world networks, where associative links mirror our everyday experiences and language use.

2.3 Comparison with Distributional Models

Network-based models are not the only approach used to model the structure of the mental lexicon. As Kumar et al. (2022) highlights, distributional models have also gained significant attention in the field. These models operate under the distributional hypothesis (Firth, 1957), which suggests that words that appear in similar contexts tend to have similar meanings.

Distributional models, such as Word2Vec (Mikolov et al., 2013), learn word representations by extracting statistical patterns from large text corpora, capturing semantic similarity through co-occurrence rather than associative links. While these models excel at processing vast amounts of linguistic data, they may lack the nuanced, context-sensitive insights argued to be captured by word associations as argued by De Deyne et al. (2016). In their study, the authors present a comprehensive comparison of the two approaches to modeling human semantic knowledge: traditional distributional models based on large-scale external text corpora (External language models or E-language models) and models derived from word associations (Internal language models or I-language models), which are more closely aligned with internal cognitive representations of words.

I-language models in this study aim to approximate the mental lexicon by leveraging word association data, reflecting how humans store and process semantic information in their minds. As the authors describe, these models view language as "language as the body of knowledge residing in the brains of its speakers" (De

Deyne et al., 2016, p. 1861). While E-language models like Word2Vec (Mikolov et al., 2013), which rely on external data sources, have been highly successful in NLP tasks and are often seen as capturing word meaning in ways that approximate human cognition, De Deyne et al. (2016) questions the assumption that such models truly reflect the way humans encode semantic information, suggesting that I-language models may offer a more accurate representation of human mental processes. Their key finding is that I-language models, trained on a rich, large-scale word association dataset, consistently outperform state-of-the-art E-language models, such as Word2Vec, in predicting human judgments of word similarity and relatedness.

For the E-language models, the authors combined four different corpora to represent a wide range of human language experiences which consisted of 2.16 billion tokens and 65,632 unique word types. These models are trained on external text corpora using two different approaches: Count-Based Model which tracks how often pairs of words cooccur within a sliding window at the sentence level and the co-occurrence frequencies are transformed using Positive Pointwise Mutual Information (PMI+), and Word Embeddings (Word2Vec) model which uses a Continuous Bag of Words (CBOW) architecture. Latter learns to predict a word from its surrounding context in a sliding window. Hyperparameters such as window size (2–10) and dimensionality (100–500) were tuned to optimize model performance.

I-language models are trained on the word association data which was described in their 2019 paper as discussed above. These models were trained with two different approaches as well: A Count-Based Model which uses PMI+ to estimate association strength between cue words and responses based on the frequency of associations in the word association dataset and a Spreading Activation Model which builds on the idea of spreading activation in semantic networks

(inspired by Collins and Loftus, 1975). When a word is presented, it activates the corresponding node in the semantic network and spreads to related nodes. A random walk is used to model the spread of activation across the semantic graph. A damping parameter (α) controls how much weight is given to shorter vs. longer paths in the graph, with shorter paths receiving more weight. This "decay" parameter is crucial in preventing the rapid activation of the whole network by limiting its spread. Following the iterative procedure in (Newman, 2010), the spreading activation is iterated for different path lengths, and a PMI+ transformation is applied to the resulting random walk graph.

(2) $$\mathbf{G_{rw}} = \sum_{r=0}^{\infty}(\alpha\mathbf{P})^r = \mathbf{I} - \alpha\mathbf{P}^{-1}$$

The models were evaluated on a set of standard human similarity and relatedness datasets as in their 2019 study as well as a new remote triads task where participants select the most related pair out of three nouns. The triads were constructed randomly, focusing on weakly related word pairs to avoid simple grouping heuristics. The cosine similarity between word vectors was used as the similarity measure for each of the models.

For the Remote Triads Task, the models were evaluated by correlating their similarity predictions with human judgments for each triad. The authors tuned the models by adjusting hyperparameters (such as window size and vector dimensionality) and evaluated their predictions against the human judgment datasets. The models were evaluated based on their Spearman rank order correlations with human similarity and relatedness judgments from the datasets. Also, the I-language models were compared with E-language models across various tasks to see how well they captured human semantic knowledge. The best-performing hyperparameters were identified for each model, and their average correlation scores were computed.

Key parameters they found were a window size of 3 for the count model while 7 for Word2Vec. Word2Vec performed best with 400 dimensions and the damping parameter for the random walk model was set to 0.75.

The authors also performed additional control experiments to further investigate the findings. They restricted the I-language model to only use the first associate from the word association data (G1), rather than the full set of three responses (G123). This showed that using multiple responses (G123) resulted in significantly better performance. The performance of the new word association dataset (G123) was compared with older datasets like the Edinburgh Association Thesaurus (EAT; Kiss, Armstrong, Milroy & Piper, 1973) and the University of South Florida (Nelson et al., 2004) word association norms. The new dataset outperformed the older datasets in predicting human similarity judgments.

Table 2. Spearman Rank Order Correlations Between Human Relatedness and Similarity Judgments (based on G123), and the Predictions From All Four Models (De Deyne et al., 2016)

			Text Corpus		Word Associations	
Data set	*n*	*n(overlap)*	Count	*word2vec*	Count	Random Walk
WordSim-353 Related	252	207	.67	.70	.77	.82
WordSim-353 Similarity	203	175	.74	.79	.84	.87
MTURK-771	771	6788	.67	.71	.81	.83
SimLex-999	998	927	.37	.43	.70	.68
Radinsky2011	287	137	.75	.78	.74	.79
RG1965	65	52	.78	.83	.93	.95
MEN	3000	2611	.75	.79	.85	.87
Remote Triads	300	300	.65	.52	.62	.74
mean			.67	.69	.78	.82

The authors demonstrate that the word association-based models provide a more accurate reflection of human cognition because they account for semantic links that are not always present in external linguistic corpora. For instance, they note that certain strong associations, such as “yellow” and “banana,” may not co-occur

frequently in corpora due to pragmatic language use (since it is common knowledge that bananas are yellow, this modifier is often omitted in natural speech), yet these associations are highly salient in the mental lexicon. This highlights one of the key limitations of E-language models: while they are excellent at handling large volumes of text, they may overlook important associations that are less frequent or contextually specific. Spreading activation models account for indirect associations that may not be captured by simple co-occurrence statistics. For example, a word like "tiger" may activate associations like "zebra" or "claws," even though these words may not frequently co-occur in texts. The random walk model, with its emphasis on indirect links, proves to be especially powerful in predicting human similarity judgments, further demonstrating that the I-language approach offers a more cognitively plausible account of semantic memory. The Remote Triads Task forces participants to rely on subtler semantic relationships, which is particularly challenging for distributional models based on co-occurrence. The superior performance of I-language models in this task further underscores their ability to capture the kinds of semantic judgments that are more reflective of human cognition.

In 2024, Cabana et al. (2024) extended the scope of word association research by providing free association norms for Rioplatense Spanish, a variant spoken in Argentina and Uruguay. The study aimed to explore the organization of semantic knowledge in a Romance language, drawing comparisons with previous studies in Dutch and English (De Deyne and Storms, 2008; De Deyne et al. 2013, 2019). The Small World of Words (SWOW) project had previously collected large-scale association norms in other languages as discussed above, but this paper marks the first attempt at providing similar norms for Rioplatense Spanish, offering unique insights into how language variant influences associative structures.

The dataset consisted of responses collected from 67,525 participants across Argentina and Uruguay, with cues covering a wide array of categories such as nouns, verbs, adjectives, and adverbs. This scale makes it the most extensive association dataset available for Spanish, and the first of its kind for Rioplatense Spanish. Participants were asked to provide three associations to each cue word. This methodology of eliciting multiple associations reflects earlier work by De Deyne et al. (2019). Similar data collection and preprocessing steps to De Deyne et al. (2019) were followed such as using a snowballing procedure to select the cue words. Starting with 1,000 high-frequency words, they added new cue words based on the responses generated by participants. This allowed the dataset to grow dynamically and ensure that the most commonly associated words were included as cues in subsequent iterations.

In terms of the response structure, only the strongest component of the association network –those words that appeared both as cue words and as responses– was retained for analysis. The network of word associations displayed a classic small-world structure, characterized by a few highly connected nodes (hubs) and many weakly connected nodes. This is consistent with previous findings in Dutch and English (De De Deyne and Storms, 2008; Deyne et al., 2013, 2019), where a small number of central words are responsible for linking large clusters of less frequently connected words.

In terms of word frequency and centrality, the most frequently given responses in the dataset were words associated with basic human needs and daily life, such as "agua" (water), "comida" (food), and "amor" (love). These results mirror previous studies in other languages, where high-frequency words often revolve around essential and familiar concepts.

The study also revealed the prevalence of gender effects in associative patterns, although the phi coefficient for gender congruency between cue and response was relatively weak ($\varphi = .23$). Feminine cues tended to elicit more feminine responses, while masculine cues triggered more masculine responses, a finding that aligns with similar patterns observed in other Romance languages (Boroditsky et al., 2003).

To validate the quality of the Rioplatense Spanish norms, the authors compared the dataset with previous word association norms in Iberian Spanish (Fernández et al., 2012). The comparison showed a high degree of overlap (about 85%) in the cue words, but notable differences in response patterns. For example, certain words that are region-specific, such as "auto" for "car" in Rioplatense Spanish versus "coche" in Iberian Spanish, highlighted the need for separate association norms for different Spanish dialects. This finding is consistent with previous studies emphasizing the importance of cultural and linguistic context in shaping associative networks (Armstrong et al., 2016).

One of the key findings of the study was the predictive power of word association data for psycholinguistic tasks such as lexical decision tasks (LDT) and semantic similarity judgments. By calculating centrality measures, such as in-degree centrality, the authors demonstrated that word association norms could explain unique variance in reaction times for LDT. Specifically, words with higher in-degree values were associated with faster decision times, reflecting their centrality in the mental lexicon. This result replicates findings in Dutch and English, where central words have been shown to facilitate quicker recognition and categorization (De Deyne et al., 2019).

In this study, Cabana et al. (2024) compared the performance of three embedding models to word association-based models, such as spreading activation and random walks, derived from their association norms. The comparison focused on how well each approach predicted human judgments of word similarity, which were assessed through tasks such as semantic similarity ratings and a remote triads task. The models are trained on large text corpora, where they capture the co-occurrence statistics of words in sentences or windows of text, generating continuous vector representations of words in a high-dimensional space.

First model used in this study was Word2Vec. This model comes in two variants: Continuous Bag of Words (CBOW) and Skip-Gram. It learns to predict either a word from its context (CBOW) or its context from a given word (Skip-Gram). Word embeddings are created such that words with similar contexts are mapped to nearby points in the vector space.

Second model was GloVe (Pennington et al., 2014), which builds on the co-occurrence matrix by calculating the ratio of co-occurrences between word pairs, emphasizing that certain words co-occur more frequently than others and that this difference in frequency reflects word similarity.

Lastly, FastText (Bojanowski et al., 2017) was used which is a variant of Word2Vec that also takes into account subword information (character-level n-grams), making it more robust for morphologically rich languages like Spanish.

One of the key tasks for comparing word associations with embeddings was predicting human similarity judgments for various word pairs. The researchers used standard datasets for Spanish, such as the Moldovan et al. (2015) dataset, the Spanish version of the MultiSimlex corpus (Simlex-ES), and a new relatedness dataset collected for Rioplatense Spanish. Cabana et al. (2024) evaluated the performance of

the different models by calculating the Spearman rank-order correlations between model predictions and human similarity judgments.

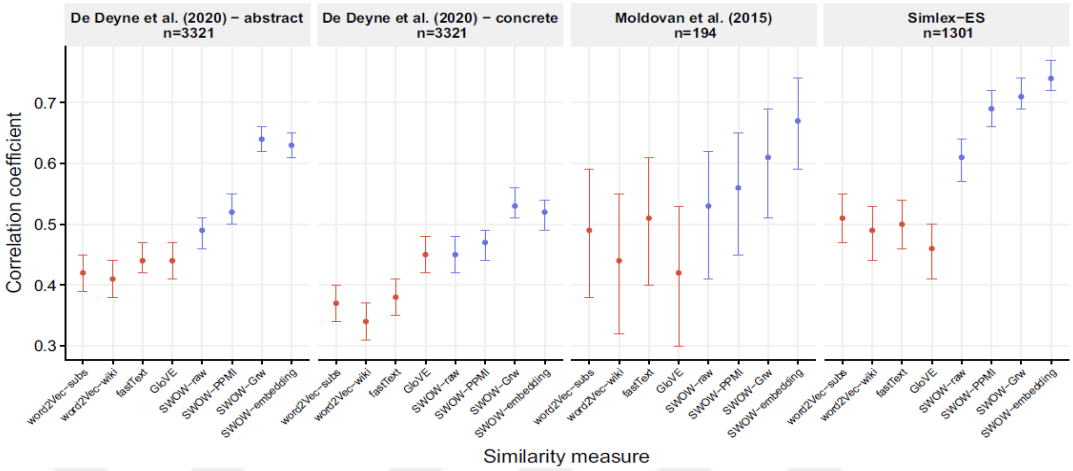


Figure 4. Correlation between human similarity judgments and similarity scores obtained from word embedding algorithms and SWOW-RP data. (Cabana et al., 2023)

The results consistently showed that SWOW-based models outperformed corpus-based word embeddings in predicting human judgments of both similarity and relatedness. The raw association data (i.e., simple counts of word associations) already performed better than corpus-based embeddings like Word2Vec and FastText. However, when more sophisticated models like random walks and positive pointwise mutual information (PPMI) were applied to the association data, their performance increased even further. These transformations allowed the models to account for indirect and more complex associations between words, providing a more cognitively plausible account of semantic memory. For the remote triads task the association-based models again outperformed the distributional models, underscoring their ability to capture subtler semantic relationships.

Another paper, Thawani et al., 2019, proposes a new intrinsic evaluation method for word embeddings using word association data from the Small World of

Words (SWOW) dataset. The study uses a subset of the SWOW dataset, which consists of cue-response pairs. The dataset contains 8,500 cue words and nearly a million cue-response pairs, with each cue associated with up to three responses from participants. For each cue-response pair (C-R), the authors used the number of participants who provided the response (R123) as the association strength between the cue (C) and the response (R). The final SWOW-8500 dataset was filtered to include only cue-response pairs with an association strength (R123.Strength) of at least 0.2, meaning that at least 20% of participants selected the response as one of the top three for the given cue.

The main task proposed for intrinsic evaluation is based on how well word embeddings can predict human word associations from the SWOW dataset. For a given cue word, word embeddings are used to predict the most likely responses. This is operationalized by using cosine similarity: given the vector representation of a cue word, the embeddings calculate which other word vectors are closest in the vector space. These closest vectors are considered the model's predictions for human associations. For each cue, the number of correct predictions is compared to the actual responses provided by participants, and performance is evaluated using traditional metrics such as precision, recall, and F1 score.

The authors evaluated six pre-trained word embeddings: Word2Vec (Skip-Gram) (trained on Google News), GloVe (trained on Wikipedia 2014 and Gigaword 5), FastText (trained on Common Crawl), ConceptNet Numberbatch (cased on a knowledge graph) (Speer et al., 2017), Baroni and Lenci (2014)'s Count-Based Embeddings (created using a large count matrix), Random Baseline (a baseline where word vectors are initialized with random values). The embeddings were evaluated on their ability to predict top-k responses for each of the 6,481 cues present

in both the vocabulary and the SWOW-8500 dataset. The value of k, the number of responses predicted, was variable and corresponded to the number of responses given for each cue in the SWOW-8500 dataset. This SWOW-based evaluation was then compared to traditional word similarity benchmarks, including datasets like WordSim-353, SimLex-999, and MEN.

In addition to intrinsic evaluation, the authors also conducted extrinsic evaluations by testing the performance of embeddings on downstream NLP tasks such as: Sentiment Analysis, Named Entity Recognition (NER), Part-of-Speech (POS) Tagging, Natural Language Inference (NLI) and Chunking. For downstream tasks, the authors used the VecEval framework, applying simple feed-forward neural networks with one or two hidden layers, rather than more complex architectures like LSTMs (Hochreiter and Schmidhuber, 1997; Jozefowicz et al., 2016) or Transformers (Vaswani, 2017), to isolate the impact of the word embeddings themselves.

The ConceptNet Numberbatch embeddings outperformed other models in both intrinsic (SWOW-8500 task) and extrinsic (downstream task) evaluations. Authors argue that this is likely due to ConceptNet's reliance on a knowledge graph that better captures human semantic relationships. SWOW-8500 demonstrated a strong correlation with existing word similarity tasks, but with the added advantage of larger scale and better statistical reliability. The authors noted that SWOW-8500 had narrower confidence intervals, making it a more reliable evaluation measure. For instance, the confidence interval for error rates in SWOW-8500 was three times narrower than the best-performing traditional similarity dataset. This statistical robustness adds credibility to SWOW-8500 as a reliable benchmark for intrinsic evaluation. For downstream tasks, embeddings that performed well on the SWOW-

8500 task, such as ConceptNet and FastText, also excelled in real-world applications like sentiment analysis and NER, suggesting that the SWOW-based evaluation is a good predictor of practical performance.

2.4 Recent applications using LLMs

While traditional word embedding models such as Word2Vec and GloVe have been instrumental in modeling semantic knowledge and benchmarking against human judgments, advancements in language modeling have brought about more sophisticated approaches. Transformer-based models, like those used in large language models, represent a significant leap forward in distributional semantics. These models are now increasingly used to simulate cognitive experiments, providing an opportunity to compare their outputs directly with human behavior in tasks. For example, Aher et al. (2023) introduces a novel methodology called a *Turing Experiment (TE)*, which evaluates how well large language models (LLMs), such as GPT (Brown et al., 2020; OpenAI, 2023), can simulate human behavior. This paper's core aim is to examine LLMs' ability to replicate findings from classic human subject studies across various domains, including behavioral economics, psycholinguistics, social psychology, and collective intelligence.

A TE differs from the traditional Turing Test, which involves simulating a single human. Instead, TEs simulate a diverse population of participants, as would be done in actual human subject research. The goal is to determine how well LLMs can replicate human behavior at scale, providing insight into the fidelity and limitations of these models when simulating humans. The Turing Experiment framework requires LLMs to simulate multiple individuals rather than a single subject, which is more reflective of real-world human subject studies. The researchers argue that this

approach is critical for understanding whether LLMs can simulate the diversity of human behaviors observed in various fields. The paper introduces a framework for conducting TEs using LLMs. A Turing Experiment uses two types of inputs: participant details and experimental conditions. For example, in one TE, the researchers represented demographic diversity by assigning names associated with different racial groups to the simulated participants. These names were derived from the 2010 U.S. Census Data (U.S. Census Bureau, 2010) and included 100 of the most common surnames for each of five racial groups, such as African American, Asian, and Native American populations. Titles like “Mr.,” “Ms.,” or “Mx.” were also used to indicate binary or non-binary gender identities. By incorporating these names and titles as inputs, the experiment modeled demographic variation, enabling analyses of how the simulated participants' responses differed across these identity markers.

Outputs are synthetic records that simulate human subject studies and their outcomes, allowing for comparison with actual human studies to evaluate how well the LLM simulates real human behavior. The experiments are zero-shot, meaning that the LLM is not explicitly trained on the specific experimental data but is expected to generate results based on general knowledge from its training data. The authors conducted four TEs using LLMs to simulate classic experiments:

- Ultimatum Game (Behavioral Economics): The experiment involves fairness and decision-making, where one participant proposes how to split a sum of money, and the other can accept or reject the offer. The LLM simulation accurately replicated gender differences in behavior.
- Garden Path Sentences (Psycholinguistics): Participants judge whether sentences are grammatical or not. Garden path sentences are sentences that

initially lead the reader to an incorrect interpretation. The LLM simulations were able to replicate the challenges humans face with such sentences.

- Milgram Shock Experiment (Social Psychology): The classic experiment where participants were instructed to administer electric shocks to another person under the authority's orders. The LLM simulation reproduced the trend of obedience observed in Milgram's original study.
- Wisdom of Crowds (Collective Intelligence): This experiment examines the phenomenon where the average answer of a group is often more accurate than most individuals' answers. The LLM simulations, especially the larger models, exhibited a “hyper-accuracy distortion” where participants provided unrealistically correct answers, suggesting an issue with the model's overalignment.

The first three TEs successfully replicated findings from human studies, showing that LLMs can simulate certain aspects of human behavior. For instance, in the Garden Path sentences, the following prompt was given:

(3) Ms. Olson was asked to indicate whether the following sentence was grammatical or ungrammatical.

Sentence: While the student read the notes that were long and boring blew off the desk.

Answer: Ms. Olson indicated that the sentence was _____

The results of the simulation support the theory that garden path sentences are more likely to be misinterpreted as ungrammatical compared to the control sentences as the validity rates from all the models are high. Overall, the study demonstrates that LLMs can simulate various aspects of human behavior across different experimental conditions. The findings suggest that while LLMs can replicate many human

behaviors, they also introduce distortions, such as hyper-accuracy, that can impact their usefulness in simulating real-world scenarios. Moreover, the paper acknowledges several limitations such as the possibility that the LLMs were trained on large datasets that might include descriptions of these experiments. The zero-shot nature of the TEs is not fully guaranteed, and the models may be influenced by prior exposure to these studies.

Two other papers also simulate linguistic experiments with LLMs. Firstly, the paper by Shin and Song (2023) explores the ability of neural language models (particularly BERT (Devlin et al., 2019)) to replicate human-like language processing, focusing on Negative Polarity Items (NPIs). The study investigates how well BERT captures syntactic and semantic constraints involved in NPI licensing, using psycholinguistic experimental paradigms. The authors conduct two experiments. “Syntactic Licensing and Grammatical Illusion” experiment is to test whether BERT captures the hierarchical relationship between NPIs and their licensors and to explore the possibility of grammatical illusion, which occurs when a human processor mistakenly accepts an incorrect NPI licensor. The stimuli were adapted from previous psycholinguistic experiments and included sentences with valid, invalid (illusory), or absent licensors. For example:

(4) a. Valid: "No scandals have ever generated a large public outcry."

b. Invalid: "*The scandals that no prominent politicians met ever generated a large public outcry."

c. No licensor: "*The scandals that the prominent politicians met ever generated a large public outcry." (Shin & Song, 2023)

In this experiment, BERT was tested using a masked language model approach, where the NPI “ever” was masked in sentences, and the model’s ability to

predict the correct word in the masked position was measured using surprisal values. Surprisal is the negative log probability of a word given its context and reflects processing difficulty. One-way ANOVA was used to compare surprisal values across conditions. Post hoc tests were performed for pairwise comparisons. In the second experiment, namely “Semantics and Scale of Licensing Strength”, purpose was to examine BERT’s sensitivity to the semantic strength of NPI licensors (e.g., "no," "few," "only"), which differ in their ability to license NPIs based on their semantic properties. Sentences with different types of licensors (e.g., classic negation "no," minimal negation "few," and zero-negation "only") were used. The study focused on whether BERT could differentiate between stronger and weaker licensors. As in Experiment 1, the cloze test method was used, and surprisal values were measured for both the licensors and NPIs in different sentence conditions. ANOVA and Tukey post hoc comparisons were used to analyze the differences in surprisal values across conditions.

