Language and Cognition (2026), 18, e67, 1–24
doi:10.1017/langcog.2026.10119

ARTICLE

Lexical organization in humans and large language models: Evidence from L1, L2 and LLM networks

Marta Sánchez-Saus Laserna

Departamento de Lengua Española, Lingüística y Teoría de la Literatura, Universidad de Sevilla, Spain
Email: msanchezsaus@us.es

(Received 30 March 2026 UTC; Revised 07 August 2026 UTC; Accepted 07 September 2026 UTC)

Abstract

This study compares paradigmatic lexical networks in native speakers of Spanish (L1), advanced learners of Spanish as a foreign language (L2) and a large language model (ChatGPT) in order to examine how the organization of model-generated networks compares to that observed in human speakers. All groups produced vocabulary appropriate to the target category and displayed coherent local clustering. However, the two human groups showed highly similar global network structures despite differences in acquisition history, whereas the model exhibited higher modularity, longer path lengths and substantially higher lexical convergence. L2 learners, despite more limited exposure to Spanish, converged with the L1 group on all key structural measures, while the model diverged from both. In addition, the large language model produced less concrete vocabulary than either human groups, and the relationship between lexical availability and familiarity observed in human speakers was not present in the model. These results indicate that similarities in lexical content do not entail comparable organization and suggest that human lexical systems converge on shared structural properties under different learning conditions. The findings also highlight the value of including speakers at different stages of acquisition in comparisons between human and artificial language systems.

Keywords: large language models; lexical availability; mental lexicon; second language acquisition; semantic networks

1. Introduction

Comparisons between large language models and human speakers have become a central strategy for evaluating the scope and limits of artificial language systems. Early discussions focused primarily on text production, asking whether models could produce fluent and contextually appropriate text. Attention has increasingly shifted,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (http://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

however, from what systems produce to how their linguistic material is organized. The question is no longer limited to output quality, but extends to the structural configuration of lexical relations across systems.

When humans generate words within a semantic category, their responses reflect patterns of association, clustering and transition that reveal how lexical items are interconnected. These patterns can be modeled as networks (Collins & Loftus, 1975), making it possible to characterize lexical organization in terms of local clustering, global integration and the degree to which responses form distinct subgroups (De Deyne & Storms, 2008; Steyvers & Tenenbaum, 2005). Semantic fluency tasks, including lexical availability tasks, provide a well-established method for eliciting such data and have been widely used to examine lexical organization in native and non-native speakers, as well as sociolinguistic variation (Bartol Hernández, 2006; Hernández Muñoz, 2006; Sánchez-Saus Laserna, 2022).

The same methodology can now be applied to artificial systems. If exposure to large-scale textual data was sufficient to produce human-like patterns of lexical organization, language models trained on massive corpora should exhibit network structures comparable to those observed in human speakers. If not, structural differences may arise even when lexical content appears similar.

Most existing empirical comparisons have relied on English data. The present study extends this line of research to Spanish, drawing on the long-standing tradition of lexical availability research in this language. We examine paradigmatic association networks elicited by the category *la casa* (the house), a concrete and familiar domain organized around the spatial structure and lived experience of a familiar environment. Three groups are compared: native speakers of Spanish (L1), advanced learners of Spanish as a foreign language at C1–C2 (CEFR) level (L2) and ChatGPT (GPT-5.2).

Native speakers develop lexical knowledge through long-term interaction with language and experience (Barsalou, 2008; Vigliocco et al., 2009), whereas advanced L2 learners develop their lexicons later and under more restricted conditions of exposure, progressively developing direct lexical–conceptual connections as proficiency increases (Kroll et al., 2010; Kroll & Stewart, 1994). The language model, by contrast, acquires lexical knowledge through self-supervised learning over massive text corpora, without the developmental and embodied experience that shapes human lexical systems (Bender et al., 2021; Brown et al., 2020).

The comparison therefore does not simply contrast human speakers with an artificial system, but systems shaped by different conditions of acquisition and use, making it possible to examine whether similar organizational patterns emerge despite that difference or whether differences in how lexical knowledge is acquired are reflected in the structure of the resulting networks.

Advanced L2 learners provide a particularly informative comparison. Although they have received substantially less cumulative exposure to Spanish than native speakers, highly proficient learners develop direct links between lexical items and conceptual representations, resulting in integrated lexical systems. If lexical organization depended primarily on exposure alone, their networks should diverge from native patterns; convergence, instead, would indicate that human lexical organization reflects constraints that are robust across acquisition histories. In this sense, L2 learners are not simply an intermediate group, but a key point of comparison for interpreting differences between human and model-generated systems.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

The central research question is therefore whether lexical networks derived from a large language model display organizational properties comparable to those observed in human speakers, and whether similar structural patterns emerge across systems shaped by different conditions of acquisition and use. The analysis considers not only lexical content but also patterns of integration, connectivity, variability and lexical profile.

This study is guided by four research questions:

RQ1 (Network organization): Do LLM-derived lexical networks differ from human-derived networks in global topology, particularly in the balance between internal cohesion and global integration?

RQ2 (Human convergence): Do native speakers and advanced L2 learners converge on comparable network structures despite differences in acquisition history and cumulative exposure?

RQ3 (Lexical diversity): Do LLM-generated responses differ from human responses in lexical diversity and lexical convergence?

RQ4 (Word type): Do LLM-generated and human-generated vocabulary differ in psycholinguistic properties such as concreteness and familiarity?

2. Theoretical frame

2.1. Large language models and lexical organization

The rapid development of large language models has renewed attention to a central question: whether linguistic systems trained primarily on large-scale textual data develop organizational properties comparable to those observed in human lexical systems or whether human lexical organization depends on additional cognitive and experiential constraints.

Early work in distributional semantics proposed that aspects of meaning can be inferred from patterns of co-occurrence in linguistic input (Firth, 1957; Harris, 1954). Contemporary computational models implement this idea by representing words in high-dimensional vector spaces derived from large corpora, capturing semantic similarity through statistical regularities (Günther et al., 2019; Lenci, 2018). Large language models extend this line of work by learning from extremely large multilingual corpora. Empirical studies report partial parallels with human data, including alignment with semantic similarity judgments and lexical decision behavior, although model-generated conceptual structures tend to be less coherent and less stable than those observed in humans (Digutsch & Kosinski, 2023; Suresh et al., 2023; Xiao et al., 2025).

From a cognitive perspective, however, these parallels do not imply equivalence at the level of network organization. Human lexical networks are shaped not only by linguistic experience but also by sensorimotor experience, communicative use and cognitive constraints on memory and retrieval (Barsalou, 2008; Hills et al., 2012; Vigliocco et al., 2009). Words are encountered through repeated interaction with physical objects and social routines, and these embodied traces influence both how they are organized and how readily they are accessed. Large language models, by contrast, lack developmental history and perceptual grounding. If such factors contribute to the characteristic configuration of human lexical networks, systems

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

trained exclusively on textual data may not reproduce the same organizational profile even when lexical content appears appropriate.

Empirical comparisons support this expectation, pointing to limitations such as reduced lexical diversity, lower variability across outputs and network structures characterized by higher modularity and poorer global integration (Abramski et al., 2023; Wang et al., 2025). Such comparisons remain largely concentrated on English-language data. Recent work applying semantic fluency paradigms in Spanish similarly points to systematic differences, with LLM outputs tending to show higher convergence and denser lexical cores (Sánchez-Saus Laserna et al., 2026), and the present study builds on this emerging line of research.

2.2. Paradigmatic organization and the mental lexicon

To understand how lexical networks are structured in human cognition, it is necessary to consider the role of paradigmatic relations in the mental lexicon. Words that can substitute for one another in similar contexts, such as *dog* and *cat* or *table* and *chair*, reflect shared semantic features or category membership and reveal how lexical items are grouped within semantic memory.