Overall, the authors found that BERT successfully captured the hierarchical relationship between NPIs and licensors, showing similar sensitivity to syntactic structure as humans. The model was able to distinguish between valid licensors and invalid ones but showed some susceptibility to grammatical illusions, similar to humans (Xiang et al., 2009; Parker and Phillips, 2016). BERT also demonstrated a significant ability to differentiate between licensors of varying semantic strength, with the strongest licensors (e.g., "no") resulting in the lowest surprisal values while BERT’s performance with the licensor "only" showed inconsistencies, possibly due to the complexity of its semantic and pragmatic roles (Zwarts, 1996; Giannakidou, 1997; Xiang et al., 2013; Chatzikontantinou et al., 2015). The study found that BERT’s processing patterns for NPIs were strikingly similar to those of humans, both

in its success and its limitations (e.g., grammatical illusion). This suggests that BERT can serve as a useful tool for studying human language processing.

Secondly, the paper by Kauf et al. (2023) investigates whether LLMs possess generalized event knowledge similar to humans, specifically focusing on how LLMs distinguish between plausible and implausible events. The study evaluates five LLMs (BERT, RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), GPT-J (Wang & Komatsuzaki, 2021), MPT (The MosaicML NLP Team, 2023)) using three datasets containing sentences with event descriptions. These sentences are categorized as possible, impossible, likely, and unlikely, with variations achieved by altering agent-patient relationships or the plausibility of the events described. The authors used log-probabilities (surprisal values) from the models to assess their ability to correctly rank plausible events higher than implausible ones. The paper also analyzed human judgments to compare how closely LLM behavior mimics human responses.

The authors found that LLMs, like humans, generally assign higher likelihoods to possible over impossible events. However, distinguishing between likely and unlikely events remains a challenge for these models, where their performance is less consistent than when judging between possible and impossible events. The models performed well when the distinctions were based on syntactic or structural clues but struggled when the difference was semantic, such as in cases of subtle implausibility. In addition, the study revealed that LLMs are influenced by surface-level sentence features like word frequency rather than just event plausibility.

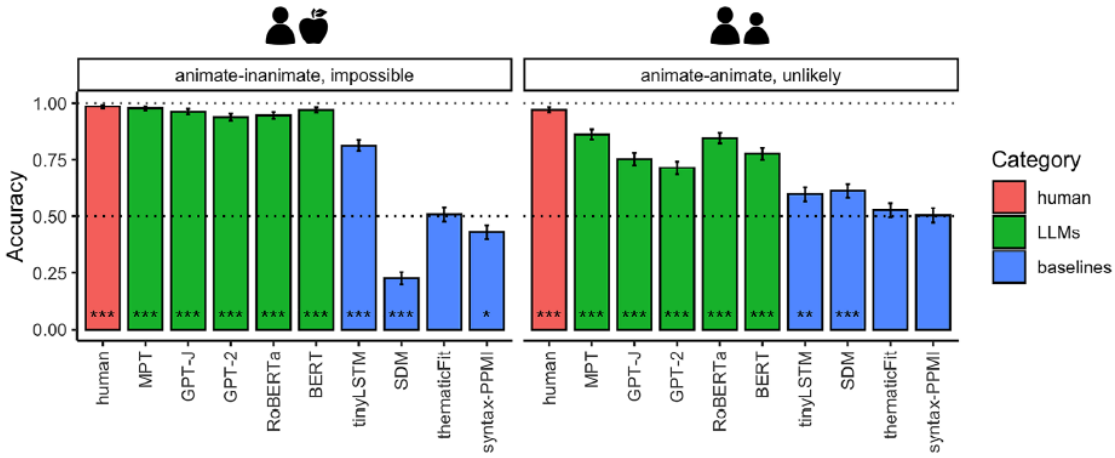


Figure 5. Main results of Dataset 1 sentences from Kauf et al. (2023)

2.5 Conclusions and relevance to the present study

Word association tasks have significantly advanced our understanding of semantic memory, revealing how associative strength and set size reflect the structure of the mental lexicon. Studies by Nelson et al. (2000, 2004) and De Deyne and Storms (2008) De Deyne et al. (2013) have demonstrated the value of both single and multiple-response paradigms in capturing the dynamic nature of word associations. Recent research has highlighted the potential of word association data to enhance computational models, providing more cognitively plausible representations of semantic relationships. As large language models begin to simulate human cognition, word associations remain a critical tool for bridging human and machine understanding of language.

Building on these insights, the present study compares human word associations with those generated by large language models using a continuous association experiment. Foundational concepts like associative strength and set size (Nelson et al., 2000, 2004) and continuous response paradigms (De Deyne and Storms, 2008) inform the design of the study, enabling a richer analysis of primary and secondary associations. Graph-based approaches, such as those used in network

studies (De Deyne et al., 2013, 2019), align closely with the current methods, where graph similarity metrics are applied to evaluate structural patterns in human and model-generated associations.

By building on these foundational and modern approaches, my study extends the LLM simulation research discussed in this chapter (e.g., Aher et al., 2023; Kauf et al., 2023; Shin & Song, 2023) into the domain of word associations. In doing so, it not only evaluates the extent to which LLMs replicate human word association patterns but also offers valuable insights into the processes underlying human semantic organization and memory.

CHAPTER 3

METHODOLOGY

This chapter outlines the methodological approach used in this study to investigate word associations in both human participants and large language models. The study aims to compare the patterns of word associations generated by humans with those produced by state-of-the-art LLMs in Turkish. The methodology is divided into three main sections: the human experiment, the LLM simulations, and the comparison criteria used to evaluate the results from both groups.

The human experiment section describes how the free association task was designed and administered to participants, detailing the recruitment, stimuli selection, and data collection procedures.

The second section outlines how the same experiment was simulated using several large language models, including GPT-4o mini (OpenAI, 2024), Trendyol LLM v1.8 (Trendyol, 2024), Cosmos LLaMA Instruct-DPO (Kesgin et al., 2024). Each model was prompted to generate word associations under conditions designed to mimic those of the human participants, allowing for a direct comparison between human responses and machine-generated responses.

Finally, the comparison section details the quantitative and qualitative methods used to analyze and compare the human and LLM data. Various metrics were employed, including cosine similarity, Jaccard similarity, and graph-based similarity metrics, to assess the structural and associative similarities between the two datasets. These analyses provide a framework for evaluating how closely LLMs replicate human-like patterns of free word associations.

By combining an experimental design with a comparison framework, this methodology aims to shed light on the capabilities and limitations of LLMs in simulating human associative linguistic behavior, particularly in the Turkish language.

3.1 Human experiment

The human experiment was conducted to investigate the natural patterns of word association among Turkish speakers. Participants were tasked with responding to a series of cue words by providing up to three distinct word associations. This experiment aimed to capture a range of associative strengths, from the most immediate and automatic associations to more secondary and less prominent ones.

3.1.1 Experiment design

The word association experiment was conducted using the PCIbex Farm platform (Zehr and Schwarz, 2023), which allows for remote administration of experiments in a web-based interface. Participants accessed the task via an internet link, which allowed them to participate at their convenience, reducing scheduling constraints and making it possible to reach a larger sample size.

The experiment was structured as a word association task where participants were prompted with a series of words and instructed to respond with the first three words that came to mind for each cue word. This design aimed to capture both the immediate and secondary associations participants make, reflecting both strong and weak associations which offers a richer insight into both direct and indirect links in semantic memory. The experiment was ethically approved by the Boğaziçi University Ethics Committee (Decision number: SBB-EAK 2023/45), and

participation was incentivized by offering students extra credit in a Linguistics course. Participants were fully informed about the nature of the study and their right to withdraw at any point. Informed consent was obtained from all participants before the experiment began. A copy of the informed consent form and a demonstration of the experiment interface can be found in Appendix A and B, respectively.

3.1.2 Participants

A total of 381 participants took part in the experiment. These included 230 women, 144 men, 6 non-binary individuals, and 1 participant who chose not to disclose their gender. The participants were recruited from the Boğaziçi University Linguistics Department, as well as from the researcher's social circle. The mean age of the participants was 21.24 years, with most participants being undergraduate students. The relatively young age of the participant pool is due to the university setting, and while this introduces certain demographic limitations, the diversity of linguistic and cognitive experiences among the participants ensures that the data remains relevant to broader analyses of word associations. No participants were excluded from the dataset, as all participants completed the task, providing a comprehensive dataset for analysis.

3.1.3 Materials

The stimuli used in this experiment were selected from the AnlamVer dataset (Ercan and Yıldız, 2018), a resource specifically designed to evaluate Turkish semantic models by distinguishing between word similarity and relatedness. The AnlamVer dataset was developed to provide a robust benchmark for distributional semantic

models in Turkish, incorporating various linguistic and semantic features unique to the language such as its complex morphology.

The AnlamVer dataset is a collection of Turkish word pairs, annotated with human judgments of both word similarity and word relatedness. It was created to evaluate the performance of distributional semantic models, specifically targeting the unique linguistic features of Turkish, including its rich morphology and agglutinative nature. This dataset was designed to balance word pairs across several attributes, such as word frequency, morphology, concreteness, and semantic relation types (e.g., synonymy, antonymy, hypernymy). These considerations ensure that the dataset can serve as a reliable evaluation tool for Turkish. Each of the 500 word pairs in the dataset was scored by 12 human annotators, with two distinct scores assigned to each pair: one for similarity and another for relatedness. This dual scoring was motivated by the similarity-relatedness confusion in the literature and previous evaluation benchmarks. A comprehensive discussion of the limitations to datasets caused by the lack of a distinction between relatedness and similarity are provided by Hill et al. (2015). To demonstrate the distinction between relatedness and similarity, Hill et al. (2015) gives the example of the pairs [car, bike] and [car, petrol]. It is said that cars are related to petrol but, cars are similar to bikes. Cars and bikes are intuitively comparable because they have similar physical characteristics, have the same purpose, or belong to a class of objects that is well defined. On the other hand, because they regularly appear together in language and space, cars and petrol are connected. To maintain this difference, the creators of the AnlamVer dataset got scores for both types of relations in their experiments. The similarity score reflects how alike the meanings of two words are, while the relatedness score captures how frequently or contextually the words are associated with one another. For example,

kırmızı (red) and *gül* (rose) are highly related in meaning but their meanings are not similar, leading to a high relatedness score but a lower similarity score.

The initial pool of words for AnlamVer was drawn from the "Türkçe Kelime Normları" (TKN) dataset (Tekcan & Göz, 2005), which contains 600 Turkish words (50% concrete, 50% abstract) with annotated concreteness values. In AnlamVer, words from this dataset were selected based on their concreteness ratings, ensuring a mix of abstract and concrete terms. In addition, 99 new derivational words were manually added to the candidate pool to enrich the dataset with more morphologically complex words.

After the initial candidate selection, the words were filtered and balanced across several key attributes, including:

- Frequency: Words were balanced according to their frequency in the Boun Corpus (Sak et al., 2011), a large-scale Turkish language corpus consisting of approximately 3.2 million token types. Words were chosen to represent five distinct frequency bands: 0-32, 32-320, 320-3200, 3200-32000, and words with a frequency greater than 32,000. This way, the frequency distribution within the dataset was carefully controlled to include both high-frequency and low-frequency words, as well as out-of-vocabulary (OOV) words and made-up words. This ensures that the dataset captures a wide range of linguistic contexts, from everyday language to rare or novel terms.
- Morphology: Words with multiple derivational and inflectional morphemes were specifically included to reflect the complexity of Turkish morphology.
- Concreteness: Words were chosen to reflect a balance between highly concrete and highly abstract terms. This variation is important for testing the

generalizability of semantic models across different types of conceptual information.

Once the word pool was finalized, the next step involved constructing word pairs. The dataset was designed to include 500 word pairs, with each word occurring in multiple pairs (2-5 times). The pairs were selected based on several semantic relations, including synonymy, antonymy, hypernymy, and meronymy. Each word pair in the AnlamVer dataset was manually annotated for its semantic relationship, with pairs representing various types of relations such as synonymy (e.g., *otomobil* – automobile vs. *araba* – car), antonymy (e.g., *sıcak* – hot vs. *soğuk* – cold), hypernymy (e.g., *kuş* – bird vs. *serçe* – sparrow), and meronymy (e.g., *kuş* – bird vs. *tüy* – feather). This construction process ensures that the dataset captures a wide range of semantic relationships, allowing for comprehensive evaluations of model performance.

For our study, we selected a total of 240 unique words from the AnlamVer dataset (see Appendix C for the full list of stimuli words). The word selection was done randomly. First, all the distinct words from the word pairs in the AnlamVer dataset were listed, and then random word selection was made from this list. The selected words were not subjected to preprocessing; both root forms and inflected forms were preserved. These words were divided into 12 groups with 20 words in each group. The selection process leveraged the rich linguistic and semantic diversity of the AnlamVer dataset, which provided a comprehensive foundation for ensuring balance across the stimuli. The aim was to select words that represented a broad spectrum of Turkish vocabulary, including both high-frequency and low-frequency words, simple and complex morphological structures, and a wide range of concrete

and abstract meanings. This was essential for providing a comprehensive set of words capable of testing word associations under varying conditions.

Word frequency is a crucial factor in human word processing speed (Brysbaert and New, 2009). The AnlamVer dataset balances word frequency across several categories, ranging from very common, high-frequency words to rare, potential out-of-vocabulary words. This balance in word frequency allows the study to assess how human participants and large language models handle both frequent and infrequent words. High-frequency words are typically encountered more often in everyday language, making them more readily accessible in memory (Brysbaert & New, 2009). In contrast, low-frequency and OOV words pose accessibility challenges for both humans and LLMs as low-frequency words generate weaker connections in memory. Rare, out-of-vocabulary words are essential for testing how participants and models handle words that they may not have encountered before. By including these challenging words, we aimed to explore the limits of word association performance in both humans and LLMs. In large language models, high-frequency words are better represented in training data, while low-frequency words may lack sufficient exposure, leading to less accurate associations. This idea was used by Vecchi et al. (2017) to assess models' phrase-level creative potential (generalization power). The main idea is that humans have the ability to understand the intended meaning of a word, even if it sounds strange to them when they hear it for the first time. This distribution is essential for understanding the robustness of word association performance across different levels of word familiarity.

Turkish is a morphologically rich language where root words can be extended through a series of affixes to create more complex words. The AnlamVer dataset reflects this complexity by including both root words and words with various degrees

of derivation and inflection. Words were categorized based on whether they were in their base (root) form or had undergone one or more derivational and inflectional transformations. This inclusion allows for the testing of how both human participants and LLMs respond to words of varying morphological complexity. For instance, a simple root word like *maymun* (monkey) contrasts with more morphologically complex words like *maymungiller* (family of monkeys), which contain derivations and inflections. A range of inflected forms is included in the stimuli, allowing the study to test how participants and models handle grammatical variations such as tense (eg., *saldırmaktaydılar* – they-were-attacking), number (eg., *kitaplıklar* – bookshelves), and possession (eg., *delilim* – my-evidence). This balance ensures that the study can assess the ability of participants and models to recognize and associate words with different levels of grammatical and morphological complexity. As the morphological complexity increases with the addition of derivational morphemes, associations may become more morphologically complex. For example, the form of the cue word *kitaplar* "books" may evoke similar forms in associations like *masalar* "tables" or *defterler* "notebooks". Morphologically rich languages like Turkish also require the participant to process not only the base meaning of the word but also the additional layers of meaning provided by the affixes. This can lead to more varied associations, as participants must integrate the meaning of the affixes with the base word. Traditional language models that rely on word-level embeddings may struggle with rich morphology in Turkish, as they treat each inflected or derived form as a separate word. This can lead to fragmented and less coherent representations for derived words, particularly if those forms are less frequent in the training data. More recent models like FastText (Bojanowski et al., 2017) or BERT (Devlin et al., 2019), which account for subword information, perform better by capturing the shared

meaning across different forms of the same root. However, even these models might struggle with the full range of Turkish morphological variation, especially in cases of complex derivations.

Words in the AnlamVer dataset are also balanced based on their concreteness. Concrete words tend to represent tangible objects (e.g., *araba* – car), while abstract words refer to concepts or emotions (e.g., *mutluluk* – happiness). The AnlamVer dataset includes a range of words from highly concrete to highly abstract, ensuring that both types are well represented. This balance of concrete and abstract words was maintained in this study as well to test how different levels of abstractness affect word associations. Abstract words rely more on cultural and personal and sensory experiences (Borghi et al., 2017) which has been argued to be captured better by word association models than purely distributional models (De Deyne et al., 2021). LLMs tend to perform better with concrete words, as their associations are more likely to be grounded in co-occurrence patterns found in large corpora. Abstract concepts in LLMs models often require more sophisticated handling, as they rely less on direct co-occurrence and more on nuanced semantic relationships that are harder to capture in distributional models (Liao et al., 2023). Including both types of words in the stimuli allows for a comprehensive evaluation of how participants and models handle different levels of word concreteness.

The words selected for this study from the AnlamVer dataset reflect a diverse range of linguistic and semantic attributes. This ensures that the study can explore word associations across different frequency levels, morphological complexities, and semantic relationships. This balanced selection lays the foundation for meaningful comparisons between human and LLM word associations in later chapters.

3.1.4 Procedure

Upon accessing the experiment, participants were provided with instructions detailing the task. They were informed that they would see a series of words on the screen, one at a time, and that they should respond with the first three words that came to mind for each word.

Stimuli words were randomly divided into 12 groups, each containing 20 words. Participants were randomly assigned to one of these groups, and the words within the assigned group were further randomized in their presentation order, ensuring that each participant received a unique sequence of stimuli. The task was self-paced, allowing participants to move on to the next word after entering their three responses for the current cue. This flexible pacing was designed to reduce stress and ensure that participants could complete the task comfortably within the estimated time frame of 20 minutes.

The decision to collect three responses per cue word was made to capture both strong, primary associations as well as weaker, secondary associations. Research has shown that secondary and tertiary responses often reveal more nuanced relationships between words, which are not immediately apparent in single-response tasks (De Deyne & Storms, 2008). By gathering multiple responses, this study aimed to capture a broader spectrum of associative strengths, providing a more detailed picture of participants' semantic networks.

3.1.5 Data collection and preprocessing

All participant responses were collected automatically via the PCIbex platform and stored in a secure database for further analysis. The preprocessing of the collected

data was critical to ensuring the accuracy and consistency of the dataset, particularly given the complexities of the Turkish language.

The responses were normalized by stripping punctuation and converting all text to lowercase. This ensured that superficial variations in formatting did not affect the analysis. For instance, the word *kitap* (book) would be treated the same regardless of whether it was capitalized as *Kitap* or typed in lowercase. Following normalization, morphological analysis was performed to retain the full morphologically complex forms of all cues and responses. This step was essential for analyses that required detailed morphological information, such as examining suffix patterns or morphological complexity, and ensured that the data preserved its rich morphological structure. For these analyses, the unstemmed versions of the responses were stored in one file for direct comparison.

Subsequently, stemming was performed using the TrnlpWord library in Python (Brolin59, n.d.), a specialized tool designed for handling the rich morphology of Turkish. Stemming was crucial for reducing words to their base forms, particularly in cases where participants provided inflected or derived forms of the cue words to allow for a part of the analysis to focus on the core semantic connections between words without being influenced by morphological variations. For example, a participant might respond to the cue word *kitap* (book) with *kitaplar* (books) or *kitaplık* (bookshelf). By reducing these words to their root form, the analysis could focus on the underlying semantic connections rather than being confounded by morphological variations. Given the complexity of Turkish morphology, stemming also allowed for a more accurate comparison of word associations. The preprocessing ensured that morphological variations were handled appropriately, allowing for a more coherent analysis of participants' associative networks.

No participants or responses were excluded from the analysis, ensuring that the dataset represented the full range of associations elicited during the experiment. This complete dataset was then prepared for subsequent analyses of associative strength, similarity, and other factors that would be compared to large language model-generated responses.

3.2 LLM simulations

The LLM simulations were conducted to replicate the human word association task using large language models. By presenting the same set of stimuli to several state-of-the-art models, the study aimed to assess the extent to which these models can mimic human associative behavior. The simulations were carefully controlled to match the conditions of the human experiment, allowing for a direct comparison between human and LLM-generated associations.

The models were used without modifying any configurations or parameters, relying entirely on their default settings to preserve the integrity of the fine-tuning already performed during their training stages. To simulate the task, various prompts were tested for effectiveness. After manually examining the outputs, the prompt *".. kelimesini okuduğunda aklına gelen birbirinden farklı ilk üç kelimeyi yaz."* (Write down the first three different words that come to your mind when you read the word ....) was selected as it produced the most suitable responses. Adding "*birbirinden farklı*" (different from each other) explicitly prompted the models to avoid generating duplicate responses for a given cue.

The entire dataset used in the human experiment was run through each model 30 times, ensuring a comparable level of data generation to the human experiment, where approximately 30 participants responded to each stimulus group. Each trial

was conducted independently to ensure unbiased outputs for every run. This setup allowed for an equitable comparison of associative diversity and response patterns between humans and models.

3.2.1 Selection of LLMs

This study used transformers-based models fine-tuned specifically for Turkish. The chosen models are decoder-only transformer architectures, optimized for language generation tasks. Selection criteria focused on models capable of handling complex Turkish language tasks.

3.2.1.1 Transformer architecture

Transformers, introduced by Vaswani et al. (2017) in their "Attention is all you need" paper, are the foundational architecture for most contemporary NLP models. Transformer based models process language sequences in parallel, contrasting with earlier sequential approaches such as LSTMs (Hochreiter and Schmidhuber, 1997; Jozefowicz et al., 2016), and are highly effective for managing complex linguistic dependencies across long sequences (Vaswani et al., 2017). This parallelization is achieved through an encoder-decoder structure, with some LLMs (including those in this study) employing a decoder-only setup optimized for language generation tasks. An encoder transforms input data into a compact representation (embeddings), capturing essential features, and is typically used in tasks requiring language understanding. A decoder takes this encoded representation and reconstructs it into an output, generating sequences or making predictions, often used in tasks like translation or text generation.

Central to the Transformer's effectiveness is the self-attention mechanism (Vaswani et al., 2017), which dynamically assesses the importance of each word in relation to every other word within the input sequence. This mechanism enables the model to develop a nuanced understanding of language context, crucial for generating context-appropriate associative responses. Each word is represented through Query (Q), Key (K), and Value (V) vectors to facilitate context-driven calculations:

- Query (Q): Defines the token in focus.
- Key (K): Represents all other tokens in the sequence to establish relationships.
- Value (V): Contains the contextual content of each token.

The model calculates attention scores between each Query-Key pair, adjusting focus on the Value vectors based on relevance. Multiple attention heads capture different contextual patterns, and stacking these heads and layers enables handling of complex linguistic structures crucial for nuanced language understanding.

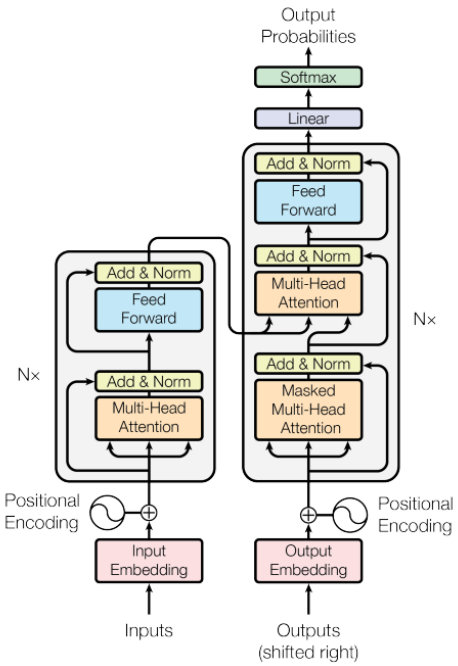


Figure 6. The transformer architecture showcasing the encoder on the left and the decoder on the right from Vaswani et al. (2017)

Some LLMs employ decoder-only Transformer architectures for tasks such as text generation and zero-shot inference. Decoder-only models are optimized to generate text by predicting each word based on previously generated words, making them ideal for autoregressive language modeling.

Decoder-only models are typically preferred in LLMs because they process sequences in a left-to-right manner, conditioning each token on prior tokens. This structure is naturally aligned with the objective of language modeling, where the model aims to predict the next word based on the preceding context (Wang et al., 2022). Unlike encoder-decoder models, which process both an input and output sequence, decoder-only models operate on a single stream of text. This setup requires fewer computational resources, reducing memory demands and facilitating the scaling to billions of parameters. As a result, decoder-only models are optimized for both performance and efficiency (Liu et al., 2018). These models simplify the fine-tuning process by focusing solely on the generative aspect. This specialization enables them to align well with task-specific requirements in text generation, such as zero-shot settings, where the model leverages its extensive pretrained knowledge to generate coherent text without additional training (Wang et al., 2022).

3.2.1.2 Fine-Tuning approaches

The large language models used in this study were pre-trained on non-Turkish data and subsequently fine-tuned to handle Turkish-specific language tasks. This approach leverages the extensive knowledge gained from general pre-training while adapting the models for Turkish through additional specialized fine-tuning. Fine-tuning techniques applied to these models, such as Direct Preference Optimization (DPO) (Rafailov et al., 2024) and Low-Rank Adaptation (LoRA) (Hu et al., 2021),

enable the models to effectively generalize across Turkish language contexts while maintaining computational efficiency.

- Low-Rank Adaptation (LoRA) (Hu et al., 2021): LoRA is a fine-tuning technique that injects a limited number of trainable parameters into pre-existing model layers. By introducing low-rank matrices, LoRA can adapt large models to new tasks without altering all parameters, making the process computationally efficient. This approach is especially beneficial for models with substantial parameter counts, as it allows effective adaptation to Turkish tasks without excessive computational demand.
- Direct Preference Optimization (DPO) (Rafailov et al., 2024): DPO aligns model outputs with human preferences by incorporating ranked feedback. During DPO training, human evaluators rank multiple responses generated by the model, and the model parameters are adjusted to prioritize higher-ranked responses. This ranking-based approach allows models to produce more relevant, coherent, and contextually accurate outputs (Rafailov et al., 2024).

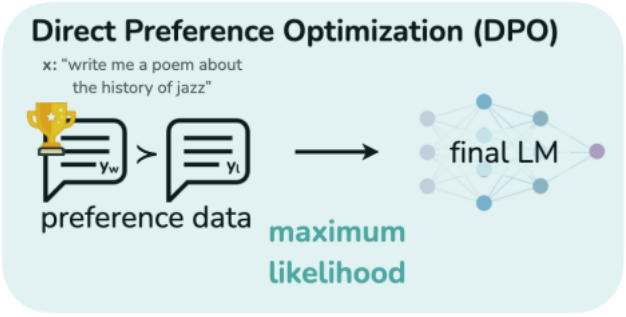


Figure 7. Illustration of the Direct Preference Optimization (DPO) Process taken from Rafailov et al., (2024)

The pre-trained models selected for this study have undergone these fine-tuning processes but were not further fine-tuned for the experimental simulations.

This ensures that the models are optimally prepared for generating Turkish language associations without additional modifications.

3.2.1.3 Selected LLMs

The selection of large language models for this study was guided by the OpenLLM Turkish leaderboard (Alhajar, 2024), which evaluates model performance across a suite of benchmark datasets. These datasets are created by Alhajar (2024) and include 5 different datasets (Hellaswag, Truthful_qa, GSM8K, Winogrande, ARC, and MMLU), covering a range of competencies from reasoning and truthfulness to knowledge and understanding of commonsense scenarios. Hence, the leaderboard offers a reliable indicator of model effectiveness in handling complex Turkish language tasks.

The top 20 models on this leaderboard were initially considered for selection, with attention to the diversity of base models (e.g., Mistral (Jiang et al., 2023), LLaMA (Touvron, 2023) and fine-tuning methodologies (e.g., LoRA (Hu et al., 2021), DPO (Rafailov et al., 2024)). This approach aimed to ensure that the chosen models collectively represent a range of architectures and fine-tuning strategies, thereby enabling a comprehensive evaluation of each model's potential for generating associative responses in Turkish.

The final models were selected based on their high performance across the leaderboard tests and their unique adaptations for Turkish language processing. Additionally, GPT-4o mini (OpenAI, 2024) was included based on its widespread popularity and proven success of GPT models in language processing (Brown et al., 2020), making it a relevant model for comparison despite not being a Turkish-specific LLM. The three models chosen for this experiment are as follows:

- GPT-4o mini (OpenAI, 2024) is one of the most widely known and used models for general-purpose language processing. It has been trained on a vast corpus of multilingual data, making it a suitable candidate for Turkish language tasks. GPT-4o mini is based on a Transformer architecture optimized for general-purpose language tasks. It utilizes a decoder-only variant of the Transformer and is trained on a diverse, multilingual corpus. Its architecture consists of numerous layers of attention mechanisms, typically with over 90 attention heads, which allows it to capture long-distance dependencies and produce coherent language outputs (Brown et al., 2020). This model's extensive general-purpose training can enable it to generate appropriate responses in associative tasks.
- Trendyol LLM v1.8 (Trendyol, 2024) is an auto-regressive large language chat model built on the Mistral 7B (Jiang et al., 2023) architecture, tailored for Turkish-language tasks. Mistral 7B, a 7-billion-parameter transformer model, is designed for efficient, high-performance language processing. It incorporates grouped-query attention (GQA) (Ainslie et al., 2023), which reduces the number of queries to enhance inference speed, and sliding window attention (SWA) (Child et al., 2019; Beltagy et al., 2020), allowing the model to process long sequences by attending to limited spans within a sliding window. These mechanisms reduce memory and computational costs, enabling Mistral 7B to handle long inputs more effectively (Jiang et al., 2023). Trendyol LLM v1.8 leverages this base model and is fine-tuned on Turkish instruction datasets using LoRA (Hu et al., 2021). LoRA facilitates task-specific tuning by adding trainable low-rank matrices, which adapt the

model effectively to Turkish tasks without adjusting the main model weights, thereby enhancing training efficiency (Wei et al., 2022).

- Cosmos LLaMA Instruct-DPO (COSMOS AI Research Group, 2024) is based on Meta's Large Language Model Meta AI (LLaMA) architecture (Touvron et al., 2023), designed to support text generation tasks. LLaMA itself is a foundational model that achieves competitive performance by training on a vast dataset of publicly available text sources, following a transformer architecture optimized for efficiency and scalability (Touvron et al., 2023). Cosmos LLaMA Instruct-DPO is fine-tuned on diverse instruction data using Direct Preference Optimization (Rafailov et al., 2024), which guides the model's responses to align with desired patterns and improve coherence. The model's broad training sources, including websites, books, and other text materials, contribute to its comprehensive language abilities.

3.2.1.4 Data preprocessing

The preprocessing of model outputs was designed to ensure consistency with the preprocessing steps applied to human data, thereby enabling a direct and meaningful comparison between the two datasets.

The first step in preprocessing involved cleaning the output of the language models. Models' outputs except GPT-4o mini (OpenAI, 2024) often included sentences structured as '*Aklıma gelen ilk üç kelime şunlardır: 1.. 2.. 3..'* (The first three words that come to mind are: 1.. 2.. 3..). To facilitate the analysis, such formatted outputs were cleaned while retaining other forms of sentence responses, which occurred less frequently. Furthermore, since the model pipeline configurations were not altered, some outputs contained repeated phrases or responses within a

single output. These duplicates were also removed to avoid redundancy and ensure the integrity of the analysis.

Rest of the preprocessing steps followed the same preprocessing pipeline used for human data. This included normalization, morphological analysis to retain unstemmed versions for certain analyses, and stemming using the TrnlpWord library in Python (Brolin59, n.d.).

3.3 Comparison criteria for human and LLM data

The comparison between human-generated word associations and the outputs produced by LLMs was conducted using several quantitative and qualitative metrics to assess the similarities and differences in their associative behavior. These metrics aimed to evaluate both the individual response tendencies and the overall structure of word associations between humans and LLMs.

3.3.1 Common responses and distinct response counts

The first stage of analysis involved identifying the most common responses in both the human and LLM datasets. For each cue word, two distinct analyses were performed:

- Analysis of First Responses (R1): The dataset was filtered to include only the first response provided by participants or generated by the LLMs for each cue word. These responses were then counted to determine the most frequently occurring first response for each cue word. This step focused on immediate associations, capturing the primary connections elicited by the cue word.
- Analysis of Combined Responses (R1, R2, R3): In this analysis, all responses from the three response positions—R1, R2, and R3—were considered

together. Each response, regardless of its position, was included in the analysis, and the frequency of each response was calculated. This allowed for a broader understanding of associative patterns by capturing both primary and secondary associations.

By comparing the results of these two analyses, it was possible to investigate how humans and LLMs converge or diverge in their associative responses to specific cue words. These steps mirror the methodologies of De Deyne et al. (2013) for Dutch, De Deyne et al. (2019) for English and Cabana et al. (2024) for Rioplatense Spanish associations, who similarly identified the most common responses in word association datasets. Their findings showed that the most frequent responses often reflect basic human needs such as money, food, or water, reflecting universal themes in associative behavior. Incorporating this perspective into the present analysis allowed for a deeper understanding of whether LLMs replicate such tendencies.

Set sizes (distinct response counts) were calculated for each cue across both datasets to understand the diversity of associations generated by humans and LLMs. This included calculating the average distinct response count across all participants and models, as well as within each response order (R1, R2, and R3). This distinction between response orders provided insight into whether the diversity diminished as participants and models moved from the first response to the second and third responses.

3.3.2 Morphological analysis

The morphological analysis of word associations was conducted to examine the relationship between the morphological properties of cues and responses. To analyze morphological complexity and suffix overlap, the Starlang Turkish Morphological

Analyzer (Yıldız et al., 2019) was used. This tool is based on finite-state transducers and performs morphological parsing by analyzing words into their constituent morphemes. It provides detailed information about each word's structure, allowing the extraction of morphological features.

The first aspect of the analysis focused on morphological complexity, measured as the number of morphemes attached to the stem of a word. The morphological complexity of cues and responses was computed, and the distribution of complexities across response orders was examined. To assess the relationship between cue and response complexities, correlations were calculated for each response order. A Poisson regression model was trained. In this model, response order (R1, R2, or R3) was treated as the independent variable, while morphological complexity was included as the dependent variable. This analysis tested whether the morphological complexity of responses varied systematically with their position in the response order and whether later responses (R2, R3) became more diverged from the complexity of the cues.

The second metric used in the analysis was suffix overlap, which measures the extent to which the morphemes of responses matched those of the cues. This was quantified using Jaccard distance, a set-based metric that computes the proportion of shared elements (in this case, morphemes) relative to the total unique elements in the cue and response.

(3) $$\text{Jaccard Distance} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Jaccard distance was particularly useful in this context, as it provides a clear and interpretable measure of morphological similarity, ranging from complete overlap (distance = 0) to no overlap (distance = 1). The distribution of Jaccard distances was analyzed across response orders to test whether suffix overlap systematically

decreased from R1 to R3. To validate these findings, statistical analysis was applied. A Friedman chi-square test was used to evaluate whether suffix overlap varied significantly across response orders. This non-parametric test was chosen because it is well-suited for repeated measures data, such as the Jaccard distances calculated for R1, R2, and R3.

3.3.3 Concreteness analysis

The concreteness analysis aimed to examine the relationship between the concreteness of cues and responses in the word association data. Concreteness scores, ranging from 1 (abstract) to 7 (concrete), were derived from the AnlamVer dataset, which itself is based on the Türkçe Kelime Normları (TKN) dataset (Tekcan & Göz, 2005). However, many words in the association dataset lacked direct concreteness ratings, necessitating an alternative approach to estimate these missing values. To account for words without direct concreteness scores in the dataset, word embeddings were utilized to estimate their concreteness values. This approach leverages findings from recent studies, including Flor (2024) and Ljubešić et al. (2018), which demonstrate that word embeddings are effective in approximating human-rated concreteness scores across languages. Based on these findings, FastText (Bojanowski et al., 2017) embeddings were employed to train a regression model that predicts concreteness scores for words missing direct ratings, enabling coverage of a larger portion of the dataset.

To predict concreteness for words without direct ratings, FastText word embeddings were extracted for all words in the dataset. These embeddings represent words as high-dimensional vectors, encoding both semantic and morphological information based on the contexts in which the words occur. For words with known

concreteness scores, their embeddings were used as input features, while their concreteness values served as the target variable. A Ridge regression model was then trained to predict concreteness, leveraging the strong association between embedding spaces and semantic properties. The workflow began with preprocessing the data by pairing known concreteness scores for 150 cues with their corresponding FastText embeddings. The dataset was split into training and testing subsets using an 80/20 split. Once the model was trained, it was applied to the embeddings of words without direct ratings to estimate their concreteness values. Ridge regression, a regularized linear regression model, was selected to handle potential multicollinearity and ensure stable predictions in high-dimensional embedding spaces.

The use of embeddings as predictors relies on the assumption that the semantic properties captured by embeddings correlate strongly with human-assigned concreteness ratings. For example, embeddings for highly concrete words like "tree" or "apple" are expected to cluster distinctly from embeddings for abstract words like "freedom" or "idea". The regression model leverages this underlying structure to infer concreteness for words without direct human ratings.

With the concreteness values assigned to all words, the relationship between response concreteness and response order was analyzed. First, the correlations between the concreteness of cues and responses were calculated to investigate whether more concrete cues elicit similarly concrete responses. In addition to correlation analysis, the distribution of concreteness scores was examined across response orders to identify potential systematic differences using Ordinary Least Squares (OLS) regression. This analysis treated response concreteness as the dependent variable and response order as the independent variable. The aim was to investigate whether the concreteness of responses systematically varied across

response orders, testing for trends that could reveal cognitive patterns in word association.

3.3.4 Association strengths and cosine similarity

To quantify the strength of associations between each cue and its responses, association strengths were calculated as conditional probabilities (Nelson et al., 2000; Nelson et al., 2004). Specifically, for each cue word (c), the association strength of a response (r) was defined as the probability that a participant or model generated that response, given the cue:

(4) Association Strength $= P(r \mid c)$

This probability reflects how strongly a specific response is associated with a given cue. For example, if the cue word is *adalet* (justice), and the response *hukuk* (law) is provided 12 times out of 108 total responses to *adalet*, the association strength is:

(5) P(hukuk | adalet) = 12 / 108 = 0.111.

These calculations were performed for both the human and LLM datasets.

The resulting data were organized into an association strength table, where each row corresponds to a cue-response pair. Table 3 below illustrates an example of this.

Table 3. Association Strength Table Example for *adalet* From Human Dataset

Cue	Response	Count	Total Responses	Association Strength
adalet	hukuk	12	108	0.111
adalet	hak	10	108	0.093
adalet	eşitlik	7	108	0.065
adalet	mahkeme	6	108	0.056
adalet	terazi	5	108	0.046
adalet	hakim	4	108	0.037
adalet	adil	3	108	0.028

Using this information, association strength vectors were created for each cue word. An association strength vector represents the strength of association between a cue word and all its possible responses. For instance, the vector for the cue word *adalet* would include the association strengths of *hukuk*, *hak*, *eşitlik*, and all other responses to *adalet*.

The association strength matrix is a structured representation of these vectors, where each row corresponds to a cue word, and each column represents a response across the dataset. Each cell in the matrix contains the association strength of a specific response for a specific cue. For example:

Table 4. Example Association Strength Matrix for Cue Words *adalet* and *terazi* From Human Dataset

Cue	hukuk	hak	eşitlik	terazi	altın	ağırlık	bakkal
adalet	0.111	0.093	0.065	0.046	0.000	0.000	0.000
terazi	0.000	0.000	0.028	0.167	0.009	0.083	0.009

Using these vectors, the cosine similarity between pairs of cue words was calculated to determine how similar their associative patterns were. The cosine similarity between pairs of cue words was then calculated using these vectors (DeDeyne et al., 2019). Cosine similarity is a metric used to measure the similarity between two vectors by calculating the cosine of the angle between them. It ranges from -1 to 1, where a value of 1 indicates identical vectors (or identical associative patterns in this context), 0 indicates orthogonal vectors (no similarity), and -1 indicates opposite vectors. This metric provided a measure of how similar the associative patterns of different cue words were, based on the strength of their associations with shared responses. Cosine Similarity between two vectors A and B is calculated as (6) where $A \cdot B$ is the dot product of the vectors, and $|A|$ and $|B|$ are the magnitudes (or Euclidean norms) of the vectors.

(6) $$\text{Cosine Similarity} = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3.3.5 Correlations with *AnlamVer* dataset

The cosine similarity values calculated from both the human and LLM data were compared against the similarity and relatedness scores provided in the AnlamVer dataset. This dataset assigns two distinct scores to each pair of words: one for similarity (how alike the meanings of the words are) and one for relatedness (how often or contextually the words are associated). By calculating the correlation between the cosine similarity values and the AnlamVer similarity/relatedness scores, it was possible to assess how well the associations generated by humans and LLMs reflected established semantic relationships in Turkish. This provided a key metric for evaluating the cognitive plausibility of the LLMs' associative behavior.

3.3.6 Human-LLM cosine similarity and edit distance

To further evaluate the alignment between human and LLM associations, the cosine similarities between cue pairs from the human data were compared with those from the LLM data. This comparison involved calculating the correlation between the cosine values assigned by humans and LLMs to the same cue pairs. High correlations would suggest that humans and LLMs are generating similar associative structures, while lower correlations would indicate divergent patterns.

Table 5. Example Cue-Pair Cosine Similarity (CS) Values Across Human, GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA Data

Cue 1	Cue 2	Human CS	GPT-4o mini CS	Trendyol LLM v1.8 CS	Cosmos LLaMA CS
yağmur	çiçek	0.038	0.284	0.034	0.0
hastane	kemik	0.075	0.013	0.0015	0.023
laikçiler	sekülerizmciler	0.490	0.723	0.097	0.353
bilimci	mühendis	0.010	0.183	0.018	0.023
hayat	nefes	0.253	0.241	0.002	0.027

Additionally, the edit distance between the human and LLM cosine dictionaries was calculated (see Table 6 for the top 5 elements of the human cosine dictionary). Edit distance, also known as Levenshtein distance, is a metric that quantifies the minimum number of changes (insertions, deletions, or substitutions) needed to transform one sequence into another:

(7) $Levenshtein\ Distance\ =$

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

For this analysis, the edit distance was normalized by dividing the raw edit distance by the length of the longest list. Applied here, the normalized edit distance measures how different the ordering or structure of cosine similarity lists is between humans and LLMs. By assessing the number of transformations needed to align the LLMs' cosine similarity list with the human list, a finer-grained measure of alignment is obtained, providing insight into whether the associative relationships follow similar patterns at the level of specific cue pairs.

Table 6. Top 5 Cosine Similarity Scores Between Cue Word Pairs From Human Data Jaccard Similarity of Non-Zero Cosines

Cue 1	Cue 2	Cosine Similarity
orman	ormanlık	0.864
damar	pıhtı	0.862
yorgun	yorgunluk	0.832
barınak	hayvan	0.793
ateistlik	teizm	0.782

To complement the edit distance analysis, Jaccard similarity was also calculated, focusing specifically on the non-zero cosine values in the cue-pair similarity lists for both humans and LLMs. Jaccard similarity is a metric that

quantifies the similarity between two sets by measuring the proportion of shared elements relative to the total elements in both sets. It is defined as:

(8) $$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

where A and B represent two sets—in this case, the sets of cue pairs with non-zero cosine similarities for humans and LLMs. The numerator, $|A \cap B|$, represents the number of shared cue pairs that both humans and LLMs assigned a non-zero cosine similarity, indicating that they consider these cue pairs to have some level of associative strength. The denominator, $|A \cup B|$, represents the total unique cue pairs with non-zero cosine values in either the human or LLM lists.

By focusing on non-zero cosine values, this Jaccard similarity calculation specifically highlights the degree of overlap in associations deemed meaningful or relevant by both humans and LLMs. In other words, it captures the extent to which both datasets identify the same cue pairs as related. High Jaccard similarity indicates a strong alignment in identifying which cue pairs are considered related, even if the strength of the association differs. Low Jaccard similarity, on the other hand, would suggest that humans and LLMs disagree on which cue pairs hold meaningful associations.

This measure provides a nuanced view of agreement between human and LLM associations, independent of the exact cosine values. It allows for an assessment of similarity that focuses on the presence or absence of associative relationships, rather than the precise strength. Therefore, Jaccard similarity offers insight into the shared structure of associations, complementing other metrics like cosine correlation, which measure similarity in strength.

3.3.7 Response order and cosine similarity

In addition to analyzing overall similarities between humans and LLMs, an analysis was conducted to investigate whether the order of responses (R1, R2, R3) affected the cosine similarity between cues and responses. Using the pretrained Turkish FastText model (Grave et al., 2018), word vectors were obtained for each cue and response in the dataset. The mean cosine similarity between the cue and its responses was calculated for each response order (R1, R2, R3) in both the human and LLM datasets.

This analysis aimed to explore whether response order influenced the semantic similarity between cues and responses. For example, stronger semantic associations were expected in R1 compared to R3, as the first responses typically reflect stronger, more immediate connections, while later responses may capture weaker or more context-dependent associations.