Classic models of semantic memory describe the lexicon as a network of interconnected nodes in which activation spreads from one item to related items (Collins & Loftus, 1975). Later work has refined this view using tools from network science, conceptualizing lexical systems as graphs in which nodes represent words and edges represent associative or similarity-based relations (De Deyne & Storms, 2008; Steyvers & Tenenbaum, 2005). According to non-decompositional models of lexical production (Levelt, 1989; Levelt et al., 1999; Roelofs, 1992), activation of one concept propagates to semantically related nodes and the word selected is the one with the highest activation level at the moment of retrieval. This mechanism explains why semantically related words tend to cluster in production sequences.

Semantic networks derived from human data typically display small-world properties, combining high local clustering with short average path lengths (De Deyne & Storms, 2008; Steyvers & Tenenbaum, 2005). This configuration supports efficient retrieval: speakers exploit dense local neighborhoods while maintaining the ability to transition across different regions of the network. Semantic fluency tasks provide a behavioral basis for constructing such networks, since sequential production patterns allow associative links to be inferred and aggregate patterns reveal broader organizational properties (Hills et al., 2012; Kenett et al., 2017).

In the Spanish-speaking research tradition, lexical availability tasks have provided a systematic framework for examining categorical production, with availability indices reflecting not only vocabulary size but also accessibility and internal organization (Bartol Hernández, 2006; Hernández Muñoz et al., 2006). More recent work applies network-based analyses to availability data (Agustín Llach & Rubio, 2024; Tomé Cornejo et al., 2025), providing a methodological foundation for the present study.

The semantic category *la casa* ('the house') is particularly well suited to this framework. Unlike natural categories such as animals or body parts, organized primarily around taxonomic relations, *la casa* functions as a structured representation of a familiar environment, integrating spatial organization, typical objects and patterns of interaction. Categories of this type have been described in cognitive

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

research as schemata (Rumelhart, 1980), frames (Minsky, 1975) and scenes (Mandler, 1984), emphasizing their organization around experiential configurations rather than taxonomic similarity.

Networks derived from semantic fluency data, including those analyzed here, are constructed from the sequential order in which words are produced. This means the resulting structure reflects two things at once: the underlying organization of the lexicon and the retrieval process used to search it. Two models describe this retrieval process differently. Under a random walk model, more strongly connected words are simply more likely to be produced one after another, so clustering emerges without any active search strategy. Under the foraging model (Hills et al., 2012), speakers instead exploit a semantically related group of words for as long as new items come easily and switch once retrieval becomes harder. This is a pattern that has generally predicted human fluency data better than random walk (Avery & Jones, 2018), with switching further shaped by properties of the words already produced (Zhang & Jones, 2022). Under either view, the sequence still reflects lexical organization, but under a foraging model, it also reflects the speaker's own decision about when to move between groups (Zemla, 2022; Zemla & Austerweil, 2017). Because the present networks are built directly from production order, their structural measures cannot separate these two sources of difference: two groups could differ in network structure because of organization, retrieval strategy or both. We return to this distinction in Section 5.

2.3. L1 and L2 lexical organization

Research on bilingual lexical processing shows that the organization of the second language lexicon develops progressively and may differ from native organization, particularly at early stages of acquisition. Initially, lexical access in the second language often relies on links to the first language, with conceptual representations accessed through translation equivalents. As proficiency increases, direct connections between lexical items and conceptual representations strengthen and reliance on the first language decreases (Kroll et al., 2010; Kroll & Stewart, 1994).

With increasing experience, L2 lexical representations become more integrated at the conceptual level. In highly proficient speakers, activation from shared conceptual representations can propagate directly to lexical nodes, leading to patterns of organization that resemble those observed in native speakers. From a network perspective, this development corresponds to a shift from sparser, less-interconnected structures to more densely connected and integrated networks.

Empirical evidence supports this convergence. Word association studies show that lower-proficiency learners tend to produce more form-based or syntagmatic responses, whereas advanced learners display predominantly paradigmatic patterns similar to those of native speakers (Fitzpatrick & Izura, 2011). Network-based analyses further indicate that lexical networks in less proficient learners are less connected and less efficiently structured, while advanced learners develop networks that increasingly approximate native organization (Borodkin et al., 2016; Sánchez-Saus Laserna, 2022). Taken together, these findings indicate that lexical organization in a second language is dynamic but can approach native-like structural configurations at higher levels of proficiency.

This convergence is theoretically informative because it emerges under conditions that differ substantially from native acquisition. Advanced L2 speakers have received

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

less cumulative exposure and have acquired the language later, yet their lexical systems can display comparable structural organization. This makes it possible to examine whether the properties observed in human lexical networks depend on how each person specifically learned the language or whether they appear regardless of how learning happened.

Lexical availability research provides additional insight into the factors that shape retrieval. In native speakers, word accessibility is shaped by several psycholinguistic variables, including typicality, familiarity, concreteness and age of acquisition (Hernández Muñoz, 2006; Morrison & Ellis, 1995; Rosch & Mervis, 1975). Because the present study compares lexical organization across human speakers and a language model trained only on text, we focus on two variables that directly characterize the experiential properties of lexical items: familiarity and concreteness. Concreteness reflects the extent to which a concept is grounded in perceptual and sensorimotor experience (Paivio, 1986), whereas familiarity reflects speakers' subjective experience with lexical items through repeated exposure and use (Gernsbacher, 1984). Other psycholinguistic variables, including typicality and age of acquisition, are also known to influence lexical availability. They were not included here because the present analyses were specifically designed to examine the experiential dimension captured by familiarity and concreteness. Together, these two variables provide complementary perspectives on the extent to which the lexical items produced by each system are grounded in perceptual experience and accumulated linguistic exposure. These variables do not only affect which words are produced but also their relative accessibility during semantic retrieval.

In L2 learners, additional factors such as age of acquisition, learning context and cross-linguistic similarity can influence lexical productivity and network structure (Hernández Muñoz & Tomé Cornejo, 2017; Sánchez-Saus Laserna & Álvarez Torres, 2024). Even so, at advanced levels of proficiency, lexical organization increasingly reflects shared conceptual structure rather than mediation through the first language.

The inclusion of advanced learners of Spanish therefore provides a particularly informative point of comparison. They represent fully functional human lexical systems shaped by different acquisition histories and degrees of exposure. If similar structural patterns emerge in both L1 and L2 networks despite these differences, this indicates that key properties of lexical organization are robust across learning trajectories. In this context, convergence between L1 and L2 networks is informative not because the groups are identical, but because they arrive at comparable structural configurations under different conditions. Differences observed between human and model-generated networks under these conditions are therefore less likely to reflect variation in experience alone and more likely to reflect differences in how lexical knowledge is organized.

3. Method

3.1. Participants

The study included three groups: native speakers of Spanish (L1), advanced learners of Spanish as a foreign language (L2) and ChatGPT (GPT-5.2) as a computational comparison system.

The L1 group comprised 50 undergraduate students from the University of Cádiz, Spain (13 male, 37 female). Their mean age was 18.5 years (SD = 0.89; range = 18–22).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

All participants were native speakers of Spanish who had completed their primary and secondary education in Spain.

The L2 group consisted of 50 advanced learners of Spanish (C1–C2 level according to the Common European Framework of Reference for Languages). The group included 18 male and 32 female. Mean age was 21.8 years (SD = 1.5; range = 19–25). At the time of testing, they were residing in Spain and had lived there for at least three months. The sample was intentionally heterogeneous with respect to first language in order to capture general properties of advanced second language lexical organization rather than effects tied to a specific L1. Participants included native speakers of English (n = 15), German (n = 14), French (n = 12) and Italian (n = 9), representing both Germanic and Romance language backgrounds.

ChatGPT (GPT-5.2) was selected as a computational comparison system because it is a widely used contemporary language model trained on large-scale multilingual data. Although the precise composition of its training corpus is not fully disclosed (OpenAI, 2023), it has been exposed to a quantity of linguistic input far exceeding that of any individual speaker. This makes it a suitable comparison system for examining whether large-scale textual training leads to patterns of lexical organization comparable to those observed in human speakers.