3.3.8 Graph construction and similarity metrics

Finally, the data from both humans and LLMs were represented as graphs, where nodes represented cues and responses, and the edges between them are weighted based on the association strengths. These graphs allow for a visual and structural comparison between human and LLM associations.

The similarity between these graphs was quantified using several metrics, including:

- Jaccard Similarity: This was used to measure the overlap between the edges in the human and LLM graphs. A higher Jaccard similarity indicated that both humans and LLMs tended to form similar connections between words.

- Sørensen–Dice Similarity: This metric was employed to compare the overall structure of the graphs by measuring the extent to which the edges between nodes (cue-response pairs) were shared between the human and LLM graphs.

  (8) $$\text{Sørensen–Dice Similarity (SDS)} = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

- Overlap Coefficient: This coefficient was used to measure the degree of overlap in associations between the human and LLM graphs. This metric provided insight into how closely aligned the associations were between the two datasets.

  (9) $$\text{Overlap Coefficient (OC)} = \frac{|A \cap B|}{\min(|A|, |B|)}$$

By constructing and comparing these graphs, the study was able to evaluate the extent to which the associative structures generated by LLMs mirrored the patterns observed in human responses. This analysis provided a holistic view of the structural similarities and differences between human and LLM word associations, contributing to the overall understanding of how well LLMs simulate human-like associative behavior.

CHAPTER 4

RESULTS

This chapter presents the findings of the study by comparing the word association data collected from human participants with those generated by large language models. The analysis aims to uncover similarities and differences in associative patterns, focusing on both descriptive and inferential statistics. By evaluating human and machine-generated word associations, the chapter provides insights into the strengths and limitations of current language models in replicating human-like semantic processes.

The chapter is organized into multiple sections. The first section details the descriptive results, providing an overview of the most frequent responses and set sizes across different response orders (R1, R2, R3). This is followed by a report on the relationship between response order and set sizes, modeled using Poisson regression. The descriptive section is crucial in establishing a baseline for understanding the associative trends in both datasets (human and LLM). Subsequent sections delve into more complex analyses, including comparative metrics such as cosine similarity, Jaccard similarity, and graph-based measures of association. These sections explore structural and functional differences between human and LLM-generated associations, revealing the extent to which LLMs align with or diverge from human response patterns. Interpretations and broader implications of these results will be discussed in detail in Chapter 5.

4.1 Descriptive results

The analysis of the first response (R1) reveals distinct patterns in how humans, GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA associate specific words with given cues. Among human participants, the most frequent first-order response was *para* (money), mentioned 70 times, followed by *ağaç* (tree) with 58 occurrences and *yemek* (food) with 57 occurrences. The fourth and fifth most frequent responses were *beyaz* (white) and *cam* (glass), occurring 46 and 44 times, respectively.

Table 7. Top 5 Most Common Responses (R1)

	Human		GPT-4o mini		Trendyol LLM		Cosmos LLaMA	
Rank	R	Freq.	R	Freq.	R	Freq.	R	Freq.
1	para	70	din	75	şeffaf	84	yemek	127
2	ağaç	58	şeffaf	71	bira	83	kitap	80
3	yemek	57	meyve	68	seyahat	56	giysi	63
4	beyaz	46	içki	64	orman	53	ağaç	62
5	cam	44	ağaç	61	araba	53	insan	62

For GPT-4o mini, the most frequent first-order responses included *din* (religion), *şeffaf* (transparent), and *meyve* (fruit), with 75, 71, and 68 occurrences, respectively. Similarly, the Trendyol LLM v1.8 frequently generated *şeffaf* (transparent) and *bira* (beer) as its top responses, with 84 and 83 occurrences, followed by *seyahat* (travel), *orman* (forest), and *araba* (car), each mentioned 56, 53, and 53 times, respectively. Cosmos LLaMA generated *yemek* with a much higher

frequency than humans, followed by *kitap* (book), *giysi* (clothing), *ağaç* and *insan* (person).

Table 8. Top 5 Most Common Responses Across All Cues (R123

	Human		GPT-4o mini		Trednyol LLM		Cosmos LLaMA	
Rank	R	Freq.	R	Freq.	R	Freq.	R	Freq.
1	yemek	159	doğa	203	bira	224	yemek	225
2	para	154	din	159	kitap	119	tatlı	171
3	beyaz	153	yemek	127	mezarlık	118	ölüm	133
4	ağaç	137	meyve	104	şüphe	112	hava	128
5	sarı	119	yaşam	101	araba	105	güçlü	120

When considering all three response orders combined (R123), the frequency patterns shift slightly. For humans, *yemek* becomes the most common response with 159 mentions, closely followed by *para* (154 occurrences) and *beyaz* (153 occurrences). *ağaç* remains a frequent association with 137 mentions, while *sarı* (yellow) enters the top five with 119 occurrences. For GPT-4o mini, *doğa* (nature) was the most common association, with 203 mentions, followed by *din* (159) and *yemek* (127). Meanwhile, the Trendyol LLM v1.8 generated *bira* (beer) as the most frequent response with 224 mentions, followed by *kitap* (book, 119) and *mezarlık* (cemetery, 118). Cosmos LLaMA's most frequent responses were *yemek* (225), *tatlı* (dessert, 171), and *ölüm* (death, 133).

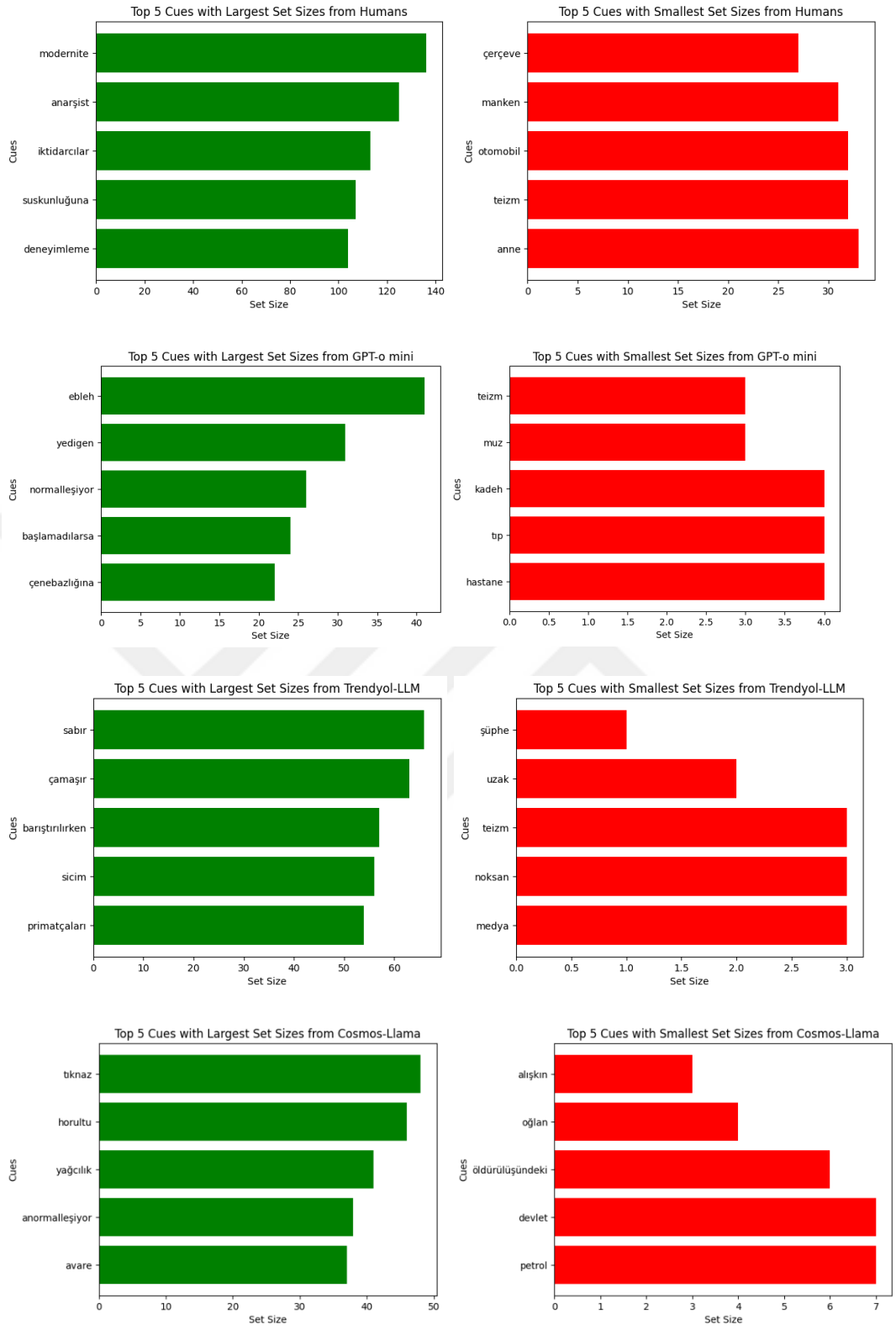


Figure 8. Horizontal bar charts for cues with the largest and smallest set sizes, with set sizes plotted on the x-axis for each dataset.

The diversity of associations generated by participants is reflected in the set sizes for each cue word. The cues eliciting the largest set sizes from humans were *modernite* (modernity), *anarşist* (anarchist), and *iktidarcılar* (rulers), with set sizes of 136, 125, and 113, respectively. These cues, often abstract, appear to evoke a wide range of responses. Conversely, the smallest set sizes were observed for cues such as *çerçeve* (frame), *manken* (model), and *otomobil* (car), which elicited 27, 31, and 32 unique responses, respectively. These cues, being more concrete, may limit the variability in participants' associations. GPT-4o mini's set sizes show variability, with the largest set size observed for *ebleh* (stupid) (41) and smaller set sizes for cues like *teizm* (theism) (3) and *muz* (banana) (3). This suggests that GPT-4o mini's associative range can vary significantly depending on the input cue. Trendyol LLM v1.8 results show smaller set sizes overall, with highly constrained responses for certain cues like *şüphe* (doubt) and *uzak* (far), contrasting with human data where even the smallest sets contained greater diversity. The cues eliciting the largest set sizes from Cosmos LLaMA were *tıknaz* (stout), *horultu* (snore), and *yağcılık* (flattery), with set sizes of 48, 46, and 41, respectively. Conversely, the smallest set sizes were observed for cues such as *alışkın* (accustomed), *oğlan* (boy), and *öldürülüşündeki* (in his/her murder), which elicited 3, 4, and 6 unique responses, respectively.

The average number of set sizes varied across response orders, increasing progressively from R1 to R3 for all datasets. For human participants, the averages were 21.24 for R1, 24.38 for R2, and 25.30 for R3, with an overall set size of 54.24 across all three response orders. To examine whether the difference in average set sizes across response orders (R1, R2, R3) was statistically significant, a Poisson regression analysis was conducted. Poisson regression analysis for human data revealed a small but significant positive relationship between response order and

count of distinct responses, with a coefficient of 0.114 ($p < 0.001$), indicating that later response orders were associated with an increased number of distinct responses. For GPT-4o mini, the averages were significantly lower than those of humans, with 3.77 for R1, 6.90 for R2, and 9.34 for R3, and an overall average of 11.96 distinct responses. The regression results for GPT-4o mini showed a stronger relation compared to human data, with a coefficient of 0.43 ($p < 0.001$), showing a moderate positive relationship between response order and distinct responses. However, the intercept is lower (0.976, $p < 0.001$) for GPT-4o mini.

The Trendyol LLM v1.8 exhibited a less limited pattern than GPT-4o mini, with averages of 8.04 for R1, 11.86 for R2, and 8.91 for R3, resulting in an overall average of 24.25 distinct responses. Poisson regression for Trendyol LLM v1.8 showed a weaker effect of response order on distinct responses, with a coefficient of 0.045 ($p = 0.005$). While all datasets exhibit a positive relationship between response order and distinct responses, the Trendyol LLM v1.8's smaller coefficient indicates that its associations are less influenced by response order.

For Cosmos LLaMA, the averages for distinct responses were 5.81 for R1, 8.48 for R2, and 10.71 for R3, resulting in an overall average of 17.68 distinct responses. Poisson regression analysis revealed a coefficient of 0.298 ($p < 0.001$), indicating a positive relationship between response order and distinct responses, though this effect was smaller compared to GPT-4o mini, it is larger than humans. The intercept (1.494, $p < 0.001$) reflects Cosmos LLaMA's intermediate position between GPT-4o mini and Trendyol LLM v1.8 in terms of generative scope.

The results are summarized in Table 9.

Table 9. Poisson Regression Results for the Relationship Between Response Order and Set Sizes

Dataset	Int.	SE	z	p	Coef	SE	z	p
Humans	2.910	0.023	125.835	<0.001	0.114	0.010	11.037	<0.001
GPT-4o mini	0.976	0.048	20.165	<0.001	0.430	0.020	21.403	<0.001
Trendyol LLM v1.8	2.171	0.035	61.92	<0.001	0.045	0.016	2.82	0.005
Cosmos LLaMA	1.494	0.041	36.388	<0.001	0.298	0.018	16.979	<0.001

The Cohen's *d* values calculated for the differences in average set sizes between human responses and those of the LLMs reveal consistently large effect sizes across all response orders (R1, R2, R3) and the combined responses. For human-GPT-4o mini comparisons, Cohen's d values are highest overall, reaching 2.85 for R1, 2.74 for R2, 2.45 for R3, and 3.08 for the combined responses. The human-Trendyol LLM v1.8 comparisons also show very large effect sizes, with Cohen's d values ranging from 2.11 for R1 to 2.53 for R3, and 2.33 for the combined responses. Trendyol LLM v1.8 demonstrates a more pronounced divergence in tertiary responses (R3) compared to primary responses (R1). In the case of human-Cosmos LlaMa comparisons, Cohen's d values are similarly large, with 2.46 for R1, 2.42 for R2, 1.77 for R3, and 2.54 for combined responses. Overall, the results highlight significant and large differences between human and LLM responses, with

GPT-4o mini displaying the largest overall divergence, followed by Cosmos LlaMa and Trendyol LLM v1.8.

A logistic regression was conducted to examine the relationship between set sizes, response order, and the likelihood of a source being human or Trendyol LLM v1.8 or GPT-4o mini or Cosmos LLaMA. The dependent variable was the source (1 for human, 0 for the LLMs), while the predictors included the set size and the response order (R1, R2, R3).

For GPT-4o mini, the model showed a strong fit with a pseudo R-squared value of 0.89, indicating that 89% of the variance in source identification was explained by the predictors. Each additional distinct response increased the odds of the source being human, with the log-odds increasing by 1.07 per additional response ($p < 0.001$). This reflects humans' tendency to produce a greater variety of responses compared to GPT-4o mini.

For Trendyol LLM v1.8, the model achieved a pseudo R-squared value of 0.51, with a one-unit increase in distinct responses increasing the odds of the source being human by 32% ($p < 0.001$), and later response orders decreasing the odds by 29% ($p = 0.002$). These results highlight significant differences in generative patterns, with humans producing more diverse and varied responses compared to both GPT-4o mini and Trendyol LLM v1.8.

For Cosmos LLaMA, the model achieved a pseudo R-squared value of 0.7475, with each additional distinct response increasing the odds of the source being human by 68.4% ($p < 0.001$), while later response orders reduced the odds with a log-odds decrease of 1.921 per unit increase in response order ($p < 0.001$). This places Cosmos LLaMA’s generative diversity closer to human data than Trendyol LLM v1.8 but still behind GPT-4o mini.

4.2 Morphological analysis results

This analysis examines how morphological complexity and suffix patterns differ in associations generated by humans and models to explore linguistic similarities and divergences.

The analysis of morphological complexities and suffix overlaps between cues and responses revealed distinct patterns across human data, GPT-4o mini, Cosmos LlaMa and Trendyol LLM v1.8. For humans, morphological complexities showed weak but statistically significant correlations between the cue and responses across all response orders. The correlations for cue and R1, R2, and R3 were $r = 0.073$, $r = 0.095$, and $r = 0.075$, respectively, with $p < 0.001$ for each. Poisson regression did not identify significant relationships between response order and response complexity ($\beta = 0.0012$, $p = 0.702$). Regarding suffix overlaps, the average Jaccard distances were 0.47 for R1, 0.49 for R2, and 0.49 for R3. The Friedman test did not find significant differences in suffix overlap across response orders ($\chi^2 = 3.119$, $p = 0.210$).

In the GPT-4o mini data, correlations between cue and response complexities were higher than those observed in human data and were all statistically significant. The correlations for cue and R1, R2, and R3 were $r = 0.12$, $r = 0.14$, and $r = 0.15$, respectively, with $p < 0.001$. The Poisson regression did not show significant relationships between response order and response complexity ($\beta = -0.0042$, $p = 0.143$). Suffix overlap analysis showed mean Jaccard distances of 0.46 for R1, 0.49 for R2, and 0.50 for R3. These differences were statistically significant according to the Friedman test ($\chi^2 = 79.741$, $p < 0.001$).

In the Trendyol LLM v1.8 data, correlations between cue and response complexities were also significant, with the highest correlation for R1 ($r = 0.19$, $p <$

0.001). For R2 and R3, the correlations were $r = 0.08$ and $r = 0.10$, respectively, both $p < 0.001$. Poisson regression identified no significant relationship between response complexity and response order ($\beta = 0.0026$, p = 0.188). The suffix overlap analysis revealed mean Jaccard distances of 0.50 for R1, 0.65 for R2, and 0.64 for R3. The Friedman test showed significant differences in suffix overlap across response orders ($\chi^2 = 951.980$, $p < 0.001$).

For the Cosmos LLaMA dataset, correlations between cue and response complexities were $r = 0.158$ (R1, $p < 0.001$), $r = 0.132$ (R2, $p < 0.001$), and $r = 0.109$ (R3, $p < 0.001$). Poisson regression indicated a weak but significant relationship ($\beta = 0.0059$, $p = 0.014$). The suffix overlaps showed mean Jaccard distances of 0.55 for R1, 0.58 for R2, and 0.58 for R3. The Friedman test confirmed significant differences ($\chi^2 = 36.513$, $p < 0.001$).

Cohen's d was calculated to further assess the magnitude of differences in morphological complexity between human and model responses across response orders and overall. For GPT-4o mini, the effect sizes were minimal, with Cohen's d values of -0.10 (R1), −0.07 (R2), −0.03 (R3), and −0.07 (overall), suggesting negligible differences in complexity between GPT-4o mini and human responses. For Trendyol LLM v1.8, moderate to large differences were observed, with Cohen's d values of 0.63 (R1), 0.58 (R2), 0.52 (R3), and 0.57 (overall), reflecting its greater morphological complexity compared to human responses. The differences for Cosmos LLaMA were also negligible, with Cohen's d values of 0.07 (R1), 0.07 (R2), −0.00 (R3), and 0.05 (overall). These results emphasize the substantial variability in alignment with human-like complexity among models, with GPT-4o mini and Cosmos LLaMA resembling human patterns, while Trendyol LLM v1.8 diverges significantly based on complexity distributions.

Statistical tests further highlight the differences in morphological complexities between humans and large language models. A Mann-Whitney U test comparing humans and GPT-4o mini responses revealed a statistically significant difference ($U = 215{,}996{,}021.5$, $p < 0.001$), although the observed difference appears small in magnitude based on the distributions. Comparing humans and Trendyol LLM v1.8, a highly significant difference was observed ($U = 314{,}199{,}041.0$, $p < 0.001$), consistent with Trendyol LLM v1.8's broader range and higher median complexities. Similarly, the comparison of humans and Cosmos LLaMA complexities showed a significant difference ($U = 248{,}742{,}921.5$, $p < 0.001$).

Pairwise comparisons of Jaccard distances between humans and LLMs also reveal varied patterns. A Mann-Whitney U test comparing humans and GPT-4o mini showed no statistically significant difference ($U = 125{,}136{,}440.5$, $p = 0.861$), suggesting that GPT-4o mini aligns with human suffix overlap patterns. However, the comparison between humans and Trendyol LLM v1.8 revealed a significant difference ($U = 116{,}959{,}389.5$, $p < 0.001$), indicating that its responses exhibit distinct suffix overlap patterns compared to humans. Similarly, the comparison of humans and Cosmos LLaMA suffix overlaps showed a significant difference ($U = 124{,}430{,}877.0$, $p < 0.001$).

These results underscore notable variations in morphological complexity and suffix overlap patterns between humans and large language models. While the correlations reported are statistically significant, the observed effect sizes are small, particularly for human data. This suggests that while morphological complexity and suffix patterns influence responses, their practical impact on the association process is likely limited.

4.3 Concreteness analysis results

This analysis investigates the relationship between the concreteness of cues and responses across human and LLM data, focusing on the correlations and systematic variations in concreteness across response orders.

For human data, Pearson correlation coefficients revealed significant positive correlations between cue concreteness and response concreteness for all response orders. The correlations were $r = 0.413$ for R1, $r = 0.385$ for R2, and $r = 0.394$ for R3, with $p < 0.001$ in all cases. Ordinary Least Squares regression further examined the systematic relationship between cue concreteness, response order, and response concreteness. The regression model indicated that both cue concreteness ($\beta = 0.347$, $p < 0.001$) and response order ($\beta = -0.044$, $p = 0.001$) were significant predictors of response concreteness. The negative coefficient for response order suggests that response concreteness slightly decreases in later response orders. The model explained 16% of the variance in response concreteness ($R^2 = 0.158$).

For GPT-4o mini data, correlations between cue and response concreteness were stronger than those observed in human data. The correlations were $r = 0.617$ for R1, $r = 0.522$ for R2, and $r = 0.510$ for R3, all with $p < 0.001$. The OLS regression model similarly revealed that cue concreteness ($\beta = 0.465$, $p < 0.001$) and response order ($\beta = -0.128$, $p < 0.001$) were significant predictors of response concreteness. The larger negative coefficient for response order in the GPT-4o mini data indicates a more pronounced decrease in response concreteness across orders compared to human data. The model explained 31% of the variance in response concreteness ($R^2 = 0.308$).

For Trendyol LLM v1.8 data, the correlations between cue and response concreteness were the highest among the three datasets. The correlations were $r =$

0.831 for R1, $r = 0.565$ for R2, and $r = 0.612$ for R3, all with $p < 0.001$. The OLS regression model revealed that both cue concreteness ($\beta = 0.626$, $p < 0.001$) and response order ($\beta = -0.177$, $p < 0.001$) significantly predicted response concreteness. The negative coefficient for response order indicates a more substantial decline in response concreteness in later response orders compared to both human and GPT-4o mini data. The model explained 45% of the variance in response concreteness ($R^2 = 0.454$).

For Cosmos LLaMA data, correlations between cue and response concreteness were slightly stronger than those observed in human data. The correlations were $r = 0.434$ for R1 ($p < 0.001$), $r = 0.409$ for R2 ($p < 0.001$), and $r = 0.416$ for R3 ($p < 0.001$). The regression model indicated that both cue concreteness ($\beta = 0.347$, $p < 0.001$) and response order ($\beta = -0.092$, $p < 0.001$) were significant predictors of response concreteness. The larger negative coefficient for response order suggests a more substantial decrease in response concreteness across orders compared to human data. The model explained 18% of the variance in response concreteness ($R^2 = 0.179$).

Pairwise regressions comparing Human-GPT-4o mini, Human-Trendyol LLM v1.8 and Human-Cosmos LLaMa data provided further insights into dataset-specific differences:

- Human-GPT Regression Results: The regression revealed significant effects of cue concreteness ($\beta = 0.403$, $p < 0.001$) and response order ($\beta = -0.083$, $p < 0.001$) on response concreteness, with GPT responses being less concrete than human responses ($\beta = -0.229$, $p < 0.001$). The model explained 23% of the variance in response concreteness ($R^2 = 0.228$).

- Human-Trendyol Regression Results: The regression showed significant effects of cue concreteness ($\beta = 0.453$, $p < 0.001$) and response order ($\beta = -0.095$, $p < 0.001$) on response concreteness, with Trendyol LLM v1.8 responses being slightly less concrete than human responses ($\beta = -0.165$, $p < 0.001$). The model explained 26% of the variance in response concreteness ($R^2 = 0.260$).
- Human-Cosmos Regression Results: The regression indicated significant effects of cue concreteness ($\beta = 0.347$, $p < 0.001$) and response order ($\beta = -0.066$, $p < 0.001$) on response concreteness, with Cosmos LLaMA responses being less concrete than human responses ($\beta = -0.239$, $p < 0.001$). The model explained 17% of the variance in response concreteness ($R^2 = 0.174$).

These findings highlight significant relationships between cue and response concreteness across all datasets, with stronger correlations and systematic decreases in concreteness across response orders observed in the language models compared to human data. Among the LLMs, Trendyol LLM v1.8 exhibited the strongest cue-response concreteness relationships and the smallest deviations from human data. GPT-4o Mini responses were the least concrete compared to human responses, showing the largest deviation. Cosmos LLaMA responses, while less concrete than human responses, were closer to human patterns than GPT but more divergent than Trendyol LLM v1.8.

4.5 Association strengths and cosine similarity with correlations to *AnlamVer*

The results of the analyses revealed distinct patterns in the association strengths and cosine similarity correlations for humans, GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA data.

For the human data, the strongest associations were observed for cue-response pairs such as *Kuyumcu-Altın* (jeweler-gold), *Çerçeve-Resim* (frame-picture), and *Pıhtı-Kan* (clot-blood), with association strength scores of 0.33, 0.31, and 0.31, respectively. Table 10 presents the top 10 strongest associations for humans.

Table 10. Association Strengths Between Cue-Response Pairs From Humans

Cue	Response	Association Strength
Kuyumcu	Altın	0.33
Çerçeve	Resim	0.31
Pıhtı	Kan	0.31
Orman	Ağaç	0.28
Kadeh	Şarap	0.28
Damar	Kan	0.28
Taksi	Sarı	0.27
Sahil	Deniz	0.26
Kazanç	Para	0.26
Primatçaları	Maymun	0.26

The calculated cosine similarity values for cue pairs from the association strength vectors showed significant correlations with the *AnlamVer* dataset, with a similarity

correlation of 0.499 ($p < 0.001$) and a relatedness correlation of 0.654 ($p < 0.001$). Regression analyses confirmed these correlations' statistical significance, with cosine similarity as a significant predictor of both similarity ($R^2 = 0.249$, $p < 0.001$) and relatedness scores ($R^2 = 0.428$, $p < 0.001$).

For GPT-4o mini, the association strength results highlighted pairs such as *Petrol-Enerji* (petrol-energy) (0.36), *Kapüşon-Başlık* (hood-headpiece) (0.35), and *Lokanta-Restoran* (eatery-restaurant) (0.35). Cosine similarity correlations with the *AnlamVer* dataset were 0.644 for similarity and 0.585 for relatedness, both statistically significant ($p < 0.001$). Regression analyses further supported these findings, with cosine similarity accounting for 41.5% of the variance in similarity scores ($p < 0.001$) and 34.2% of the variance in relatedness scores ($p < 0.001$). These results indicate that GPT-4o mini's associative patterns closely mimic semantic and relatedness measures in the dataset, though with slight variations compared to human data.

Table 11. Association Strengths Between Cue-Response Pairs From GPT-4o mini

Cue	Response	Association Strength
Petrol	Enerji	0.361446
Kapüşon	Başlık	0.348837
Lokanta	Restoran	0.348315
Dükkan	Mağaza	0.344828
Hayat	Yaşam	0.344444
Yoğurt	Süt	0.340909
Lokanta	Yemek	0.337079
Uzak	Mesafe	0.337079
Kuşku	Şüphe	0.333333
Kitaplıklar	Kitap	0.333333

For the Trendyol LLM v1.8, the strongest associations included *Uzak-Uzak* (0.98), *Noksan-Eksik* (0.95), and *Adalet-Adalet* (0.88). Correlations with the *AnlamVer* dataset were slightly lower than those for human and GPT-4o mini data, with 0.462 for similarity and 0.350 for relatedness ($p < 0.001$). The regression analyses indicated that cosine similarity explained 21.4% of the variance in similarity scores ($p < 0.001$) and 12.2% of the variance in relatedness scores ($p < 0.001$). While the associations generated by Trendyol LLM v1.8 were less aligned with the *AnlamVer* dataset than those of humans and GPT-4o mini, they still demonstrated some degree of cognitive plausibility.

Table 12. Association Strengths Between Cue-Response Pairs From Trendyol LLM v1.8

Cue	Response	Association Strength
Uzak	Uzak	0.977778
Noksan	Eksik	0.954545
Adalet	Adalet	0.877778
Masum	Masum	0.877778
Otobüs	Otobüs	0.853933
Gözlük	Gözlük	0.816092
Titiz	Dikkatli	0.802326
Şüphe	Şüphe	0.800000
Savunmaktaydılar	Savunmak	0.792208
Kadeh	Bira	0.777778

For Cosmos LLaMA, the strongest associations included pairs such as *öldürülüşündeki-ölüm* (in-the-death), *alışkın-alışkanlık* (accustomed-habit), and *başlamadılarsa-başlangıç* (if-they-did-not-start-beginning), with association strength scores of 0.90, 0.71, and 0.61, respectively. Table 13 below presents the top 10 strongest associations for Cosmos LLaMA:

Table 13. Association Strengths Between Cue-Response Pairs From Cosmos LLaMA

Cue	Response	Association Strength
öldürülüşündeki	ölüm	0.900
alışkın	alışkanlık	0.711
başlamadılarsa	başlangıç	0.611
yorgun	yorgunluk	0.528
asosyalleştirdikleri	sosyal	0.474
yeni	yenilik	0.455
süt	beyaz	0.419
su	su	0.412
kıtlıklarda	yoksunluk	0.402
dükkan	mağaza	0.398

The cosine similarity values calculated from the association strength vectors for Cosmos LLaMA also showed significant correlations with the AnlamVer dataset. The correlation for similarity was 0.566 ($p < 0.001$), and for relatedness, it was 0.539 ($p < 0.001$). The regression analyses further examined the relationship between cosine similarity and the AnlamVer metrics. For similarity, cosine similarity was a significant predictor, explaining 32% of the variance ($p < 0.001$). For relatedness, cosine similarity accounted for 29% of the variance ($p < 0.001$).

In summary, all datasets demonstrated significant correlations with semantic and relatedness measures, though the degree of alignment varied across humans, GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA, reflecting differences in their associative mechanisms.

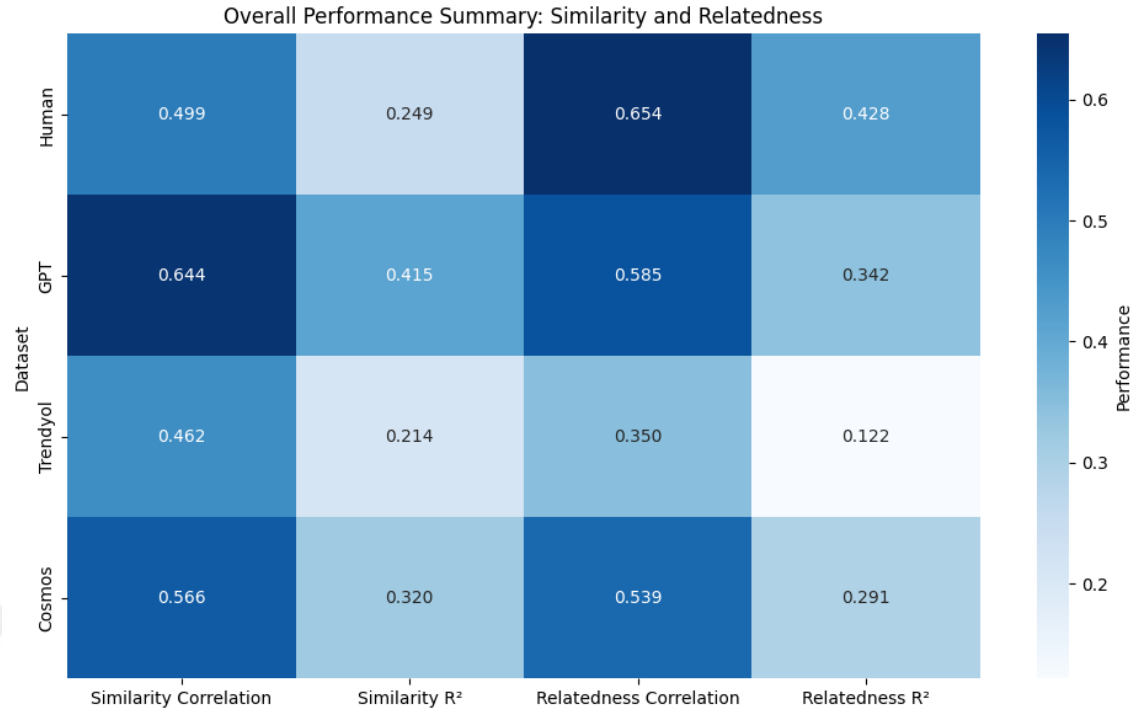


Figure 9. Performance Summary of Correlations and R^2 Values for Similarity and Relatedness Across Datasets

4.6 Human-LLM cosine similarity: edit distance and jaccard similarity of non-zeros

The cosine similarity values for cue word pairs were analyzed to evaluate the alignment between human associations and those generated by GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA. In the methodology chapter, the top five most similar cue pairs from the human data, based on association strength vectors, were listed. These pairs were *orman–ormanlık* (0.864), *damar–pıhtı* (0.8618), *yorgun–yorgunluk* (0.8315), *barınak–hayvan* (0.7931), and *ateistlik–teizm* (0.782).

The corresponding top five cue pairs from GPT-4o mini, ranked by cosine similarity values calculated from association strength vectors, are presented in the table below:

Table 14. Top 5 Cosine Similarity Scores Between Cue Word Pairs From GPT-4o mini

Cue 1	Cue 2	Cosine Similarity
orman	ormanlık	0.986340
saydam	transparan	0.900829
atatürkist	kemalci	0.898518
nefes	oksijen	0.853096
inançlılık	teizm	0.825222

Similarly, the top five cue pairs for Trendyol LLM v1.8, ranked by their cosine similarity values, are shown in Table 15.

Table 15. Top 5 Cosine Similarity Scores Between Cue Word Pairs From Trendyol LLM v1.8

Cue 1	Cue 2	Cosine Similarity
saydam	transparan	0.984068
boza	kadeh	0.940724
orman	ormanlık	0.916009
kadeh	yiyecek	0.913909
araba	otomobil	0.908988

For Cosmos LLaMA, the top five cue pairs based on cosine similarity values are presented below:

Table 16. Top 5 Cosine Similarity Scores Between Cue Word Pairs From Cosmos LLaMA

Cue 1	Cue 2	Cosine Similarity
kitaphane	kitaplıklar	0.755288
mezar	öldürülüşündeki	0.701662
ikindi	öğle	0.681777
orman	zeytin	0.666687
kuşku	şüphe	0.665076

To assess the alignment between human and LLM cosine similarity scores, the correlations between the cosine similarity values of the same cue pairs were calculated. For GPT-4o mini, the Spearman correlation coefficient was 0.3341 and Kendall's tau was 0.2271, both with a *p*-value of 0.0, indicating statistical significance. For Trendyol LLM v1.8, the Spearman correlation coefficient was 0.3286 and Kendall's tau was 0.2212, also with a *p*-value of 0.0, indicating a significant but moderate positive correlation. For Cosmos LLaMA, the correlations were lower, with a Spearman correlation coefficient of 0.113 ($p < 0.001$) and Kendall's tau of 0.081 ($p < 0.001$). These results suggest a weaker alignment of Cosmos LLaMA's associative structures with human data compared to GPT-4o mini and Trendyol LLM v1.8.

Edit distance was calculated to further examine structural differences between the ranked lists of cosine similarities. These lists were ranked in descending order of cosine similarity values within their respective datasets. The normalized Levenshtein distance between the sorted human cue pair cosine similarity list and all of the LLMs were 0.99. These values provide a measure of the number of transformations required

to align the LLM lists with the human list, indicating a comparable degree of structural deviation for both models.

From the ranked lists, non-zero elements were extracted to compute Jaccard similarity, which quantifies the overlap between the sets of cue pairs with non-zero cosine similarity values. The Jaccard similarity between the human and GPT-4o mini lists was 0.1589, while the similarity between the human and Trendyol LLM v1.8 lists was 0.0793 and for the Cosmos LLaMA model the similarity was 0.1372. These results suggest that GPT-4o mini exhibits the highest degree of overlap with human data, followed by Cosmos LLaMA, while Trendyol LLM v1.8 shows the least overlap.

4.7 Response order and cosine similarity

This analysis explored the influence of response order (R1, R2, R3) on the semantic similarity between cues and responses as measured by cosine similarity. Using the pretrained Turkish FastText model (Bojanowski et al., 2017), word vectors were obtained for each cue and response, and the mean cosine similarity was calculated for each response order. This analysis was conducted on human data as well as the GPT-4o mini, Trendyol LLM v1.8 and Cosmos LLaMA datasets.

The mean cosine similarity values for each response order in the human dataset were as follows: 0.3121 for R1, 0.3156 for R2, and 0.3005 for R3. The results indicate slight variations in semantic similarity across response orders, with a minor increase observed in R2 compared to R1, followed by a decrease in R3.

In the GPT-4o mini dataset, the mean cosine similarity values were 0.4623 for R1, 0.4089 for R2, and 0.3910 for R3. These results demonstrate a notable decrease

in semantic similarity as the response order progresses, with R1 exhibiting stronger semantic associations compared to R3.

For Trendyol LLM v1.8, the mean cosine similarity values were 0.5875 for R1, 0.3827 for R2, and 0.2964 for R3. Similar to GPT-4o mini, there is a clear decrease in semantic similarity with increasing response order, with the strongest associations observed in R1 and the weakest in R3.

In the Cosmos LLaMA dataset, the mean cosine similarity values were 0.4049 for R1, 0.3576 for R2, and 0.3331 for R3. These results indicate a consistent decline in similarity as the response order progresses, resembling the trend observed in other LLMs.

Regression analyses were performed to examine the relationship between response order and cosine similarity in each dataset. In the human dataset, the regression model revealed a small but statistically significant negative relationship between response order and cosine similarity (coefficient = −0.0058, $p < 0.001$), with a very low R-squared value (0.001), indicating minimal variance explained by response order.

For GPT-4o mini, the regression model showed a stronger negative relationship between response order and cosine similarity (coefficient = −0.0357, $p < 0.001$) with an R-squared value of 0.036, suggesting that response order plays a more substantial role in explaining variance in semantic similarity compared to the human dataset.

The Trendyol LLM v1.8 dataset exhibited the strongest negative relationship between response order and cosine similarity (coefficient = −0.1455, $p < 0.001$), with an R-squared value of 0.116, indicating that response order significantly impacts the semantic similarity in this dataset. The higher coefficient and R-squared values

highlight a more pronounced effect of response order on cosine similarity for Trendyol LLM v1.8 than for GPT-4o mini or the human data.

The regression model for Cosmos LLaMA showed a statistically significant negative relationship between response order and cosine similarity (coefficient = −0.0359, $p < 0.001$), with an R-squared value of 0.027. This suggests that while response order influences semantic similarity in Cosmos LLaMA, its impact is less pronounced than in Trendyol LLM v1.8 but more substantial than in the human dataset.

These results collectively suggest that response order influences semantic similarity differently across datasets, with stronger effects observed in the LLMs compared to the human data. The decreasing trend in cosine similarity from R1 to R3 is evident in all datasets but is most pronounced in the Trendyol LLM v1.8 data, followed by GPT-4o mini and Cosmos LLaMA. Human data exhibits the least pronounced decline, indicating a relatively stable level of semantic similarity across response orders.

4.8 Graph construction and similarity metrics

The associative data from humans and large language models were transformed into graph structures. Each graph represented cues and responses as nodes, with edges weighted according to the association strengths. Graphs were constructed separately for non-stemmed (raw) and stemmed datasets for both human and LLM data, resulting in eight distinct graphs: human-raw, human-stemmed, GPT-4o mini-raw, GPT-4o mini-stemmed, Trendyol LLM v1.8-raw, Trendyol LLM v1.8-stemmed, and Cosmos LLaMA raw and Cosmos LLaMA stemmed datasets. The similarity between these graphs was quantified using Jaccard similarity, Sørensen–Dice similarity, and

overlap coefficients to evaluate the structural alignment of associations between humans and LLMs.

Graphs constructed for the cue word "*mücevher*" (gem) serve as visual examples below, showcasing differences between human and LLM-generated associations.

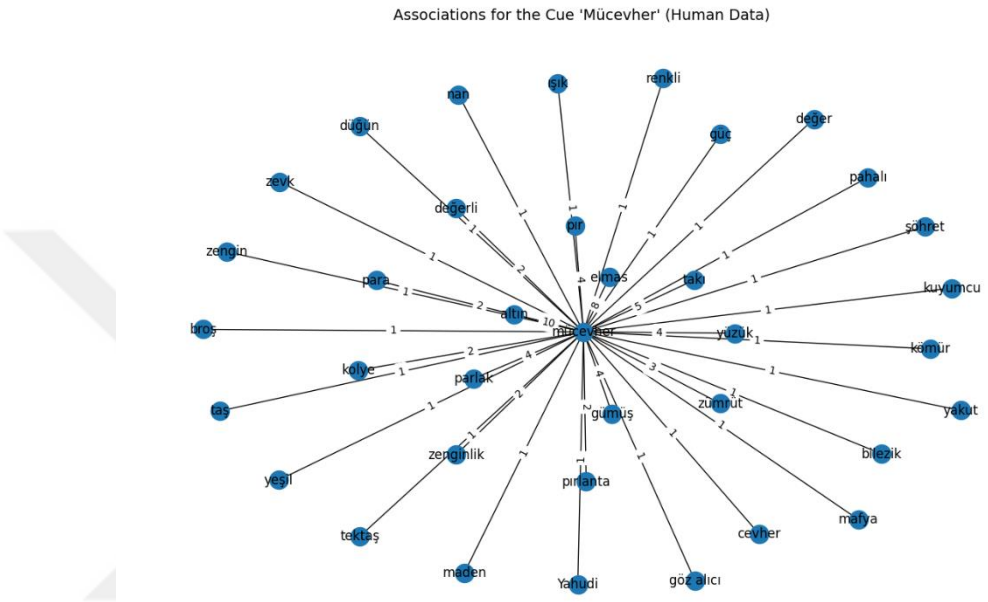


Figure 10. Graph Representation of associations for the cue "*Mücevher*" in stemmed human dataset

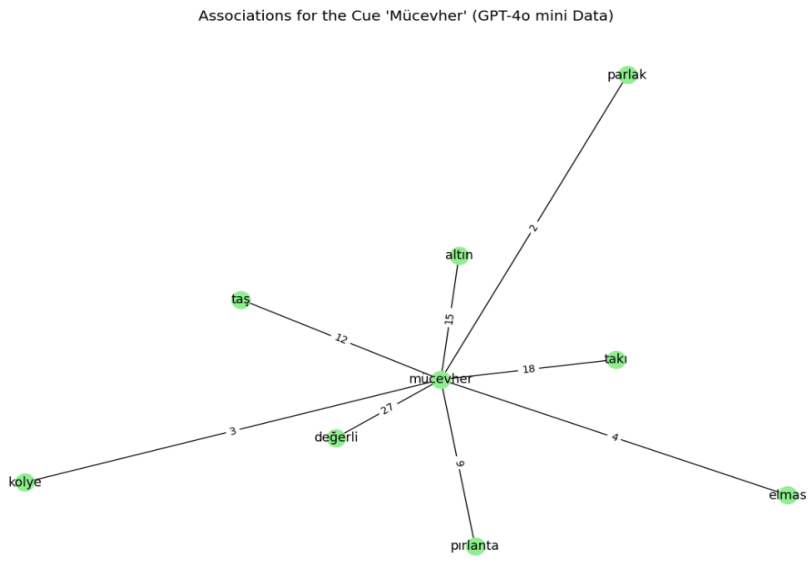


Figure 11. Graph representation of associations for the cue "*Mücevher*" in stemmed GPT-4o mini dataset

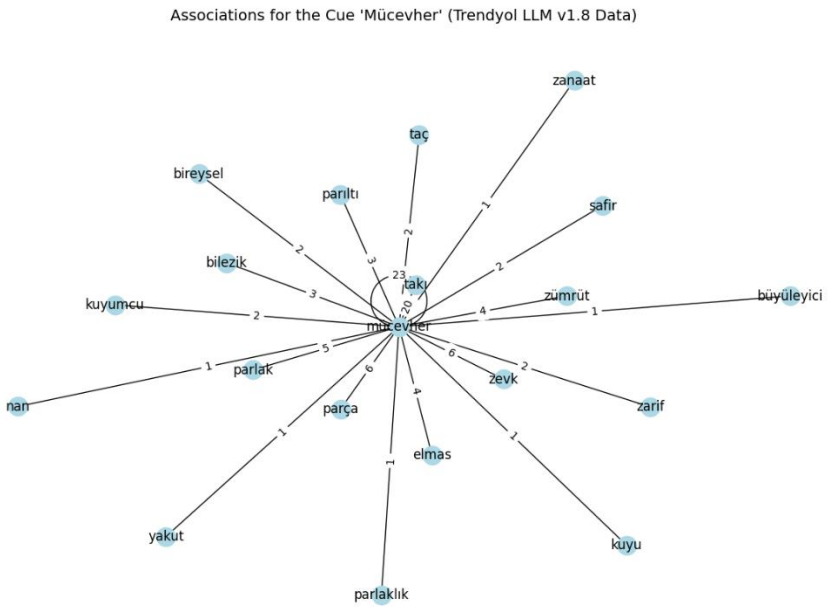


Figure 12. Graph representation of associations for the cue "*Mücevhe*r" in stemmed Trendyol LLM v1.8 dataset

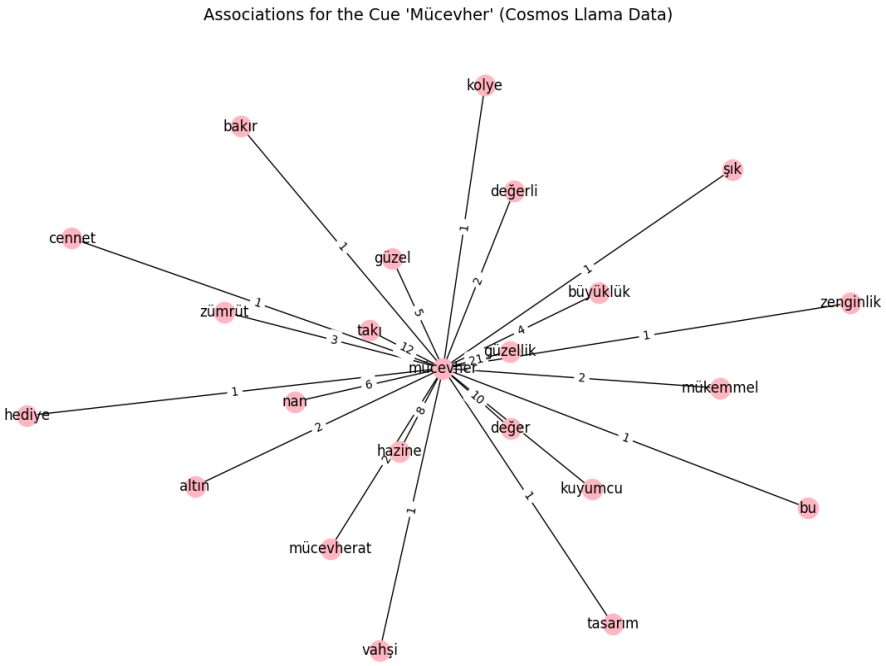


Figure 13. Graph representations of associations for the cue "*Mücevhe*r" in stemmed Cosmos LLaMA dataset

Jaccard similarity scores for node sets indicated that Cosmos LLaMA was closer to human data than Trendyol LLM v1.8 in both raw and stemmed datasets but

slightly below GPT-4o mini. For node sets in the raw dataset, GPT-4o mini achieved a score of 0.1695, Cosmos LLaMA scored 0.1640, and Trendyol LLM v1.8 scored 0.0951. In the stemmed dataset, the scores were 0.2093, 0.2018, and 0.1312 for GPT-4o mini, Cosmos LLaMA, and Trendyol LLM v1.8, respectively.