3.2. Materials and procedure

All groups completed a semantic fluency task based on the category *la casa* (the house), selected because it is well represented in lexical availability research in Spanish and typically elicits vocabulary organized around spatial and experiential relations within a familiar environment.

Human participants wrote down individual words related to *la casa* during a two-minute interval, producing single words in the order in which they came to mind. For the LLM condition, the following prompt was used: 'Dame una lista de las primeras 40 palabras que se te vengan a la mente cuando piensas en "la casa"' ('Give me a list of the first 40 words that come to mind when you think of "the house"'). Fifty runs were conducted independently from separate sessions using different accounts to minimize carryover effects.

Repeated elicitation from a large language model differs fundamentally from sampling across human participants: each run comes from the same model under fixed conditions, not from a distinct individual with its own linguistic history. Language models generate text through a stochastic decoding process, sampling each word from a probability distribution over possible continuations rather than choosing it deterministically (Holtzman et al., 2020). Because the prompt was identical across all 50 runs, each generated in a separate session with no memory of previous runs, we treat the runs not as 50 participants but as independent samples from the model's response distribution under this fixed prompt (Wang et al., 2023). Our aim is accordingly to characterize that distribution, not to estimate variability between individual runs. This distinction matters here because LLM outputs are known to be markedly less variable than human responses, even across repeated sampling and different models (Qiu et al., 2025). Limited variability across runs under an identical prompt is therefore an expected property of this sampling process, not a design flaw: varying the prompt would have changed which distribution was being sampled, rather than adding observations from the same one.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

The model was prompted to generate 40 words per run, exceeding the average human output of approximately 20 words, in order to sample the response distribution with sufficient resolution to capture its less probable tail. This decision was informed by a pilot analysis examining output stability across positions in the list. The first 15–20 items showed high overlap across runs, indicating the higher-probability core of the distribution, whereas later items introduced greater variability, reflecting its long tail. Extending the output length therefore allowed a broader sampling of this distribution and reduced ceiling effects in lexical convergence (see Supplementary Materials).

One parameter of this sampling process could not be reported: the consumer-facing ChatGPT interface used for data collection does not expose or document temperature or top-*p*, and OpenAI does not disclose a default value for this interface. However, recent work directly manipulating temperature across eight settings while comparing human and LLM semantic search dynamics found that the human–LLM divergence persisted across the full range tested, with no configuration fully reproducing the human profile (Paris-Colombo et al., 2026), suggesting that this parameter is unlikely to fully account for the differences reported below. Future replications using API access, where such parameters can be set and reported explicitly, would allow more precise control over this aspect of the design.

3.3. Data processing

All responses were preprocessed prior to analysis using the same procedure for human and model-generated data. Repeated words within each individual response were removed, retaining only the first occurrence per participant or model run. The responses were then lemmatized: nouns and adjectives were converted to singular form, with gender-marked variants normalized to the masculine and verbs to the infinitive.

The processed data were analyzed using LexPro (Hernández Muñoz et al., 2025). The Lexical Availability Index (IDL) was calculated for each word following the López-Chávez and Strassburger (2000) formula, integrating frequency and position within individual lists and assigning higher values to words produced by many participants early in their responses. A cohesion index was also computed as the ratio between mean words per participant and total unique types (*M/T*), with higher values indicating greater convergence across lists.

3.4. Identification of central nodes and network construction

To establish a principled, non-arbitrary node-selection cutoff, words were additionally classified into centrality levels using DispoCen (v. 0.1.0) (Ávila Muñoz et al., 2021), which implements the fuzzy-set model of Ávila-Muñoz and Sánchez-Sáez (2014).

To ensure structural comparability, networks were constructed with the same number of nodes across all conditions. This number was determined once, from the L1 dataset, by classifying words into centrality levels (0–6) via the Ávila-Muñoz and Sánchez-Sáez (2014) model and retaining levels 6–3 (core and first periphery), which gave 26 items. L1 was used as the reference for this cutoff because it serves as the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

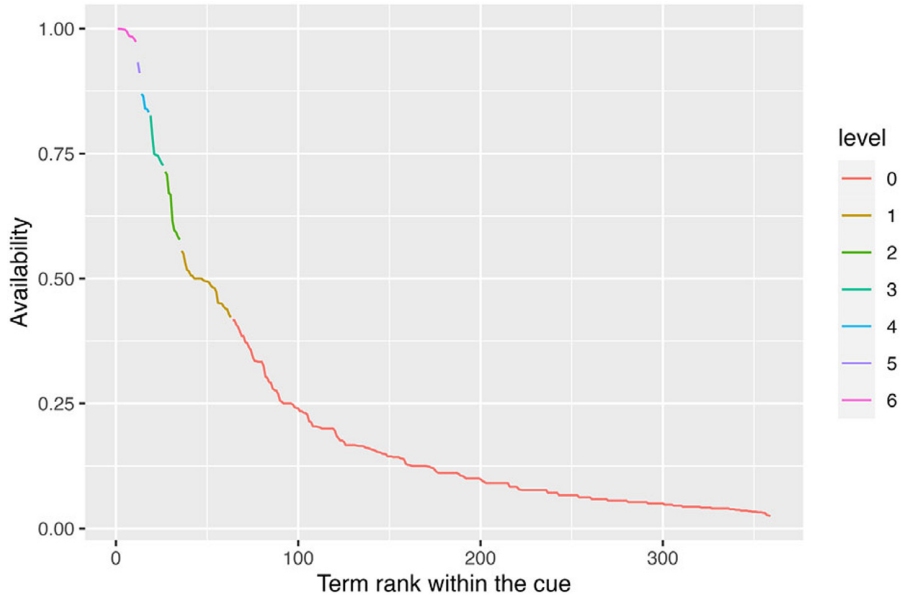


Figure 1. Availability curve for the category *la casa* (L1 group). Words are ordered by decreasing Lexical Availability Index (IDL). Colors indicate centrality levels assigned by DispoCen (levels 6 to 0).

empirical standard against which the L2 and LLM conditions are compared. This matches the comparative aim of the study.

This fixed node count ($N = 26$) was then applied independently to each group's own dataset: for each condition, the top-N items by López-Chávez and Strassburger (2000) availability were selected as network nodes. DispoCen provided a non-arbitrary way to decide how many nodes to include, while LexPro's continuous ranking (the standard formula in lexical availability research) determined which specific words filled those slots (Figure 1). As a result, the specific lexical items forming each network's node set differ across groups, since each condition has its own distribution of available vocabulary even though the number of nodes is fixed.

Undirected networks were constructed based on sequential adjacency in the fluency responses. Consecutive words within each list were identified as bigrams and treated as undirected connections, regardless of production order. This approach captures co-occurrence relations while focusing on overall connectivity patterns rather than directional transitions. Edge weights reflected the total frequency with which two words appeared consecutively across participants or model outputs, aggregated to produce one group-level network per condition. Node size in the graphical representations was proportional to IDL values, scaled using a common range across groups and edge thickness reflected adjacency frequency.

3.5. Network metrics

To characterize the topological properties of the networks, several standard graph-theoretic measures were computed on the undirected graphs. Average degree reflected the mean number of connections per node. Average shortest path length

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

(ASPL) captured the mean distance between nodes and provided an estimate of global integration. The clustering coefficient quantified the extent to which neighboring nodes were interconnected. Modularity assessed the degree to which the network could be partitioned into internally dense communities with relatively fewer intercommunity connections, calculated using the Louvain algorithm (Blondel et al., 2008). Betweenness centrality was also calculated for each node, estimating how frequently a node lies on shortest paths between other nodes and thus indicating its role in connecting different regions of the network. Together, these measures provide complementary descriptions of the local and global organizations of semantic networks and are widely used for this purpose in studies based on semantic fluency and word association data (De Deyne & Storms, 2008; Kenett et al., 2017; Steyvers & Tenenbaum, 2005). These measures were selected because they capture properties directly linked to the psychological account of lexical retrieval outlined in Section 2.2: ASPL and clustering index the small-world configuration associated with efficient retrieval, while modularity and betweenness centrality reflect the balance between locally cohesive subgroups and the cross-cluster connections that a random walk or foraging process must exploit to move between them. All metrics were computed using identical procedures across groups.