For edge sets, GPT-4o mini again outperformed the other models, with scores of 0.0637 for raw and 0.0711 for stemmed datasets. Cosmos LLaMA followed with scores of 0.0442 (raw) and 0.0484 (stemmed), while Trendyol LLM v1.8 exhibited the lowest scores, with 0.0250 (raw) and 0.0395 (stemmed). Weighted Jaccard similarity for edges in the stemmed dataset further emphasized this trend, with scores of 0.1207 for GPT-4o mini, 0.0850 for Cosmos LLaMA, and 0.0420 for Trendyol LLM v1.8.

Sørensen–Dice similarity provided additional insights into shared structural elements. For node sets in the raw dataset, GPT-4o mini scored 0.2899, followed by Cosmos LLaMA at 0.2817 and Trendyol LLM v1.8 at 0.1737. In the stemmed dataset, GPT-4o mini reached 0.3461, with Cosmos LLaMA closely following at 0.3359 and Trendyol LLM v1.8 trailing at 0.2319. For edge sets, GPT-4o mini achieved scores of 0.1197 (raw) and 0.1328 (stemmed). Cosmos LLaMA displayed slightly weaker alignment, with scores of 0.0846 (raw) and 0.0923 (stemmed). Trendyol LLM v1.8 demonstrated the lowest similarity, scoring 0.0487 (raw) and 0.0760 (stemmed).

The overlap coefficient also demonstrated closer alignment of GPT-4o mini to human data. For node sets, GPT-4o mini scored 0.6941 (raw) and 0.7251 (stemmed), while Cosmos LLaMA followed with scores of 0.5293 (raw) and 0.5845 (stemmed). Trendyol LLM v1.8 exhibited the lowest alignment, with scores of 0.2160 (raw) and 0.3577 (stemmed). For edge sets, GPT-4o mini also outperformed

other models, with scores of 0.3573 (raw) and 0.3659 (stemmed). Cosmos LLaMA scored 0.1750 (raw) and 0.1830 (stemmed), while Trendyol LLM v1.8 again trailed with scores of 0.0797 (raw) and 0.1430 (stemmed).

Overall, the analysis revealed that the similarity between human and LLM-generated association graphs was generally low across all metrics, indicating differences in associative patterns. GPT-4o mini demonstrated the highest structural similarity to human data, followed by Cosmos LLaMA, with Trendyol LLM v1.8 showing the lowest similarity. Cosmos LLaMA consistently outperformed Trendyol LLM v1.8 across all metrics and datasets (raw and stemmed), with particularly notable improvements in node-based similarities.

CHAPTER 5

DISCUSSION

This study set out to explore the nature of word associations in Turkish, comparing human responses to those generated by several state-of-the-art large language models. The primary goal was to understand the extent to which these models align with human association patterns, particularly in the context of Turkish. By examining patterns of associative strength and cue-response similarity, response diversity, morphological complexity, and concreteness effects, this research contributes to both linguistic theory and the development of more sophisticated natural language processing (NLP) tools.

This study revealed significant differences between human and LLM-generated word associations, emphasizing the complexity and richness of human associative networks compared to the systematic but limited patterns of LLMs. Key findings include the broader diversity and greater semantic coherence of human responses across response orders, as well as differences in how morphological complexity and concreteness influence associations. The investigation provides valuable insights into semantic memory and the capabilities of current AI models. The following sections discuss these findings in detail, their broader implications, and potential directions for future research.

5.1 Key findings and their implications

5.1.1 Insights from human data and cross-linguistic comparisons

5.1.1.2 Descriptive results

The most common responses in Turkish reveal universal patterns and language-specific differences when compared to larger datasets for English SWOW-EN (De Deyne et al., 2019) and Rioplatense Spanish SWOW-RP (Cabana et al., 2024). Table 17 summarizes the top 5 responses for R1 and R123 across these datasets. It is important to note that the Spanish and English datasets are much larger than the Turkish dataset in terms of the number of cue words and the number of participants, which may influence the diversity and frequency patterns observed in their results.

Table 17. Top 5 Most Common Responses in Turkish, Spanish (SWOW-RP), and English (SWOW-EN) Datasets (R1 and R123)

Rank	Turkish (R1)	Spanish (SWOW-RP R1)	English (SWOW-EN R1)	Turkish (R123)	Spanish (SWOW-RP R123)	English (SWOW-EN R123)
1	para (money)	agua (water)	money	yemek (food)	agua (water)	food
2	ağaç (tree)	comida (food)	food	para (money)	comida (food)	water
3	yemek (food)	amor (love)	water	beyaz (white)	amor (love)	love
4	beyaz (white)	animal (animal)	car	ağaç (tree)	trabajo (work)	car
5	cam (glass)	trabajo (work)	music	sarı (yellow)	dolor (pain)	red

The top responses in all datasets primarily relate to basic human needs, such as money, food, and water. This consistency suggests universal cognitive priorities in word associations. The Turkish dataset includes visually salient responses such as *beyaz* (white) and *sarı* (yellow), which are absent in the top responses for Spanish and English. These differences may reflect the linguistic and cultural contexts influencing associative patterns in Turkish.

The analysis of set sizes revealed notable differences between cues with the largest and smallest set sizes. Cues with the largest set sizes (e.g., *modernite* (modernity), *anarşist* (anarchist)) tended to be abstract concepts, eliciting a broader range of responses. In contrast, cues with the smallest set sizes (e.g., *çerçeve* (frame), *otomobil* (car)) were more concrete, yielding fewer distinct associations. This pattern may suggest that abstract concepts allow for more interpretive variability, while concrete concepts evoke more stable associations.

A key finding of this study was the progressive increase in average set sizes across response orders, from R1 to R3. For example, in the human data, the average set sizes increased from 21.24 in R1, to 24.38 in R2, and finally to 25.30 in R3, resulting in a combined average of 54.24 when all responses were considered. This pattern aligns with findings by De Deyne and Storms (2008), who demonstrated that asking for multiple responses allows researchers to capture weaker, secondary associations that would not surface in single-response tasks. In their study, set sizes grew substantially with additional responses, from 11 associations in R1 to 31 associations when all responses were combined. This increase in set size is argued to reflect the broader semantic network accessed in a continuous response task as later responses (R2, R3) often represent weaker associations (De Deyne and Storms, 2008). The diversity of later responses also highlights the exploratory nature of

associative processes, where participants move beyond immediate connections in R1 to uncover less prominent associations in R2 and R3.

5.1.1.3 Morphological analysis

The morphological analysis revealed several notable patterns in the relationship between cues and their responses. The main questions addressed here were:

- Does the morphological complexity of cues, measured by morpheme count, influence the morphological complexity of responses, and does this influence vary across response orders? Specifically, it is expected that responses to morphologically complex cues would also exhibit higher morphological complexity, and this effect might vary across response orders (R1, R2, R3).
- To what extent are the morphemes of cues retained in the morphemes of responses, and does this retention change across response orders?

Results showed that morphological complexities showed weak but statistically significant correlations between the cue and responses across all response orders, while Poisson regression did not identify a significant relationship between response order and response complexity. These findings suggest that the morphological complexity of the cue has a minimal influence on the complexity of the responses, and this relationship remains stable across different response orders.

Although suffix overlap appeared to decrease slightly as the response order progressed, approximately 50% of the cue suffixes were retained across all response orders. The Friedman test found no significant differences in suffix overlap across response orders ($\chi^2 = 3.119$, $p = 0.210$). These results indicate a consistent morphological relationship between cues and responses and that participants often

maintain some morphological connection to the cue, even as they move on to later responses.

The findings do not indicate that the morphological complexity of cues systematically influences the complexity of responses. Instead, the stable retention of morphemes suggests a consistent morphological connection between cues and responses. This pattern implies that participants rely on morphological features to some degree when forming associations, even though response complexity does not vary significantly across response orders.

5.1.1.4 Concreteness analysis

The concreteness analysis revealed significant positive correlations between cue and response concreteness across all response orders. These findings indicate that cue concreteness affects response concreteness, with more concrete cues eliciting more concrete responses, and more abstract cues prompting more abstract responses. Results from the regression model further supported this conclusion with cue concreteness having a relatively large positive coefficient ($\beta = 0.347$), meaning that the concreteness of the cue has a noticeable impact on how concrete the responses are. Regression results also showed that response order affects response concreteness. Specifically, concreteness was observed to slightly decrease as the response order progressed from R1 to R3.

The decreasing trend in correlations and concreteness across response orders suggests that the influence of cue concreteness diminishes in later responses. While initial responses may closely reflect the concreteness of the cue, later responses are likely influenced by weaker associations, leading to more abstract connections, although the effect is small. This transition aligns with the spreading activation

theory (Collins & Loftus, 1975) of semantic memory, where activation spreads further from the initial cue, reaching less related nodes in the network.

5.1.1.5 Correlations with *AnlamVer* dataset

The analysis revealed strong correlations between human-generated associations and the *AnlamVer* dataset for both similarity and relatedness scores that were gathered from human participants. These findings suggest that human associative patterns closely align with established semantic relationships in Turkish.

The association-based cosine similarities correlated more strongly with relatedness scores than with similarity scores. Relatedness refers to the broader conceptual connections between words (e.g., "dog" and "bark"), while similarity captures shared attributes or features (e.g., "dog" and "wolf"). Associations seem to often reflect conceptual links rather than shared attributes. For example, a cue like *barınak* “shelter” might elicit *köpek* “dog” or *kedi* “cat” as responses, which are conceptually related but not similar.

5.1.2 Performance of LLMs

The comparison between human data and LLM-generated responses provides insights into the capabilities and limitations of large language models. Three models were evaluated: GPT-4o mini, Trendyol LLM v1.8, and Cosmos LLaMA. Each model's results are discussed below, including a comprehensive comparison to human data, followed by a conclusion highlighting the best-performing model.

5.1.2.1 GPT-4o mini

5.1.2.1.1 Cue-Response similarity and diversity

GPT4o Mini displayed limited alignment with human data in terms of the most and least common responses. While some frequent human responses, such as *yemek* “food” and *ağaç* “tree”, were replicated by the model, their rank was lower compared to human data. Moreover, only *yemek* was in common to other languages such as English and Rioplatense Spanish. In terms of set sizes, GPT-4o mini generated significantly smaller average set sizes compared to humans across all response orders (R1, R2, R3). For example, while human participants displayed an average set size of 21.24 in R1, GPT-4o mini produced an average of 14.32, a substantial difference (d = 2.85). A possible explanation for this phenomenon is that language model outputs can be viewed as an averaged representation over multiple potential samples, as pretraining involves texts from numerous authors with varying styles and contexts. This averaging effect, combined with the model's tendency to prioritize high-probability associations during sampling, might result in less diverse outputs. The model’s Poisson regression results supported this result, which exhibited a lower intercept (0.976, $p < 0.001$) compared to human data (2.91, $p < 0.001$). GPT-4o mini seems to not expand its associative network as effectively as humans across subsequent response orders.

A logistic regression analysis further revealed that humans' greater variety of responses significantly distinguished them from GPT-4o mini. With a pseudo R-squared value of 0.89, the model showed that each additional distinct response increased the likelihood of the source being human by 1.07 log-odds ($p < 0.001$), underscoring humans' richer and more varied associative capabilities.

GPT-4o mini's strongest cue-response associations, such as *petrol-enerji* (petrol-energy) and *lokanta-restoran* (eatery-restaurant), demonstrate that the model is capable of identifying semantically logical relationships. However, these associations differ from human patterns.

The model exhibited strong correlations with the *AnlamVer* dataset's similarity and relatedness scores, aligning well with structured semantic relationships. Interestingly, unlike human data, GPT-4o mini showed stronger correlations for similarity scores than relatedness scores. This discrepancy may arise from humans' tendency to provide more contextually grounded associations, whereas the model appears to depend more heavily on shared features or attributes, reflecting its computationally driven approach.

5.1.2.1.2 Morphological analysis

In terms of morphological analysis, GPT-4o mini demonstrated stronger correlations between cue and response complexities compared to human data, indicating a greater sensitivity to morphological features when generating responses. This conclusion is supported by a Mann-Whitney U test, which revealed a statistically significant difference in morphological complexities between GPT-4o mini and human responses, though the effect size was small (Cohen's d values of −0.10 (R1), −0.07 (R2), −0.03 (R3), and −0.07 (overall)). However, response complexity remained stable across response orders, mirroring human behavior, as the Poisson regression did not identify significant relationships for either source. For suffix overlap, GPT-4o mini displayed patterns closely aligning with humans, as statistical comparisons showed no significant differences. This suggests that the model retains a similar

proportion of suffixes across response orders, mirroring human-like behavior in this aspect.

5.1.2.1.3 Concreteness analysis

Notably, GPT-4o mini demonstrated stronger correlations between cue and response concreteness compared to human data, indicating that the model relies more heavily on cue concreteness when generating responses. This heightened reliance is evident in the regression model, which identified cue concreteness and response order as significant predictors of response concreteness. However, the larger negative coefficient for response order in GPT-4o mini data suggests that the model experiences a more pronounced decline in concreteness across response orders than humans. This could reflect the model's tendency to shift toward increasingly abstract associations more rapidly than human participants. When comparing human and GPT-4o mini responses directly, regression results revealed that GPT-4o mini responses were overall less concrete than human responses. While both cue concreteness and response order significantly influenced response concreteness in both datasets, the negative coefficient for GPT responses further highlights the model's limited capacity to maintain concreteness in later responses. The explained variance was higher for GPT-4o mini data (31%) than human data (23%), suggesting that the model's response patterns are more systematically tied to cue properties and response order, but at the cost of diversity and flexibility.

5.1.2.1.4 Structural and sequential alignment

The analysis of cosine similarities of cue pairs, highlighted differences in how GPT-4o mini aligns with human patterns. This analysis was based on the cosine rankings of cue pairs acquired from their associative strength vectors, and was used to assess the alignment of humans and LLMs on these rankings. While both datasets identified similar high-ranking pairs, such as *orman–ormanlık* (forest–forested area) and *saydam–transparan* (transparent–transparent), the Spearman correlation coefficient (0.3341) and Kendall's tau (0.2271) between human and GPT-4o mini cosine similarity scores indicated only moderate alignment. These values suggest that while GPT-4o mini captures some human-like associations, it exhibits notable divergences in how association strengths are distributed.

The edit distance analysis, which compared the ranked lists of cosine similarities, revealed a normalized Levenshtein distance of 0.99 between GPT-4o mini and human data. This high value indicates substantial structural deviations in the ranked association lists, suggesting that the sequence of association strengths differs significantly between the two sources.

Jaccard similarity further quantified the overlap between sets of cue pairs with non-zero cosine similarity values. With a Jaccard similarity of 0.1589, the overlap between human and GPT-4o mini was limited, underscoring the model's constrained ability to replicate the diversity of associative patterns seen in human data.

Response order analysis revealed notable differences in semantic similarity trends between humans and GPT-4o mini. While human data showed minor variations across response orders, GPT-4o mini exhibited a pronounced decline in cosine similarity from R1 (0.4623) to R3 (0.3910). Regression analyses supported

this finding, with GPT4o Mini showing a stronger negative relationship between response order and cosine similarity (coefficient = −0.0357, R-squared = 0.036) compared to humans (coefficient = −0.0058, R-squared = 0.001). These results suggest that GPT-4o mini struggles to maintain semantic coherence in later response orders, reflecting a systematic limitation.

In summary, while GPT-4o mini aligns with human data in capturing some high-probability associations, its structural deviations, limited overlap in association patterns, and declining semantic similarity across response orders highlight its challenges in mimicking the complexity and flexibility of human associative networks.

5.1.2.1.5 Graph-based similarity metrics

The graph-based analysis revealed that GPT-4o mini demonstrated higher structural similarity to human data compared to other models, though overall alignment with human graphs remained low. Jaccard similarity scores for node and edge sets indicated that GPT-4o mini consistently achieved higher overlap with human data. Notably, the stemmed dataset improved alignment, with GPT-4o mini outperforming other models in both node (0.2093) and edge sets (0.0711). These findings suggest that stemmed responses help bridge some of the structural gaps between GPT-4o mini and human association networks.

Sørensen–Dice similarity scores further highlighted GPT-4o mini's relative success in replicating human-like structural elements. The model achieved higher alignment for both node and edge sets compared to other models, particularly in the stemmed dataset (0.3461 for nodes and 0.1328 for edges). The overlap coefficient

reinforced this trend, with GPT-4o mini scoring highest across all metrics, reflecting its ability to replicate associative connections.

Despite these strengths, the overall similarity between GPT-4o mini and human-generated association graphs remained limited. The lower scores across most metrics suggest fundamental differences in the associative processes of the model compared to humans. While GPT-4o mini shows promise in capturing some structural aspects of human associative networks, especially when aided by preprocessing techniques like stemming, its limitations highlight the challenges of fully mimicking the complexity and nuance of human associations.

5.1.2.2 Trendyol LLM v1.8

5.1.2.2.1 Cue-Response similarity and diversity

Trendyol LLM v1.8 displayed no overlap with human data or other languages in the most or least common responses, indicating distinct associative patterns. Among the most frequent responses, it shared only *şeffaf* (transparent) with GPT-4o mini, highlighting a limited alignment even between models. The model generated smaller average set sizes compared to humans but larger ones than GPT-4o mini across all response orders (R1, R2, R3). For instance, while humans exhibited an average set size of 21.24 in R1, Trendyol LLM v1.8 produced an average of 8.04 (d = 2.11). These constrained set sizes suggest that the model's semantic network is less extensive than humans, though slightly broader than GPT-4o mini's.

A logistic regression analysis revealed that Trendyol LLM v1.8's distinct responses were significantly less varied than humans'. With a pseudo R-squared value of 0.51, the analysis showed a smaller increase in log-odds for the source being

human compared to GPT-4o mini, underscoring the model's limited associative diversity relative to humans.

Trendyol LLM v1.8's strongest cue-response associations, such as *uzak-uzak* (far-far) and *noksan-eksik* (deficient-lacking), reflect a reliance on high-probability and semantically similar pairings. This pattern may stem from the model's sampling strategies, which prioritize associations with higher confidence.

The model exhibited only slightly weaker correlations with the AnlamVer dataset's similarity and relatedness scores compared to humans. Similar to GPT-4o mini, Trendyol LLM v1.8's correlations with similarity scores were stronger than with relatedness scores, diverging from human patterns where relatedness correlations were typically higher. This suggests that, like GPT-4o mini, Trendyol LLM v1.8 relies more on shared features or attributes rather than contextual or broader associative reasoning.

5.1.2.2.2 Morphological analysis

This analysis investigated how morphological complexity and suffix patterns influence the generative processes of Trendyol LLM v1.8 compared to human responses. Trendyol LLM v1.8 demonstrated stronger correlations between cue and response morphological complexities compared to human data, particularly in R1 ($r = 0.19$ vs. $r = 0.073$ for humans). While these correlations diminished slightly in R2 and R3, they remained higher than the weak yet consistent correlations observed in humans across all response orders.

Despite these stronger correlations, Poisson regression identified no significant relationship between response complexity and response order for Trendyol LLM v1.8 ($\beta = 0.0026$, $p = 0.188$), indicating that the model does not

systematically vary morphological complexity as response orders progress. This aligns with human data, where no significant relationship was observed either.

In terms of suffix overlap, Trendyol LLM v1.8 exhibited greater Jaccard distances compared to humans across all response orders, with a notable increase in later orders. These larger distances reflect a reduced tendency to retain morphological features of the cues, diverging from the more stable suffix overlap patterns observed in humans. Statistical comparisons confirmed significant differences in suffix overlap patterns between Trendyol LLM v1.8 and human data, consistent with the model's broader generative patterns and higher morphological variability.

Trendyol LLM v1.8 exhibits a heightened sensitivity to morphological complexity compared to humans, but its inability to establish a systematic relationship between complexity and response order, coupled with greater divergence in suffix overlap patterns, underscores its limitations in achieving human-like morphological consistency. These results suggest that the model relies on surface-level morphological cues rather than deeper linguistic structures, leading to a less cohesive generative process.

5.1.2.2.3 Concreteness analysis

Trendyol LLM v1.8 demonstrated the strongest correlations between cue and response concreteness among all datasets, indicating a pronounced reliance on the concreteness of cues when generating responses. This suggests that the model uses cue concreteness as a key factor in guiding its associative process. While regression analysis revealed that the concreteness of Trendyol LLM v1.8's responses declined across response orders, this decline was more pronounced than in human data,

reflecting the model's tendency to shift toward less concrete associations as it generates further responses. Despite this pattern, Trendyol LLM v1.8’s responses remained closer to human concreteness levels compared to GPT-4o mini, highlighting a stronger alignment with human-like patterns in this dimension. These findings suggest that while Trendyol LLM v.18 is influenced by cue concreteness, its generative approach may lack the contextual flexibility humans exhibit, leading to a more systematic but less nuanced handling of concreteness across response orders.

5.1.2.2.4 Structural and sequential alignment

Cosine similarity analysis revealed that Trendyol LLM v1.8 exhibited limited alignment with human association patterns. The Spearman correlation coefficient (0.3286) and Kendall’s tau (0.2212) for Trendyol LLM v1.8’s cue similarity rankings were comparable to GPT-4o mini but still demonstrated significant divergence from human data. These moderate correlations suggest that while the model captures some patterns in human-like associative strength, the overall structure of its associations is distinct.

Edit distance and Jaccard similarity analysis further highlighted these structural differences. Trendyol LLM v1.8's Jaccard similarity for non-zero cosine similarity cue pairs was the lowest among the models evaluated (0.0793), reflecting minimal overlap in the associations it forms compared to human data. The normalized edit distance between the ranked lists of cosine similarities for Trendyol LLM v1.8 and humans was 0.99, indicating substantial structural deviations in the order of ranked associations. For instance, while Trendyol LLM v1.8 did generate some semantically related pairs such as *orman–ormanlık* (forest–forested area) and *saydam–transparan* (transparent–transparent), these pairs were ranked differently

from human data, where both pairs were among the strongest associations. This difference in ranking suggests that while Trendyol LLM v1.8 can replicate certain high-probability associations, the importance it assigns to such pairs diverges from human patterns.

Semantic similarity trends across response orders showed the most pronounced decline in cosine similarity for Trendyol LLM v1.8 among all datasets. The mean cosine similarity decreased from R1 (0.5875) to R3 (0.2964), indicating that the model struggled to maintain semantic coherence as it generated responses sequentially. In contrast, humans showed greater stability in their semantic similarity across response orders, with only a slight variation from R1 (0.3121) to R3 (0.3005).

Regression analysis supported these findings, with Trendyol LLM v1.8 exhibiting the strongest negative relationship between response order and cosine similarity (coefficient = −0.1455, R^2 = 0.116). This indicates that response order plays a much larger role in the semantic decline for Trendyol LLM v1.8 compared to GPT-4o mini or human data. Such a sharp decline underscores a key limitation of Trendyol LLM v1.8: its inability to sustain meaningful and coherent associations over sequential generations.

These findings collectively illustrate that while Trendyol LLM v1.8 can replicate certain human-like associations, its overall structural and sequential alignment with human data remains limited. The sharp decline in semantic similarity across response orders, its constrained overlap in association patterns, and the high edit distance suggest a reliance on statistical regularities rather than the richer, more flexible associative processes observed in human cognition.

5.1.2.2.5 Graph-based similarity metrics

Graph-based analysis revealed significant differences between the associative networks of Trendyol LLM v1.8 and human participants. When comparing the graphs, Trendyol LLM v1.8 consistently achieved the lowest similarity scores among the models evaluated, highlighting fundamental divergences in its associative processes.

For node-based comparisons, Jaccard similarity scores for Trendyol LLM v1.8 were 0.0951 in the raw dataset and 0.1312 in the stemmed dataset. These values were substantially lower than those for GPT-4o mini or Cosmos LLaMA, indicating a more limited overlap in the sets of cues and responses generated by Trendyol LLM v1.8 and humans. Sørensen–Dice similarity scores further reinforced this trend, with Trendyol LLM v1.8 scoring 0.1737 (raw) and 0.2319 (stemmed). The increase in similarity for the stemmed dataset suggests that reducing morphological variation slightly improved alignment between the model and human graphs, but the overall gap remained significant.

Edge-based comparisons provided further evidence of structural divergence. Low scores across these comparisons reflect Trendyol LLM v1.8's tendency to form associations that differ not only in content but also in strength from human-generated associations. The overlap coefficient, which measures the alignment of associations, revealed a similar pattern. For node sets, Trendyol LLM v1.8 scored 0.2160 (raw) and 0.3577 (stemmed), significantly lower than both GPT-4o mini and Cosmos LLaMA. For edge sets, Trendyol LLM v1.8 exhibited the lowest scores among all models, with 0.0797 (raw) and 0.1430 (stemmed). This suggests that its associative connections failed to capture the richness and variability observed in human networks.

Overall, the graph-based analysis underscores the challenges Trendyol LLM v1.8 faces in replicating human-like associative networks. Although stemming improved alignment slightly, results highlight significant gaps in its ability to mirror the complex and contextually rich associations seen in human cognition.

5.1.2.3 Cosmos LLaMA

5.1.2.3.1 Cue-Response similarity and diversity

Cosmos LLaMA exhibited intermediate alignment with human associative patterns, outperforming Trendyol LLM v1.8 but underperforming compared to GPT-4o mini. The model generated diverse associations, with the most frequent responses such as *yemek* (food) and *ağaç* (tree) overlapping with high-frequency human responses and *yemek* overlapping with English and Rioplatense Spanish as well. However, notable differences in rank and set size highlight areas where Cosmos LLaMA diverges from human generative tendencies. For instance, the average set sizes of 5.81 (R1), 8.48 (R2), and 10.71 (R3) were significantly smaller than human averages (Cohen's d values as 2.46 for R1, 2.42 for R2, 1.77 for R3, and 2.54 for combined responses). The Poisson regression model, which revealed a coefficient of 0.298 ($p < 0.001$), suggests that Cosmos LLaMA exhibits a moderate expansion in associative scope across response orders, though this remains less pronounced than human patterns.

Cosmos LLaMA's logistic regression results further emphasize its limited diversity, with each additional distinct response increasing the likelihood of the source being human by 68.4%. This value, while higher than Trendyol LLM v1.8, remains below that of GPT-4o mini, underscoring its intermediate performance in replicating human-like diversity.

Cosmos LLaMA's correlations with the AnlamVer dataset provide further insights into its alignment with human semantic processes. While Cosmos LLaMA's similarity correlation ($r = 0.566$, $p < 0.001$) exceeded that of human data, its relatedness correlation ($r = 0.539$, $p < 0.001$) was notably weaker. This disparity suggests that Cosmos LLaMA emphasizes shared features and semantic similarity more than broader conceptual connections, which are more characteristic of human associative patterns.

Regression analyses further highlighted these differences. For human data, cosine similarity explained 25% of the variance in similarity scores and 43% in relatedness scores, reflecting humans' ability to balance specific semantic features with broader relational contexts. For Cosmos LLaMA, cosine similarity accounted for 32% of the variance in similarity scores and 29% in relatedness scores. These findings indicate that while Cosmos LLaMA demonstrates a systematic approach to semantic similarity, it struggles to replicate the nuanced interplay of similarity and relatedness evident in human cognition.

5.1.2.3.2 Morphological analysis

Cosmos LLaMA's approach to morphological complexity reveals both alignment with human associative processes and critical divergences. While humans exhibit weak correlations between cue and response morphological complexity (e.g., $r = 0.073$ in R1), Cosmos LLaMA's correlations are consistently stronger, reaching $r = 0.158$ in R1 and decreasing slightly across later response orders. While the effect sizes between human and Cosmos LLaMa's morphological complexities are negligible (e.g., $d = 0.07$ for R1), the difference is statistically significant ($U = 124{,}430{,}877.0$, $p < 0.001$). This pattern suggests that Cosmos LLaMA relies more on

morphological cues when forming associations compared to humans, whose associations are likely influenced by a broader interplay of linguistic and conceptual factors.

Interestingly, Cosmos LLaMA mirrors human tendencies in retaining suffixes across response orders, as indicated by stable Jaccard distances (e.g., 0.55 in R1, 0.58 in R2, and 0.58 in R3). However, despite this alignment, the overall distances remain different from human patterns, reflecting a less nuanced integration of morphological features. Humans appear to use morphology as one of many tools in their associative repertoire, whereas Cosmos LLaMA's reliance on morphology appears more systematic and less flexible.

Another notable difference lies in how response order influences morphological complexity. For humans, response complexity remains relatively stable across orders, with no significant trends observed in regression analyses (β = 0.0012, p = 0.702). In contrast, Cosmos LLaMA exhibits a weak but significant positive relationship (β = 0.0059, p = 0.014), suggesting that as the model generates later responses, it slightly increases morphological variation.

These results underscore a key limitation of Cosmos LLaMA: while it captures some morphological patterns, its reliance on these features as a primary associative mechanism contrasts with the more flexible and diverse strategies employed by humans. This suggests that the model's associative processes are still constrained more by structural regularities rather than the nuanced, contextually rich pathways that characterize human cognition.

5.1.2.3.3 Concreteness analysis

Humans demonstrate a natural tendency to align response concreteness with cue concreteness, but this relationship is balanced by contextual flexibility. This dynamic is captured by positive correlations between cue and response concreteness in human data, but these correlations diminish as response order progresses, reflecting humans' ability to move beyond immediate and tangible connections. Cosmos LLaMA, in contrast, exhibits stronger correlations between cue and response concreteness across response orders than humans, which highlights a more systematic dependence on the concreteness of cues. While this reliance may enable Cosmos LLaMA to produce consistent, logical associations, it limits the model's capacity to generate the diverse and contextually rich responses typical of human cognition. For example, the model's concrete response patterns diminish more slowly across response orders compared to human data, suggesting it struggles to break away from initial, high-concreteness anchors. This trend aligns with the regression results, where Cosmos LLaMA exhibits a more pronounced reliance on cue properties, as reflected in its regression coefficients and the variance explained by cue concreteness.

5.1.2.3.4 Structural and sequential alignment

In Cosmos LLaMA, none of the top-ranking pairs overlap with those observed in human data. Instead, the model prioritizes pairs such as *öldürülüşündeki–ölüm* (in-the-death) and *alışkın–alışkanlık* (accustomed–habit). This absence of shared high-ranking pairs indicates that Cosmos LLaMA’s associative processes diverge fundamentally from human cognition, reflecting a more constrained focus on surface-level relationships rather than the broader conceptual links evident in human associations.

The correlations between human and Cosmos LLaMA cosine similarity scores further emphasize this divergence. Cosmos LLaMA achieved a Spearman correlation of $r = 0.113$ and a Kendall's tau of $\tau = 0.081$ (both $p < 0.001$, significantly lower than the correlations observed for GPT-4o mini and closer to Trendyol LLM v1.8. These weak correlations suggest that Cosmos LLaMA struggles to align its associative strengths with the nuanced patterns of human cognition, which balance feature-based and context-driven associations more effectively.

Structural differences between human and Cosmos LLaMA associative networks are further underscored by edit distance and Jaccard similarity analyses. The normalized Levenshtein distance of 0.99 between their ranked cosine similarity lists highlights substantial structural deviations, indicating that Cosmos LLaMA's rankings require extensive reordering to align with human data. This high distance reflects the model's struggle to capture the nuanced prioritization of associations observed in human rankings.

Jaccard similarity, measuring the overlap in non-zero cosine similarity cue pairs, shows moderate alignment between Cosmos LLaMA and human data (0.1372), outperforming Trendyol LLM v1.8 but falling behind GPT-4o mini. The absence of shared high-ranking pairs highlights a significant limitation: while Cosmos LLaMA captures some common associations in lower-ranking pairs, it fails to replicate the conceptual diversity and flexibility inherent in human associative patterns.

The influence of response order on semantic similarity reveals another key limitation in Cosmos LLaMA's associative processes. Humans maintain relatively stable semantic similarity across response orders, with only minor decreases from R1 (0.3121) to R3 (0.3005). Cosmos LLaMA, by contrast, exhibits a more pronounced decline in cosine similarity as response order progresses, starting at 0.4049 in R1 and

dropping to 0.3331 in R3. Regression analyses reveal a significant negative relationship between response order and cosine similarity ($\beta = -0.0359$, $R^2 = 0.027$, $p < 0.0011$). This decline suggests that Cosmos LLaMA's ability to sustain semantically similar associations diminishes more rapidly than humans', indicating a systematic limitation in its capacity to generate contextually similar responses over time.

5.1.2.3.5 Graph-based metrics

Cosmos LLaMA achieves moderate alignment with human data at the node level, with Jaccard similarity scores of 0.1640 for raw graphs and 0.2018 for stemmed graphs. These scores, while higher than those of Trendyol LLM v1.8, remain lower than GPT-4o mini's scores. The improvement in the stemmed dataset indicates that reducing morphological variation enhances the structural alignment between Cosmos LLaMA and human associative networks. However, the scores suggest that Cosmos LLaMA still struggles to fully capture the diversity of nodes –representing cue-response pairs– evident in human data.

Sørensen–Dice similarity, which accounts for shared elements weighted by their total size, paints a similar picture. Cosmos LLaMA (0.0846 for raw graph and 0.0923 for stemmed graph) outperforms Trendyol LLM v1.8 but falls short of GPT-4o mini. These results suggest that while Cosmos LLaMA captures some small portion of human-like node distributions, its associative network lacks the breadth and variability of human responses.

At the edge level, which represents the associative strengths between nodes, Cosmos LLaMA performs less effectively. Jaccard similarity scores for edges are 0.0442 for raw graphs and 0.0484 for stemmed graphs, indicating minimal overlap

with human data. These scores are slightly higher than Trendyol LLM v1.8 (0.0250 raw, 0.0395 stemmed) but significantly lower than GPT-4o mini (0.0637 raw, 0.0711 stemmed). This limited alignment reflects Cosmos LLaMA's difficulty in replicating the nuanced strength relationships between cues and responses seen in human data. Sørensen–Dice similarity reinforces this finding, with Cosmos LLaMA again outperforming Trendyol LLM v1.8 but underperforming compared to GPT-4o mini. These low scores emphasize the model's struggle to replicate the structural richness of human associative networks, where edges not only connect nodes but also reflect varied strengths that capture deeper semantic and contextual relationships.

The overlap coefficient provides further insights. For node sets, Cosmos LLaMA scores 0.5293 for raw graphs and 0.5845 for stemmed graphs, outperforming Trendyol LLM v1.8 (0.2160 raw, 0.3577 stemmed) but again lagging behind GPT-4o mini (0.6941 raw, 0.7251 stemmed). These results indicate that Cosmos LLaMA captures some associations similar to humans. For edge sets, Cosmos LLaMA's scores are significantly lower, with 0.1750 for raw graphs and 0.1830 for stemmed graphs. This further highlights the model's inability to replicate the nuanced distributions between cue-response pairs that characterize human associative networks.

5.2 Summary

The study provides critical insights into Turkish word associations and their comparison to LLM-generated patterns. Human responses revealed both universal and language-specific trends. Universally, associations such as references to food and money aligned with English and Spanish datasets, reflecting shared cognitive priorities. Uniquely, Turkish responses included visually salient terms like *beyaz*

(white) and *sarı* (yellow), shaped by linguistic and cultural contexts. Abstract cues elicited broader response sets, while concrete cues generated fewer and more stable associations. Response diversity increased across response orders, reflecting the depth of human semantic exploration.

Morphological analyses of human data demonstrated weak but consistent correlations between cue and response complexities, suggesting limited influence of morphological complexity on associative patterns. The retention of suffixes was stable across response orders, with approximately half of the cue suffixes maintained in responses, indicating a cognitive preference for morphological ties. Concreteness analyses revealed strong positive correlations between cue and response concreteness, where concrete cues elicited more concrete responses. This influence diminished slightly in later responses, suggesting that secondary associations become less bound to the concreteness of the cue.

While all LLMs showed some degree of alignment with human data, the overall similarities and structural matches were limited. Among the LLMs, GPT-4o mini demonstrated the highest alignment but still struggled to replicate the diversity and semantic coherence of human associations, particularly across later response orders. Trendyol LLM v1.8, fine-tuned using LoRA (Hu et al., 2021) on the Mistral 7B model (Jiang et al., 2023), performed well in leveraging cue concreteness but exhibited constrained diversity and weak structural alignment. Cosmos LLaMA, fine-tuned using the Direct Preference Optimization (DPO) (Rafailov et al., 2024) technique, displayed intermediate performance, showing some alignment in morphological retention but falling short in flexibility and associative richness. The use of DPO may have contributed to Cosmos LLaMA's responses being closer to

human patterns, as this technique is designed to optimize alignment with human preferences.

In summary, GPT-4o mini outperformed other models in structural alignment and graph-based metrics, while Trendyol LLM v1.8 and Cosmos LLaMA demonstrated strengths in narrower aspects such as concreteness and morphological features. However, no model fully captured the complexity, diversity, and nuanced patterns evident in human associations. These findings emphasize the importance of advancing LLMs to better simulate human cognition, particularly for morphologically rich and culturally unique languages like Turkish.

5.3 Broader implications

This study makes contributions to linguistic theory and the advancement of natural language processing systems. The findings emphasize the intricate interplay between linguistic and cognitive processes in shaping word associations, offering valuable insights into the mental lexicon and guiding improvements in AI language models.

By uncovering the unique characteristics of Turkish word associations, this research deepens our understanding of the mental lexicon. It highlights the balance between universal cognitive patterns and language-specific influences, as well as the interplay of abstractness, concreteness, and morphological consistency in shaping semantic networks. These findings reveal how Turkish speakers navigate their semantic environments in nuanced and structured ways.

The study also identifies areas for enhancing the performance of large language models, particularly in their ability to generate more diverse associations. Incorporating linguistic insights from Turkish can help LLMs adapt more effectively

to morphologically rich languages, improving both accuracy and contextual relevance.

Furthermore, the practical applications of these insights extend to tasks such as semantic search, contextual text generation, and cognitive modeling. They also hold potential for enhancing AI-driven educational tools and psycholinguistic simulations, where a deeper understanding of associative processes may be essential.

5.4 Limitations and future directions

This study offers valuable insights into human and LLM word associations; however, several limitations must be acknowledged.

First, while the participant sample represents Turkish speakers, it does not fully capture the linguistic and demographic diversity of the population. The sample, primarily composed of university students, may have influenced associative patterns due to their age and higher education levels. Factors such as regional dialects, age, and socio-cultural differences can result in varying word associations. For example, children, older adults, and individuals from diverse educational backgrounds may exhibit distinct associative patterns. To enhance generalizability, future research should include a broader and more diverse range of demographic groups.

Second, the quality of word association simulations with LLMs is highly dependent on the design of prompts and the configuration of model parameters. Tailored prompts that elicit more diverse and contextually nuanced associations could improve output quality such as those in Aher et al. (2023)'s study where demographic information was included in the prompts. Additionally, hyper-parameters, such as temperature and top-p settings, significantly influence model behavior. Higher temperature values encourage diverse and creative responses, while

top-p settings control the probability distribution of potential outputs, balancing the creativity and the sample space considered during response generation. These factors may explain GPT4o-mini's greater variability in its outputs compared to other models. This underscores the importance of optimizing prompts and tuning model parameters to align with specific word association task objectives. Future studies should focus on refining these elements to better harness LLMs' potential in generating human-like associations.

Third, in this study, each LLM was used with the same prompt to generate multiple trials, simulating an experiment with multiple individuals. Future research could explore differences between outputs generated in one human-to-one LLM comparisons, offering a more granular perspective on how closely an LLM's responses align with those of an individual human participant.

The use of FastText embeddings in certain analyses may present another limitation. While FastText excels at capturing string-level similarities, it may underperform in representing deeper semantic relationships compared to other embedding methods. This could have influenced specific results. However, prior research, such as Cabana et al. (2023), indicates that FastText embeddings align closely with models like word2vec and GloVe in word similarity tasks, offering some validation for their use in word association studies.

Lastly, an unexplored analytical avenue involves utilizing the raw probabilities of words generated by LLMs in response to prompts. Examining the top *n* probable words for a cue word could reveal richer insights into LLM-derived word associations and facilitate more equitable comparisons with human responses, warranting further exploration in future studies.

Future research should address these limitations by expanding participant diversity, optimizing prompt and model configurations. These efforts will deepen our understanding of the interplay between human cognition and artificial intelligence in word associations, contributing to advancements in both fields.

CHAPTER 6

CONCLUSION

This thesis investigated the similarities and differences between human word association patterns and those generated by large language models (LLMs), with a focus on set size, associative strengths, linguistic sensitivity to morphology and concreteness, and the structural alignment of their resulting associative networks. The primary objective was to assess how well LLMs simulate human-like behavior in word association tasks, particularly in the Turkish language, and to understand the broader implications of these findings for cognitive and linguistic research.

To address this research question, a comprehensive approach was adopted. A word association experiment was conducted with Turkish-speaking human participants, who were tasked with providing up to three associative responses to a series of cue words. The same task was simulated using selected LLMs, including GPT-4o mini, Trendyol-LLM v1.8, and Cosmos LLaMA, under conditions designed to mirror those of the human experiment. The resulting datasets were then analyzed and compared using a range of metrics, including set size, association strength, cosine similarity, Jaccard similarity, and graph-based metrics, to explore the overlap and divergence in associative patterns.

The results revealed fundamental differences between human and LLM-generated word associations, despite some minor overlaps. Humans demonstrated a richer diversity in responses, reflecting the nuanced interplay of cultural, experiential, and cognitive factors in shaping their associative networks. Additionally, their responses exhibited some universal patterns as with other languages, suggesting shared cognitive and linguistic mechanisms that transcend

individual differences. Conversely, LLMs produced responses that were less diverse, relying more on statistical patterns from training data such as word similarity, but less so on relatedness. Moreover, human responses exhibited greater stability in morphological properties across responses and showed weak correlations to the cue's morphological characteristics. In contrast, some LLMs displayed heightened sensitivity to these morphological patterns, more pronounced than in humans, but lacked consistency across different cues. This variability suggests that while LLMs can detect morphological patterns to some extent, they exhibit differing patterns of stability compared to humans. Furthermore, human responses demonstrated a nuanced interaction with linguistic features like concreteness which LLMs were unable to simulate successfully, reflecting distinct patterns of association influenced by abstractness levels. LLMs showed a more noticeable drop in concreteness across response orders and greater correlations between cue and response concreteness. This highlights LLMs' current limitations in capturing the adaptive and context-sensitive nature of human associative behavior. The graph-based analyses further underscored these differences, showing that while LLMs could approximate certain structural features of associative networks, they failed to capture the dynamic and interconnected nature of human semantic memory.

This study contributes to the literature by providing a systematic comparison of human and LLM word association patterns in Turkish, a morphologically rich language. By integrating methodologies from psycholinguistics, cognitive science, and computational linguistics, this work not only advances our understanding of LLM capabilities but also underscores the enduring complexity of human cognition. The findings have implications for both theoretical research and the development of more human-like AI systems, particularly in linguistically diverse contexts.