In addition to network metrics, lexical diversity was characterized using total unique word types, mean words per participant and the cohesion index (*M/T*), where higher values indicate greater convergence across responses.

3.6. Statistical analysis

Because analyses were conducted on group-level network structures rather than item-level responses, mixed-effects modeling was not applicable. Instead, variability was assessed using resampling procedures at the participant level.

To assess the stability of network metrics and estimate uncertainty at the group level, a bootstrap procedure was implemented separately for each condition. A case bootstrap approach was used, resampling whole participant response lists rather than individual bigrams, since words within one list are not independent of each other, whereby participants were resampled with replacement within each group while maintaining the original sample size (n = 50). For each of 1,000 iterations, a new group-level dataset was created by resampling participants, the network was reconstructed and all metrics were recalculated. Empirical confidence intervals were derived from the 2.5th and 97.5th percentiles of the bootstrap distribution.

Pairwise differences between groups were evaluated using bootstrap difference distributions. For each iteration, the metric obtained for one group was subtracted from that of the comparison group, yielding an empirical distribution of differences (Δ). Differences were considered statistically meaningful when the 95% confidence interval did not include zero, and two-tailed p-values were estimated from the proportion of bootstrap samples in which Δ crossed zero.

Differences in lexical diversity were evaluated through total unique word types, mean words per participant and the cohesion index (*M/T*). Because cohesion is a group-level aggregate statistic, differences between groups were evaluated using permutation tests (5,000 permutations), which preserve the internal structure of participant lists while avoiding distortions introduced by resampling with replacement. In each permutation, participant labels were randomly reassigned across

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

groups while preserving individual response lists, and the cohesion index was recalculated for each permuted dataset.

Concreteness and familiarity ratings were obtained using a two-step procedure. The primary source was Guasch et al. (2016), which provides human ratings on a 1–7 scale for 1,400 Spanish words. For words not included in this database, ratings were obtained from Martínez et al. (2025), who provide estimates generated using GPT-4o and validated against human norms ($r \approx 0.80$). Of the 679 unique lexical items in the dataset, 156 (23%) received ratings from Guasch et al. and 493 (73%) from Martínez et al. A total of 30 items (4.4%) lacked concreteness ratings and 25 (3.7%) lacked familiarity ratings; these were predominantly multi-word expressions (e.g., *cama de matrimonio* [double bed], *cuarto de invitados* [guest room]) and proper nouns (e.g., *La casa de papel* [Money Heist], *Los Simpson* [The Simpsons]) not covered by either source. Missing data were distributed across groups and are unlikely to bias comparisons.

Analyses were conducted at the word level, with group differences assessed over distributions of lexical items within each condition. Group differences were evaluated using Kruskal–Wallis tests. When significant, Dunn's *post hoc* comparisons with Bonferroni correction were conducted and effect sizes were estimated using ε^2. Within each group, Spearman correlations were computed between IDL values and both psycholinguistic variables.

All analyses were conducted in R version 4.3.2 (R Core Team, 2023). Network metrics were computed using the igraph package version 1.5.1 (Csardi & Nepusz, 2006). *Post hoc* comparisons were performed using the dunn.test package version 1.3.5 (Dinno, 2017). Bootstrap and permutation procedures were implemented using base R functions.

3.7. Robustness checks

The robustness of the reported network topology was evaluated in three ways: whether the result depends on the statistical estimation procedure, on how edge weight is represented or on the specific method used to infer the network itself.

All metrics were also recomputed once on the full, non-resampled dataset to verify that the case-bootstrap procedure described above (Section 3.6) does not itself drive the reported differences (Supplementary Table SM1). Results pointed in the same direction, though bootstrap-derived ASPL was consistently higher across all three groups (a known side-effect of case resampling, which can occasionally fragment small networks, unrelated to the group comparisons of interest).

Because our aim is to characterize which lexical items are reliably linked and how those links are globally organized, rather than how strong each connection is, degree, clustering and path length were computed on the unweighted network; modularity retained edge weights, as the Louvain algorithm incorporates them by default. Fully weighted versions of all four measures were computed as a sensitivity analysis (Supplementary Table SM1), using the weighted clustering coefficient of Barrat et al. (2004); edge weights were normalized within each group before inversion to path length, given that raw weights differed substantially in scale across groups.

We also evaluated statistical edge significance following Goñi et al. (2011), but this left too few significant edges for the L1 and L2 networks to stay connected, fragmenting them into several disconnected pieces and preventing a meaningful comparison of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

global measures (see Supplementary Materials for the full power analysis). We therefore re-estimated each group's network using U-INVITE (Zemla & Austerweil, 2018), a generative model developed for semantic fluency data and implemented in the snafupy package (Zemla et al., 2020), applied to each group's complete vocabulary to preserve genuine sequential adjacency (Supplementary Table SM2).

Across the original analysis and the three alternative estimations, the three groups kept the same relative order on modularity and average path length. This consistency indicates that the reported pattern reflects a genuine property of the data, present across different ways of building and estimating the network.

All data and materials are available at https://osf.io/emf6t/.

4. Results

4.1. Network topology

Descriptive statistics for the network metrics are presented in Table 1, and the results of the bootstrap pairwise comparisons are reported in Table 2. All networks contained 26 nodes (see Section 3.4).

As shown in Table 1, modularity was highest in the LLM network ($M = 0.533$, $SD = 0.013$), compared to L1 ($M = 0.320$, $SD = 0.030$) and L2 ($M = 0.349$, $SD = 0.024$). The bootstrap analyses reported in Table 2 confirm that these differences are statistically reliable. The LLM exhibited significantly higher modularity than both L1 ($\Delta = 0.21$, 95% CI [0.15, 0.27], $p < 0.001$) and L2 ($\Delta = 0.18$, 95% CI [0.13, 0.24], $p < 0.001$), whereas no reliable difference emerged between the two human groups ($\Delta = -0.03$, 95% CI [−0.10, 0.05], $p = 0.43$), addressing RQ2.

A similar pattern appears for ASPL. The LLM shows longer paths ($M = 4.674$, $SD = 0.665$) than L1 ($M = 2.871$, $SD = 0.377$) and L2 ($M = 2.619$, $SD = 0.250$), with statistically reliable differences for both comparisons (LLM–L1: $\Delta = 1.80$, 95% CI [0.51, 3.43], $p = 0.006$; LLM–L2: $\Delta = 2.05$, 95% CI [0.79, 3.56], $p < 0.001$), while L1 and L2 did not differ ($p = 0.59$), consistent with the pattern already observed for modularity.

Average degree presents a different configuration. The L2 network had the highest connectivity ($M = 12.201$, $SD = 0.715$), exceeding both the LLM ($M = 9.581$, $SD = 0.444$) and L1 ($M = 8.254$, $SD = 0.740$). Pairwise comparisons indicate that L2 differed from both LLM ($\Delta = 2.62$, $p = 0.002$) and L1 ($\Delta = 3.95$, $p < 0.001$), whereas LLM and L1 do not differ ($p = 0.12$). This pattern indicates a denser configuration of local connectivity in the L2 network.

Table 1. Descriptive network metrics by group

Metric	L1 (*M*, *SD*)	L2 (*M*, *SD*)	LLM (*M*, *SD*)
Modularity	0.320 (0.030)	0.349 (0.024)	0.533 (0.013)
Average shortest path length	2.871 (0.377)	2.619 (0.250)	4.674 (0.665)
Average degree	8.254 (0.740)	12.201 (0.715)	9.581 (0.444)
Clustering coefficient	0.382 (0.033)	0.469 (0.022)	0.465 (0.025)

Note: Values represent group-level metrics computed on undirected networks (26 nodes per network; $n = 50$ per group). Modularity was computed using edge weights; average degree, average shortest path length, and clustering coefficient were computed on the unweighted (binary) network (see Section 3.7 for weighted versions reported as a robustness check). Standard deviations are derived from 1,000 bootstrap resamples.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Table 2. Bootstrap pairwise differences in network metrics

Metric	Comparison	Δ	95% CI	*p*
Modularity	LLM – L1	0.21	[0.15, 0.27]	< 0.001
	LLM – L2	0.18	[0.13, 0.24]	< 0.001
	L1 – L2	−0.03	[−0.10, 0.05]	0.43
Average shortest path length	LLM – L1	1.80	[0.51, 3.43]	0.006
	LLM – L2	2.05	[0.79, 3.56]	< 0.001
	L1 – L2	0.25	[−0.59, 1.25]	0.59
Average degree	LLM – L1	1.33	[−0.31, 3.00]	0.12
	L2 – LLM	2.62	[0.92, 4.16]	0.002
	L2 – L1	3.95	[1.92, 6.08]	< 0.001
Clustering coefficient	LLM – L1	0.08	[0.01, 0.16]	0.03
	LLM – L2	−0.004	[−0.07, 0.06]	0.94
	L2 – L1	0.09	[0.01, 0.17]	0.02

Note: Δ represents the mean difference between groups (first minus second term in the comparison), computed across 1,000 paired bootstrap iterations. Confidence intervals correspond to the 2.5th and 97.5th percentiles of the bootstrap distribution. Two-tailed *p*-values were derived from the empirical distribution of differences.

Clustering coefficients show a partially similar pattern. L2 ($M = 0.469$, $SD = 0.022$) and LLM ($M = 0.465$, $SD = 0.025$) both exceed L1 ($M = 0.382$, $SD = 0.033$). Pairwise comparisons indicate that LLM differed from L1 ($\Delta = 0.08$, $p = 0.03$), as did L2 ($\Delta = 0.09$, $p = 0.02$), while LLM and L2 did not differ ($p = 0.94$).

Figure 2 provides a visual representation of the three networks, consistent with the pattern reported above. The LLM network is more clearly segmented into distinct clusters, often grouping items not typically produced together in human responses (e.g., *familia* [family]–*hogar* [home]–*techo* [ceiling]–*puerta* [door]), whereas the human networks show denser cross-connections between regions of the graph. The L2 network, in particular, exhibits the highest overall connectivity, consistent with its elevated average degree.

Across metrics, a consistent pattern emerges for both research questions: the LLM network shows higher modularity and longer path lengths than either human network, addressing RQ1, whereas the two human networks display comparable values on these measures, addressing RQ2. Differences in degree and clustering indicate variation in connectivity density, particularly in the L2 group.

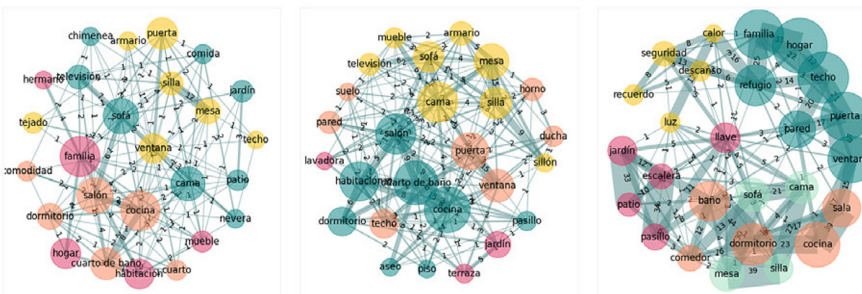


Figure 2. Semantic networks for L1 (left), L2 (center) and LLM (right). Node size reflects Lexical Availability Index (IDL), scaled using a common range across groups. Edge thickness represents adjacency frequency. Node colors indicate community membership as detected by the Louvain algorithm.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

A descriptive examination of betweenness centrality offers a complementary perspective. In the human networks, nodes with relatively elevated betweenness and moderate degree corresponded to spatial or transitional elements (e.g., *puerta* [door], *chimenea* [fireplace] in L1; *pasillo* [hallway], *jardín* [garden] in L2). In the LLM network, only *descanso* [rest] showed a comparable profile, and betweenness values were more unevenly distributed overall. Mean betweenness did not differ across groups (Kruskal–Wallis: $H = 0.36$, $p = 0.84$).

4.2. Lexical diversity and psycholinguistic measures

4.2.1. Cohesion

Descriptive measures of lexical diversity are presented in Table 3. The LLM displayed the highest cohesion (0.286), followed by L2 learners (0.109), whereas L1 speakers showed the lowest value (0.062). This pattern reflects progressively greater convergence in lexical selection from L1 to L2 and to the LLM. In parallel, the total number of unique types was highest in L1 (360), substantially lower in L2 (179) and lowest in the LLM group (140). Mean words per participant differed between the human groups (22.32 in L1; 19.48 in L2), while the LLM produced a fixed output of 40 items per run.

To evaluate whether the observed differences in cohesion were statistically reliable, permutation tests were conducted (5,000 permutations). The results are reported in Table 4. All pairwise comparisons were significant. The LLM showed higher cohesion than both L1 ($\Delta = 0.224$, $p < 0.001$) and L2 ($\Delta = 0.177$, $p < 0.001$), and L2 showed higher cohesion than L1 ($\Delta = -0.047$, $p < 0.001$), confirming that this progression is statistically reliable and not due to chance (RQ3).

4.2.2. Most available words

To further characterize the pattern of lexical convergence observed above, the highest values of the IDL were examined in each group. Table 5 presents the five most available words per group, together with their percentage of appearance across participants or model runs.

Table 3. Descriptive lexical diversity measures by group

Group	Cohesion index (*M/T*)	Total unique types	Mean words per participant
L1	0.062	360	22.32
L2	0.109	179	19.48
LLM	0.286	140	40.00

Note: Cohesion is calculated as the ratio between mean words per participant and total unique types (*M/T*). Higher values indicate greater lexical convergence.

Table 4. Permutation test results for the cohesion index

Comparison	Observed Δ	*p*
LLM – L1	0.224	<0.001
LLM – L2	0.177	<0.001
L1 – L2	−0.047	<0.001

Note: Δ represents the difference between groups (first minus second term). *p*-values are based on 5,000 permutations.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Table 5. Most available words by group

Group	Word	IDL	% Appearance
L1	*familia* (family)	0.50	64%
	cocina (kitchen)	0.49	68%
	salón (living room)	0.41	56%
	sofá (sofa)	0.40	58%
	cama (bed)	0.37	54%
L2	*cocina* (kitchen)	0.68	86%
	baño (bathroom)	0.63	92%
	salón (living room)	0.52	70%
	cama (bed)	0.52	88%
	habitación (room)	0.50	70%
LLM	*hogar* (home)	0.98	100%
	techo (ceiling)	0.86	100%
	familia (family)	0.82	100%
	puerta (door)	0.78	100%
	ventana (window)	0.74	100%

Note: Percentages represent the proportion of participants ($n = 50$) or model runs ($n = 50$) in which each word was produced.

The highest IDL value was observed in the LLM group (0.98 for *hogar* [home]). In comparison, the maximum values were 0.68 in the L2 group (*cocina* [kitchen]) and 0.50 in the L1 group (*familia* [family]).

In the LLM group, several words were produced in 100% of runs, including *hogar*, *techo*, *familia*, *puerta* and *ventana*. No lexical item reached full convergence in either human group. In L2, the highest percentage of appearance was 92% (*baño* [bathroom]), while in L1 the highest value was 68% (*cocina*).

Across groups, the most available items differ in their lexical profile. The LLM includes both concrete elements (e.g., *techo*, *puerta* and *ventana*) and more general terms such as *hogar*. In contrast, the most available items in the human groups primarily refer to elements of the domestic environment (e.g., *cocina*, *baño*, *cama* and *salón*), with some inclusion of non-physical concepts such as *familia* in L1.