Nevertheless, some limitations should be acknowledged. The study focused on a limited set of LLMs and cue words, which may not fully represent the variability across models and linguistic stimuli. Additionally, we did not configure or fine-tune these models, which may have constrained their performance. Future research could address these limitations by optimizing model configurations, developing more diverse and population-simulating prompts, and fine-tuning LLMs with word association data to better capture human-like patterns.

In summary, this thesis highlights the challenges and opportunities in leveraging LLMs to simulate human word associations. While these models exhibit remarkable capabilities, they remain far from achieving the depth and flexibility of human cognition, particularly in complex linguistic tasks. This underscores the need for continued interdisciplinary efforts to bridge the gap between artificial and human intelligence.

APPENDIX A

INFORMED CONSENT FORM

KATILIMCI BİLGİ ve ONAM FORMU

Araştırmayı destekleyen kurum: Boğaziçi Üniversitesi

Araştırmanın adı: Kelime Çağrışımları: Yapay Zeka ve İnsan Çağrışım Normlarının Karşılaştırılması

Proje Yürütücüsü: Dr. Öğretim Üyesi Ümit Atlamaz

E-mail adresi:

Telefonu: -

Araştırmacının adı: Ceren Oksal

E-mail adresi:

Telefonu: -

Proje konusu: Kelimeler zihnimizdekileri başka insanlara iletmemizi sağlar. Bu bağlamda, kelimeler zihnimizdeki kavramların temsilidir. Bu çalışmanın amacı bu temsilin özelliklerini ve böylece kelimelerin zihnimizdeki yapısını anlamaktır.

Onam: Çalışmaya katılmayı kabul ettiğiniz takdirde çalışmaya başlamadan önce kendiniz hakkında bilgiler vermeniz istenecektir. Bu bilgiler yaşınız, cinsiyetiniz, mesleğiniz, eğitim durumunuz ve yaşadığınız şehirdir. Bu bilgiler sadece araştırmanın amaçları dahilinde kullanılacak olup bilgileriniz gizli kalacaktır ve araştırmanın sonunda silinecektir.

Çalışmaya ders kredisi karşılığında giriyorsanız çalışma sonunda kredi almak istediğiniz dersin adı ve öğrenci numaranız sorulacaktır. Çalışmaya katılmak size

bunun dışında maddi bir fayda sağlamayacaktır. Çalışmadan istediğiniz anda çekilme hakkına sahipsiniz. Sizden gelecek verileri ancak deneyi tamamladığınız takdirde kullanma şansımız olacaktır. Bu sebeple kredi alacak öğrenciler ancak çalışmayı tamamladıkları takdirde kredi alabileceklerdir. Çekildiğiniz ana kadar kaydedilen veriler hiçbir şekilde çalışmada kullanılmayacaktır.

Yapmak istediğimiz araştırma, çalışma hakkında fikir edinmenizi sağlayarak sizin için faydalı olabilir. Eğer araştırma ilginizi çekiyorsa, yukarıda belirtilen iletişim bilgilerini kullanarak araştırmacıyla iletişime geçerek araştırma hakkında daha fazla bilgi edinebilirsiniz. Bu araştırmanın size herhangi bir risk getirmesi beklenmemektedir.

Çalışma yaklaşık 30 dakika sürer. Çalışmadaki soruların hiçbirinin doğru ya da yanlış yanıtı yoktur; performansınız değerlendirilmeyecektir. Bu nedenle stres altında bulunmadan sorulara olabildiğince hızlı ve dürüst yanıt vermeniz beklenmektedir.

Bu formu imzalamadan önce veya çalışmadan sonra çalışmayla ilgili sorularınız olursa lütfen araştırmacı Ceren Oksal'a ceren.oksal@boun.edu.tr adresinden ulaşınız. Araştırmayla ilgili haklarınız konusunda Boğaziçi Üniversitesi Sosyal ve Beşeri Bilimler Yüksek Lisans ve Doktora Tezleri Etik İnceleme Komisyonu'na (SOBETİK) (sbe-ethics@boun.edu.tr) danışabilirsiniz.

Bana anlatılanları ve yukarıda yazılanları anladım. Bu formun bir örneğini aldım / almak istemiyorum (bu durumda araştırmacı bu kopyayı saklar).

Çalışmaya katılmayı kabul ediyorum.

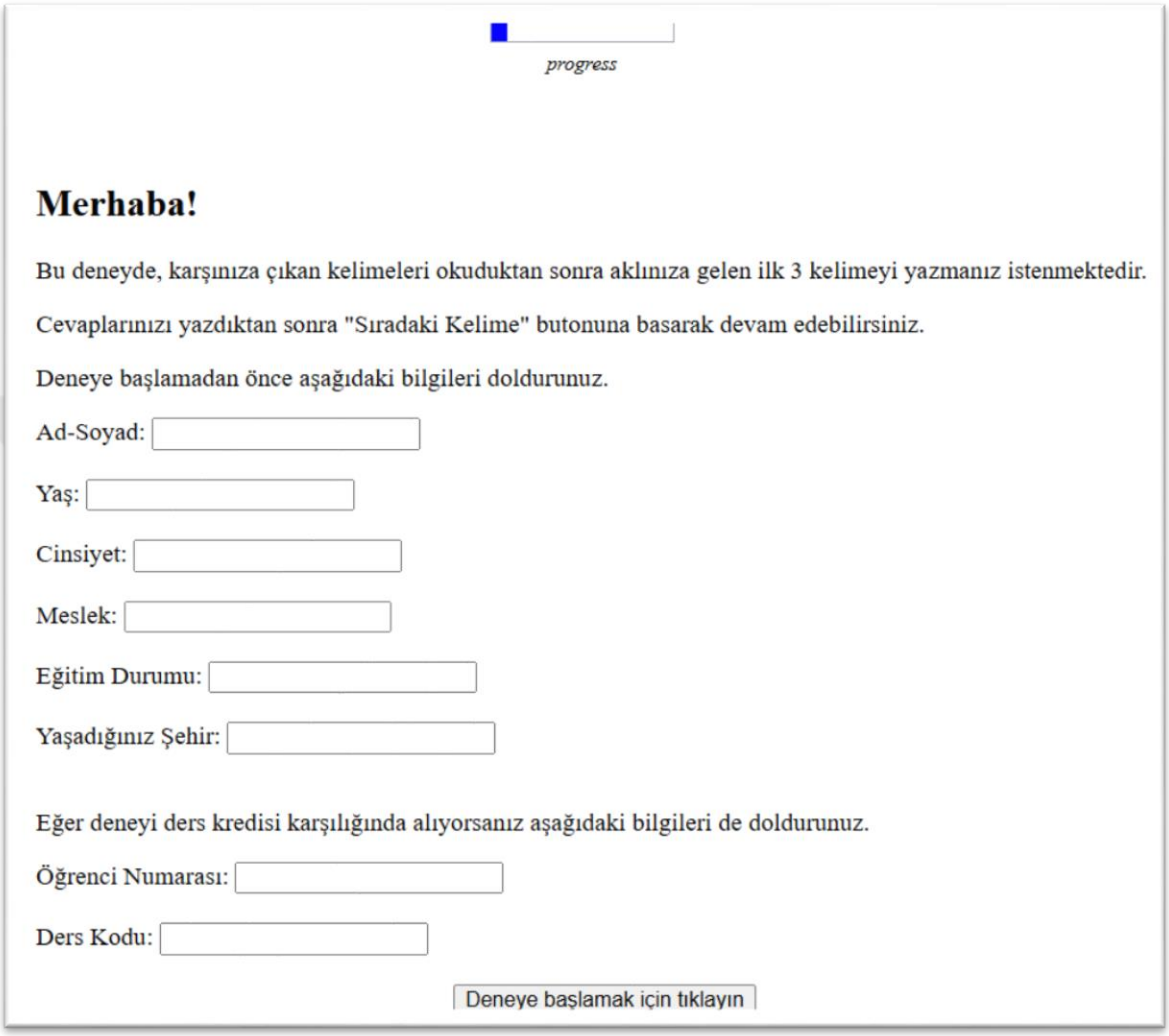
Katılımcı Adı-Soyadı:……..

İmzası:……

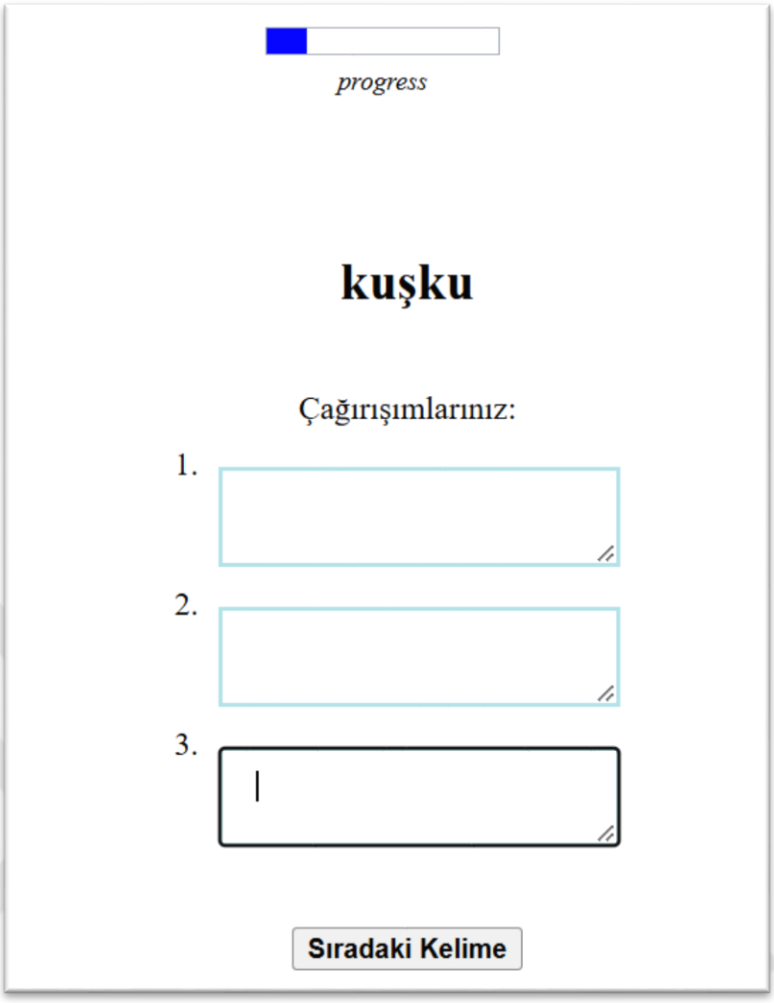
Tarih (gün/ay/yıl):...../...../.....

APPENDIX B

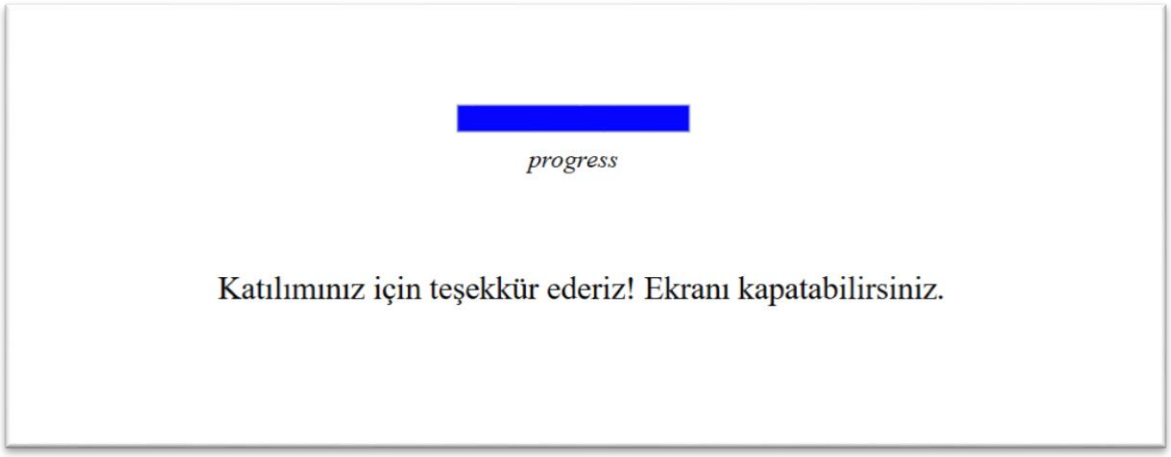
EXPERIMENT INTERFACE



After providing their consent through the online consent form, participants encountered a page explaining the experiment and requesting some information from them.



At each step, they were presented with a page containing a word and three response fields.



At the end of the experiment, a thank-you page was displayed for their participation.

APPENDIX C

EXPERIMENT MATERIALS

g1	g2	g3	g4	g5	g6
çare	savaştırılırken	pimpirikli	deneyimleme	memur	uzun
kahin	borçlanmak	cihaz	araba	saydam	kırmızı
muz	otobüs	dostçasına	sıcak	yardım	kuşku
tıknaz	opak	keşiş	savunmaktaydılar	gömlek	masum
ipek	kum	suskunluğuna	sahil	kelebek	oruç
tüketim	barıştırılırken	tımarhane	muz	meyve	gülümseme
hayat	serin	kirlice	terazi	anormalleşiyor	deli
şüphe	öğrenci	palto	bıçak	barınak	taksi
biçimlendirilmişlerdir	suçlu	sanatçı	orman	saldırmaktaydılar	fırın
teizm	kemalci	sabırsızlık	opak	kordon	sabahlardaki
ateistlik	oğlan	enerjik	hayat	kuş	adalet
barınak	delilim	horoz	pratikleştirsin	kemik	süresiz
kapüşon	ormanlık	öldürülüşündeki	devlet	şiddetli	pıhtı
çerçeve	viski	borçlanmak	atletik	saray	radyo
primatçaları	nezle	mücevher	süt	iktidarcılar	bağışlayışı
lokanta	şair	geçirgen	evrensel	anne	iktidarcılar
şiddetli	hamile	laikçiler	alacaklanmak	sos	mezar
heykel	sicim	cevher	ikindi	kitaphane	eksik
ispatım	beyaz	turizm	asosyalleştirdikleri	cenin	geceleyin
titiz	tıp	ses	saldırmaktaydılar	ocak	kral

g7	g8	g9	g10	g11	g12
yağcılık	otomobil	noksan	anarşist	kirlice	kazanç
çamaşır	savaştırılırken	kuyumcu	polis	bilimci	pamuk
yoğurt	enginar	sekülerizmciler	öncülük	külot	öğle
manavınkiler	transparan	kötülüyorsun	sefer	başıboş	avare
normalleşiyor	okul	çamaşır	erkekçene	damar	gürültü
köle	brüt	yiyecek	sabırsızlık	yağmur	uygulamak
başlamadılarsa	tabure	horultu	yedigen	varlık	hayvan
fotoğraf	gürültü	okul	yeni	atatürkist	uzak
suskunluğuna	saydam	yorgunluk	kitap	enerjik	dua
yüzük	manken	gazete	sarsıcı	süreli	seyahat
maymunsuları	kıtlıklarda	modernite	anarşist	nefes	alışkın
külot	su	mühendis	gözlük	gülümseme	zeytin
petrol	dükkan	orman	pamuk	ahşap	nezle
devlet	radyo	mezarhane	ocak	transparan	horoz
papatya	kitaplıklar	inançlılık	yorgun	ikindi	deneyimleme
bilişim	çenebazlığına	palto	müzik	özgür	ısınmış
bilimci	gevşek	sabır	modernite	kadeh	boza
mızrap	cihaz	medya	çare	çiçek	böcek
oksijen	sosyalleştirdikleri	hastane	kan	yağcılık	tecrübelenme
tepe	para	ses	sebze	kemalci	ebleh

REFERENCES

Abott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*, 558–559.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pas¸ca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 19–27).

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023, July). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning* (pp. 337-371). PMLR.

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Armstrong, B. C., Zugarramurdi, C., Cabana, Á., Valle Lisboa, J., & Plaut, D. C. (2016). Relative meaning frequencies for 578 homonyms in two Spanish dialects: A cross-linguistic extension of the English eDom norms. *Behavior Research Methods*, *48*, 950-962.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445-459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238-247).

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263.

Boring, E. (1950). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.

Boroditsky, L., & Schmidt, L. A. (2000). Sex, syntax, and semantics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *22*(22), 61–79 https://escholarship.org/uc/item/0jt9w8zf

Brolin59. (n.d.). *Trnlp: Turkish natural language processing library*. GitHub. Retrieved November 23, 2024, from https://github.com/brolin59/trnlp

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in Technicolor. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: *Long Papers- Volume, 1*, 136–145.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language & Cognitive Processes*, *12*, 177-210.

Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., & De Deyne, S. (2024). The" small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, *56*(2), 968-985.

Chatzikontantinou, A., Giannakidou, A., and Manouilidou, C. (2015). "Gradient Strength of NPI-Licensers in Greek." in Paper Presented at the *12th International Conference on Greek Linguistics* (Potsdam: University of Potsdam).

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*, 198–205.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*, 480-498.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987-1006.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp.1861–1870).

Deese, J. (1965). The structure of associations in language and thought. Baltimore: Johns Hopkins Press.

deGroot, A. M. (1995). Determinants of bilingual lexicosemantic organisation. *Computer Assisted Language Learning*, *8*(2–3), 151–180.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems, 36*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171-4186. https://doi.org/10.48850/arXiv.1810.04805

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*, *53*(8), 1560–1570.

Elias Costa, M., Bonomo, F., & Sigman, M. (2009). Scale-invariant transition probabilities in free word association trajectories. *Frontiers in Integrative Neuroscience*, *3*, 716.

Ercan, G., & Yıldız, O. T. (2018, August). Anlamver: Semantic model evaluation dataset for turkish-word similarity and relatedness. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3819-3836).

Fernández, A., Díez, E., & Alonso, M. A. (2012). Normas de Asociación libre en castellano de la Universidad de Salamanca [online database]. http://www.usal.es/gimc/nalc

Firth, J. R. A Synopsis of Linguistic Theory, 1930–1955. In Firth, J. R. (ed.), *Studies in Linguistic Analysis*, pp. 1–31. Blackwell, Oxford, United Kingdom, 1957.

Flor, M. M. (2024, May). Three Studies on Predicting Word Concreteness with Embedding Vectors. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024* (pp. 140-150).

Giannakidou, A. (1997). The landscape of polarity items. Ph.D. Groningen: University of Groningen.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining. (pp. 1406–1414).

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput. 9*, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, Albert Q., et al. "Mistral 7B." *arXiv preprint arXiv:2310.06825 (2023)*.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv 1602.02410 [Preprint]. https://arxiv.org/pdf/1602.02410.pdf

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., ... & Lenci, A. (2023). Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (*Vol. 1*, p. 2).

Kesgin, H. T., Yuce, M. K., Dogan, E., Uzun, M. E., Uz, A., İnce, E., ... & Amasyali, M. F. (2024, October). Optimizing Large Language Models for Turkish: New Methodologies in Corpus Selection and Training. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). The computer and literacy studies. In Aitken, A. J., Bailey, R. W., & Hamilton-Smith, N. (Eds.), (pp. 153–165): Edinburgh University Press.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2022). A critical review of network□ based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *14*(1), 54-77.

Landauer, T. K., & 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.

Liao, J., Chen, X., & Du, L. (2023). Concept understanding in large language models: An empirical study.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint arXiv:1907.11692*.

Ljubešić, N., Fišer, D., & Peti-Stantić, A. (2018). Predicting concreteness and imageability of words within and across languages via word embeddings. *arXiv preprint arXiv:1807.02903*.

Lund, C. B. A. K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and cognitive processes*, *12*(2-3), 177-210.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

McCoy, R. T., Min, J., & Linzen, T. (2019). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781v3

Miller, G. A., & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language & Cognitive Processes*, *6*, 1–28.

Moldovan, C. D., Ferré, P., Demestre, J., & Sánchez-Casas, R. (2015). Semantic similarity: Normative ratings for 185 Spanish noun triplets. *Behavior Research Methods*, *47*(3), 788–799.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*, 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*,402-407.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press, Inc.

OpenAI. (2024, July 18). *GPT-4o mini* [Large language model]. OpenAI. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Parker, D., and Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition 157*, 321–339. doi: 10.1016/j.cognition.2016.08.016

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*, 407–417.

Prior, A., & Bentin, S. (2008). Word associations are formed incidentally during sentential semantic integration. *Acta Psychologica*, *127*(1), 57-71.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, (pp. 337–346).

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627–633. https://doi.org/10.1145/365628.365657

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506-515.

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings* (pp. 417-427). Springer Berlin Heidelberg.

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta psychologica*, *119*(2), 159-187.

Shin, U., Yi, E., & Song, S. (2023). Investigating a neural language model's replicability of psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology*, *14*, 937656.

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

Steyvers, M., & Tenenbaum, J. B. (2005). Graph theoretic analyses of semantic networks: Small worlds in semantic networks. *Cognitive Science*, *29*(1).

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and Its Applications* (pp. 237–249). American Psychological Association. https://doi.org/10.1037/10895-018

Szalay, L. B., & Deese, J. (1978). Subjective meaning and culture: an assessment through word associations. Hillsdale, NJ: Lawrence Erlbaum.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., ... & Ronstrom, S. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Tekcan, A. İ., & Göz, İ. (2005). *Türkçe kelime normları: 600 kelimenin imgelem, somutluk, sıklık değerleri ve çağrışım setleri* (No. 849). Boğaziçi Üniversitesi.

Thawani, A., Srivastava, B., & Singh, A. (2019, June). Swow-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 43-51).

TheMosaicML NLP Team. (2023). MPT-30B: Raising the bar for open-source foundation models. Retrieved from https://www.mosaicml.com/blog/mpt-30b

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprintarXiv:2302.13971*.

Trendyol. (2024). *Trendyol-LLM-7b-chat-v1.8* [Large Language Model]. Hugging Face. Retrieved from https://huggingface.co/Trendyol/Trendyol-LLM-7b-chat-v1.8

U.S. Census Bureau. (2010). *2010 Census data*. Retrieved from https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, *41*(1), 102-136.

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model. Retrieved from https://github.com/kingoflolz/mesh-transformer-jax

Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H. W., Beltagy, I., ... & Raffel, C. (2022, June). What language model architecture and pretraining objective works best for zero-shot generalization?. In *International Conference on Machine Learning* (pp. 22964-22984). PMLR.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang*. *108*, 40–55. doi:10.1016/j.bandl.2008.10.002

Xiang, M., Grove, J., and Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. Front. Psychol. 4:708. doi:10.3389/fpsyg.2013.00708

Yıldız, O. T., Avar, B., & Ercan, G. (2019, September). An open, extendible, and fast Turkish morphological analyzer. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 1364-1372).

Zehr, J., & Schwarz, F. (2023, January 7). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832

Zwarts, F. (1996). "A hierarchy of negative expressions" in Negation: A Notion in Focus. ed. H. Wansing (Berlin: De Gruyter), 169–194.