4.3. Concreteness and familiarity

Concreteness and familiarity ratings are reported in Table 6. For concreteness, the L2 group showed the highest mean values ($M = 6.16$, $SD = 1.08$), followed by L1 ($M = 5.63$, $SD = 1.50$) and LLM ($M = 5.31$, $SD = 1.61$). A Kruskal–Wallis test indicated a significant difference among groups ($H = 27.04$, $p < 0.001$, $\varepsilon^2 = 0.04$), and *post hoc* comparisons using Dunn's test with Bonferroni correction confirmed that all pairwise differences were significant (L2 vs. L1: $p = 0.0003$; L1 vs. LLM: $p = 0.028$; L2 vs. LLM: $p < 0.0001$). These results indicate that the LLM produced, on average, less concrete words than both human groups, addressing RQ4.

For familiarity, the LLM showed the highest mean values ($M = 6.66$, $SD = 0.45$), followed by L2 ($M = 6.46$, $SD = 0.69$) and L1 ($M = 6.31$, $SD = 0.87$). A Kruskal–Wallis test indicated a significant overall difference ($H = 12.14$, $p = 0.002$, $\varepsilon^2 = 0.02$), with *post hoc* comparisons showing that the LLM differed significantly from both L1 ($p = 0.0008$) and L2 ($p = 0.017$), whereas the two human groups did not differ ($p = 0.78$). The absence of a difference between L1 and L2 matches the convergence already observed for network structure (Section 4.1), again addressing RQ2.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Table 6. Concreteness and familiarity by group

Measure	Group	*M*	*SD*	*n*
Concreteness	L1	5.63	1.50	341
	L2	6.16	1.08	173
	LLM	5.31	1.61	135
Familiarity	L1	6.31	0.87	345
	L2	6.46	0.69	173
	LLM	6.66	0.45	136

Note: Ratings on a 1–7 scale. Concreteness and familiarity values were obtained from Guasch et al. (2016) and Martínez et al. (2025). *n* indicates the number of lexical items with available ratings in each group.

Overall, the results show that LLM-generated vocabulary is characterized by high familiarity combined with lower concreteness relative to human responses.

4.4. Psycholinguistic correlates with lexical availability

To examine the relationship between lexical availability and psycholinguistic properties, Spearman rank-order correlations (ρ) were computed between the IDL and both concreteness and familiarity within each group (Table 7).

For concreteness, a significant positive correlation was observed in the L2 group ($\rho = 0.31$, $p < 0.001$), whereas no significant correlation was found in the L1 group ($\rho = 0.04$, $p = 0.48$) or the LLM group ($\rho = 0.16$, $p = 0.057$). For familiarity, both human groups showed significant positive correlations with availability (L1: $\rho = 0.31$, $p < 0.001$; L2: $\rho = 0.24$, $p = 0.002$), whereas no significant correlation was observed in the LLM group ($\rho = 0.05$, $p = 0.56$).

These results show that lexical availability in human speakers is systematically associated with familiarity, whereas no such relationship is observed in the LLM. Although the model produces highly familiar words overall (see Section 4.3), familiarity does not structure the relative availability of items within its output.

Together, the analyses reported above answer each of the four research questions guiding this study. Regarding RQ1, the LLM network showed significantly higher modularity and longer average path lengths than either human network, indicating a more compartmentalized and less globally integrated structure. Regarding RQ2, L1 and L2 networks did not differ reliably in modularity or average path length, and the two human groups showed a comparable relationship between familiarity and lexical availability, indicating convergence despite differences in acquisition history. Regarding RQ3, the LLM showed markedly lower lexical diversity and higher convergence across outputs than either human group. Regarding RQ4, the LLM produced systematically less concrete vocabulary than both human groups. Unlike the human groups, however, the LLM's higher overall familiarity ratings did not correspond to which items were more available within its own output.

Table 7. Spearman correlations between availability index and psycholinguistic variables

Variable	L1	L2	LLM
Concreteness	0.04	0.31***	0.16
Familiarity	0.31***	0.24**	0.05

Note: **$p < 0.01$, ***$p < 0.001$.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

5. Discussion

The present study compared paradigmatic association networks in native speakers of Spanish, advanced learners of Spanish as a foreign language and a large language model (ChatGPT) in order to characterize similarities and differences in lexical organization across human and model-generated systems. Two main patterns emerged across analyses. First, the two human groups displayed highly comparable network structures despite substantial differences in acquisition history and cumulative exposure to Spanish. Second, the model network differed systematically from both human groups in modularity, ASPL, lexical convergence, and in the lack of a relationship between lexical availability and familiarity. These findings bear on ongoing debates about the nature of linguistic knowledge in large language models and the conditions under which different forms of lexical organization emerge. While all systems produced lexically appropriate responses for the target category, similarities at the level of lexical content did not extend to the configuration of connections through which those items were organized. As discussed in Section 2.2, however, this configuration reflects not only how items are organized but also how that organization was searched during production. Accordingly, the patterns described below are consistent with organizational differences between systems, though our measures cannot fully separate the two.

The contrast between human convergence and model divergence helps to interpret the properties of model-generated networks. Advanced L2 learners have received substantially less cumulative exposure to Spanish than native speakers, yet their networks converge with those of native speakers on key structural measures. The model, by contrast, was trained on corpora far exceeding the input available to any individual speaker, but its network still diverges systematically from both human groups. This asymmetry suggests that the organizational properties observed in the model reflect characteristics of text-trained systems rather than being explained by differences in input quantity. The inclusion of advanced L2 learners is particularly informative in this respect, since purely L1–LLM comparisons leave open the possibility that differences reflect idiosyncratic properties of native acquisition rather than more general contrasts between human and model-generated lexical organization.

5.1. Lexical structure in the model network

The model network is informative about the type of lexical organization that emerges from large-scale textual training. The vocabulary generated by ChatGPT was clearly appropriate for the category *la casa*, including concrete elements of domestic space (e.g., *techo*, *puerta*, *ventana*) as well as more general terms referring to habitation (e.g., *hogar*, *refugio*). All groups showed coherent local clustering, indicating that distributional co-occurrence is sufficient to form coherent lexical groupings within semantic subdomains. Several items appeared in every run, and the cohesion index was markedly higher than in either human group, reflecting strong convergence on a restricted set of highly accessible items. Large-scale textual training therefore supports coherent categorical responses and stable local associations.

At the same time, the model network showed less global integration than either human network. It displayed stronger partitioning into internally dense communities with fewer cross-community connections, and nodes were on average more distant from one another. The model generated substantially fewer unique types despite

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

producing more tokens overall, and no human lexical item reached full convergence across participants.

The relationship between lexical availability and familiarity observed in both human groups was not present in the model. Although the LLM produced highly familiar vocabulary overall, familiarity did not structure the relative accessibility of items within its output. In human speakers, by contrast, more familiar words were also more available, indicating a graded relationship between experiential frequency and lexical retrieval that is not captured by the model. This pattern suggests that model-generated lexical organization reflects strong convergence on distributionally salient items, but lacks the graded experiential structuring observed in human retrieval.

A related pattern can be observed in the distribution of betweenness centrality. In the human networks, nodes corresponding to spatial or transitional elements (e.g., *puerta*, *pasillo*) contributed to linking different regions of the network, whereas in the model this connective role was less clearly distributed.

These results also bear on recent discussions about the use of large language models as surrogate participants in cognitive research. Treating model outputs as direct proxies for human behavior may create an illusion of generalizability, whereby surface-level similarity in responses masks deeper differences in the underlying organization of lexical knowledge (Crockett & Messeri, 2026).

5.2. Human convergence and theoretical implications

Against this background, the convergence between L1 and L2 networks helps to clarify the interpretation of the model results. Despite differences in acquisition trajectory and cumulative exposure, native speakers and advanced learners displayed comparable modularity and path length values, and both groups showed a consistent relationship between lexical availability and familiarity. This convergence indicates that the more integrated structure observed in human networks is not limited to native acquisition or early exposure, but emerges under diverse learning conditions. The divergence of the model therefore cannot be attributed solely to differences in input quantity, and instead points to systematic differences in how lexical associations become organized across systems.

The present findings are broadly consistent with recent comparative studies. Wang et al. (2025) report that networks derived from LLM outputs show weaker global integration and less flexible exploration of the semantic space than human networks. Xiao et al. (2025), using word association tasks, observe stronger clustering and lower diversity in LLM responses than in human data. Sánchez-Saus Laserna et al. 2026) similarly find that model-generated responses in Spanish semantic fluency tasks form denser and more cohesive networks, whereas human responses are more heterogeneous. The structural differences identified here therefore appear to reflect a broader pattern in model-generated lexical organization rather than an artifact of the present design. The inclusion of advanced L2 learners strengthens this interpretation by showing that convergence across human groups persists despite differences in acquisition history.

Overall, these findings suggest that the contrast between human and model-generated networks lies not in the availability of appropriate lexical items, but in the structural principles governing their organization. Both native speakers and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

advanced L2 learners converge on relatively integrated lexical systems despite differences in acquisition history, whereas the model shows stronger compartmentalization and higher convergence on a restricted set of items. This pattern indicates that text-trained language models capture coherent categorical knowledge but do not reproduce the graded, experience-based organization observed in human lexical systems. More broadly, the results highlight that similarities in output do not necessarily reflect shared organizational structure, and that network-level analyses provide a useful approach for identifying these differences.

Beyond the caution against treating model outputs as direct proxies for human lexical organization, the present contrast also points to a constructive use of language models in cognitive research. Text-trained models may provide a theoretically informative comparison condition: a system in which rich linguistic co-occurrence statistics are available in the absence of human developmental, perceptual and communicative experience. Comparing the network structures generated by such systems with those observed in human speakers may therefore help disentangle which aspects of lexical organization are compatible with distributional learning alone and which are associated with additional characteristics of human cognition. The present results should be interpreted as a step in that direction rather than as evidence that any single factor fully explains the observed differences.

6. Limitations and future directions

Several limitations should be acknowledged. First, the study focused on a single semantic category (*la casa* [the house]). Although this category elicits a broad range of associations, including concrete spatial elements as well as contextually grounded terms, future research should examine whether similar patterns emerge across other semantic domains.

Second, the analysis was restricted to a single language model (ChatGPT [GPT-5.2]). Although this choice was theoretically motivated and allows the present findings to be situated in direct relation to existing comparative work (Abramski et al., 2023; Suresh et al., 2023; Wang et al., 2025; Xiao et al., 2025), it limits the ability to determine whether the observed structural patterns are specific to this architecture or generalize across models, which remains an important direction for future research.

Third, although L2 participants represented multiple L1 backgrounds, potential L1-specific transfer effects were not analyzed separately. Future research could examine whether network structure varies systematically as a function of typological distance or L1 background.

In addition, there was an asymmetry in task constraints between human participants and the model. Human participants completed a timed task, whereas the model was prompted to produce a fixed number of items. Although this design choice was motivated by pilot analyses and by the need to obtain stable model output, differences in elicitation format may influence response distributions. The fixed-length prompt likely provided the model with greater opportunity to produce a wider range of associations than the timed human task. The persistence of higher modularity and convergence under these conditions therefore suggests that the observed differences are unlikely to be reducible to output length alone.

Finally, the networks were constructed using group-specific sets of highly available words, so the node sets were not identical across conditions. Structural metrics such

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

as modularity and path length, however, describe how the network as a whole is organized, not which specific words fill it, and the finding that human networks showed lower modularity and shorter path lengths than the model network held across four independent estimation approaches (see Section 3.7 and Supplementary Materials), including a generative model (U-INVITE; Zemla & Austerweil, 2018) that does not depend on frequency-based edge weighting or positional bigram adjacency. This indicates that the reported differences are not an artifact of the specific network-construction procedure. As discussed in Section 5, however, verbal fluency also reflects retrieval processes that these measures cannot isolate, so the pattern is consistent with organizational differences without being direct proof of them.

Moreover, the original networks were constructed using immediate sequential adjacency, an approach commonly adopted in semantic fluency research because of its simplicity and interpretability. Alternative edge definitions based on broader co-occurrence windows or model-based inference may capture additional semantic relationships not reflected by immediate adjacency alone. However, the convergence between the original analyses and the U-INVITE robustness analyses (Section 3.7), which do not rely on this specific operationalization, suggests that the main conclusions are unlikely to depend solely on this methodological choice. Future work could still compare window-based edge definitions directly against the adjacency criterion used here, since the present study only tested the model-based alternative through U-INVITE.

7. Conclusion

The present findings reveal systematic differences between human speakers and a text-trained language model in how their networks are structured. The model's responses gave rise to networks with more compartmentalized thematic clusters, whereas networks derived from human speakers' responses were more densely interconnected across different areas of the category, resulting in a more globally integrated configuration. A key result concerns the convergence between the two human groups: despite differences in acquisition history and cumulative exposure, native speakers and advanced L2 learners displayed comparable network structures, while the model showed stronger partitioning and higher convergence on a restricted set of items. These differences were robust across several alternative network-estimation procedures (see Section 3.7 and Supplementary Materials), suggesting that the overall pattern does not depend on a specific method of network construction. As discussed in Section 5, this pattern is consistent with differences in lexical organization, though our measures cannot fully separate this from differences in how words were retrieved during production.

These findings suggest that text-trained language models generate category-appropriate lexical responses but differ from human speakers in the global organization of the resulting lexical networks. The model also produced vocabulary that, while appropriate to the category, appeared less experientially grounded than human responses, as reflected in the absence of a familiarity–availability relationship observed in both human groups. The contrast between human convergence and model divergence highlights that similarities in lexical content do not necessarily reflect shared structural organization. These results also caution against treating model outputs as direct proxies for human lexical organization, since surface-level similarity in responses may mask deeper structural differences.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Examining network-level properties therefore provides a useful perspective for characterizing how lexical knowledge is structured across human and artificial systems. Future research should examine whether the patterns observed here generalize across semantic categories, languages and models, helping to identify which aspects of lexical organization are most sensitive to how a system acquires and uses language.

Supplementary material. The supplementary material for this article can be found at https://osf.io/emf6t/.

Data availability statement. The data, materials and analysis scripts supporting this study are available on the Open Science Framework at https://osf.io/emf6t/ (Sánchez-Saus Laserna, 2026).

Competing interests. The author declares none.

Declaration of Generative AI and AI-assisted technologies in the writing process. During the preparation of this manuscript, the author used Claude Sonnet 5, accessed via claude.ai between June and August 2026, to assist with literature searches. After using this tool, the author reviewed and verified all content and takes full responsibility for the accuracy of the published work.

References

Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3), 124. https://doi.org/10.3390/bdcc7030124

Agustín Llach, M. P., & Rubio, J. (2024). Navigating the mental lexicon: Network structures, lexical search and lexical retrieval. *Journal of Psycholinguistic Research*, 53(2), 21. https://doi.org/10.1007/s10936-024-10059-8

Avery, J. E., & Jones, M. N. (2018). Comparing models of semantic fluency: Do humans forage optimally, or walk randomly? In *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 118–123). University of California.

Ávila Muñoz, A. M., Sánchez Sáez, J. M., & Odishelidze, N. (2021). DispoCen. Mucho más que un programa para el cálculo de la disponibilidad léxica. *ELUA*, 35, 9–36. https://doi.org/10.14198/ELUA2021.35.1

Ávila-Muñoz, A. M., & Sánchez-Sáez, J. M. (2014). Fuzzy sets and prototype theory: Representational model of cognitive community structures based on lexical availability trials. *Review of Cognitive Linguistics*, 12(1), 133–159. https://doi.org/10.1075/rcl.12.1.05avi

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bartol Hernández, J. A. (2006). La disponibilidad léxica. *Revista Española de Lingüística*, 36(1), 349–378.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borodkin, K., Kenett, Y. N., Faust, M., & Mashal, N. (2016). When pumpkin is closer to onion than to squash: The structure of the second language lexicon. *Cognition*, 156, 60–70. https://doi.org/10.1016/j.cognition.2016.07.014

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Crockett, M., & Messeri, L. (2026). AI surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences*, 30(3), 203–215. https://doi.org/10.1016/j.tics.2025.09.012

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, 1–9.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1), 213–231. https://doi.org/10.3758/BRM.40.1.213

Digutsch, J., & Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Scientific Reports*, 13(1), 5035. https://doi.org/10.1038/s41598-023-32248-6

Dinno, A. (2017). *dunn.test: Dunn's test of multiple comparisons using rank sums* (R package version 1.3.5) [Computer software]. https://CRAN.R-project.org/package=dunn.test

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Blackwell.

Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2. *Studies in Second Language Acquisition*, 33(3), 373–398. https://doi.org/10.1017/S0272263111000027

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2), 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., Vélez de Mendizábal, N., Corominas-Murtra, B., Bejarano, B., Ardanza-Trevijano, S., Peraita, H., Wall, D. P., & Villoslada, P. (2011). The semantic organization of the animal category: Evidence from semantic verbal fluency and network theory. *Cognitive Processing*, 12(2), 183–196. https://doi.org/10.1007/s10339-010-0372-x

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, 48(4), 1358–1369. https://doi.org/10.3758/s13428-015-0684-y

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hernández Muñoz, N. (2006). *Hacia una teoría cognitiva integrada de la disponibilidad léxica: El léxico disponible de los estudiantes castellano-manchegos*. Ediciones de la Universidad de Castilla-La Mancha.

Hernández Muñoz, N., Izura, C., & Ellis, A. W. (2006). Cognitive aspects of lexical availability. *European Journal of Cognitive Psychology*, 18(5), 730–755. https://doi.org/10.1080/09541440500339119

Hernández Muñoz, N., & Tomé Cornejo, C. (2017). Léxico disponible en primera y segunda lengua: bases cognitivas. In F. del Barrio de la Rosa (Ed.), *Palabras Vocabulario Léxico* (pp. 99–102). Edizioni Ca' Foscari.

Hernández Muñoz, N., Tomé Cornejo, C., & López García, M. (2025). Análisis de las herramientas automáticas para el léxico disponible: LexPro. In I. C. Santos Díaz & A. M. Ávila Muñoz (Eds.), *Avances y desarrollo de los estudios sobre el léxico disponible. Una aproximación desde la variación léxica* (pp. 71–91). Peter Lang.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431–440. https://doi.org/10.1037/a0027373

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International conference on learning representations (ICLR 2020)*. https://openreview.net/forum?id=rygGQyrFvH

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(9), 1470–1489.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3), 373–381. https://doi.org/10.1017/S136672891000009X

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–38. https://doi.org/10.1017/S0140525X99001776

López-Chávez, J., & Strassburger, C. (2000). El diseño de una fórmula matemática para obtener un índice de disponibilidad léxica confiable. *Anuario de Letras*, 38, 227–251.

Mandler, J. M. (1984). *Stories, scripts, and scenes: Aspects of schema theory*. Lawrence Erlbaum Associates.

Martínez, G., Conde, J., Reviriego, P., & Brysbaert, M. (2025). AI-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 Spanish words. *Quarterly Journal of Experimental Psychology*, 78(10), 2272–2283. https://doi.org/10.1177/17470218241306694

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–278). McGraw-Hill.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 116–133.

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://arxiv.org/abs/2303.08774

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Paris-Colombo, G., Cabral-Carvalho, R. M., & Toro-Hernández, F. D. (2026). Comparing semantic navigation in humans and large language models using natural language processing. *arXiv preprint*. https://doi.org/10.48550/arXiv.2607.12195

Qiu, M., Brisebois, Z., & Sun, S. (2025). Can LLMs simulate human behavioral variability? A case study in the phonemic fluency task. *arXiv preprint*. https://arxiv.org/abs/2505.16164

R Core Team. (2023). *R:* A language and environment for statistical computing (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, 42(1–3), 107–142. https://doi.org/10.1016/0010-0277(92)90041-F

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–58). Erlbaum.

Sánchez-Saus Laserna, M. (2022). Redes semánticas, léxico disponible y didáctica del vocabulario en ELE: un análisis por niveles de español/Semantic networks, available lexicon and vocabulary didactics in SFL. An analysis in different Spanish learning levels. *Tejuelo*, 35(3), 167–204. https://doi.org/10.17398/1988-8430.35.3.167

Sánchez-Saus Laserna, M. (2026). *Data and analysis scripts: Lexical organization in humans and large language models* [Data set]. Open Science Framework. https://osf.io/emf6t/

Sánchez-Saus Laserna, M., & Álvarez Torres, V. (2024). Influencia de los contextos de aprendizaje en lexicón mental: productividad léxica y redes semánticas en estudiantes de ELE. *Revista de Lingüística y Lenguas Aplicadas*, 19, 204–217. https://doi.org/10.4995/rlyla.2024.20741

Sánchez-Saus Laserna, M., Crespo Miguel, M., Escoriza Morera, L., & Álvarez Torres, V. (2026). Semantic networks in humans and Large Language Models in Spanish: a semantic fluency approach. *Diacritica*, 40(1), 118–146. https://doi.org/10.21814/diacritica.6867

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

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L., & Rogers, T. T. (2023). Conceptual structure coheres in human cognition but not in large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 EMNLP* (pp. 722–738). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.47

Tomé Cornejo, C., Hernández Muñoz, N., & López García, M. (2025). The network as a process: On the role of the speaker in the construction of semantic networks in Spanish L1 and L2. *Caplletra*, 78, 33–52. https://doi.org/10.7203/caplletra.78.30264

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, 1(2), 219–247. https://doi.org/10.1515/LANGCOG.2009.011

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *The eleventh international conference on learning representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2203.11171

Wang, Y., Deng, Y., Wang, G., Li, T., Xiao, H., & Zhang, Y. (2025). The fluency-based semantic network of LLMs differs from humans. *Computers in Human Behavior: Artificial Humans*, 3, 100103. https://doi.org/10.1016/j.chbah.2024.100103

Xiao, B., Duan, X., Haslett, D. A., & Cai, Z. (2025). Human-likeness of LLMs in the mental lexicon. In *Proceedings of the 29th CoNLL* (pp. 586–601). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.conll-1.38

Zemla, J. C. (2022). Knowledge representations derived from semantic fluency data. *Frontiers in Psychology*, 13, 815860. https://doi.org/10.3389/fpsyg.2022.815860

Zemla, J. C., & Austerweil, J. L. (2017). Modeling semantic fluency data as search on a semantic network. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), Proceedings of the 39th annual conference of the Cognitive Science Society (pp. 3646–3651). *Cognitive Science Society*.

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational Brain & Behavior*, 1(1), 36–58. https://doi.org/10.1007/s42113-018-0003-7

Zemla, J. C., Cao, K., Mueller, K. D., & Austerweil, J. L. (2020). SNAFU: The semantic network and fluency utility. *Behavior Research Methods*, 52(4), 1681–1699. https://doi.org/10.3758/s13428-019-01343-w

Zhang, L., & Jones, M. N. (2022). Using "semantic scent" to predict item-specific clustering and switching patterns in memory search. In Proceedings of the 44th annual meeting of the cognitive science society (pp. 3236–3242). Cognitive Science Society.

Cite this article: Sánchez-Saus Laserna, M. (2026). Lexical organization in humans and large language models: Evidence from L1, L2 and LLM networks, *Language and Cognition*, 18, e67, 1–24. https://doi.org/10.1017/langcog.2026.10119

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 18:49:22, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/langcog.2026.10119