Bilingualism: Language and Cognition

cambridge.org/bil

Seeing, hearing, learning: Using pictorial and auditory stimuli to enhance second language acquisition

Melodie Bellegarda[1,2], Ana Belén García-Gámez[3] and Pedro Macizo[1,2]

[1]Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain; [2]Department of Experimental Psychology, Faculty of Psychology, University of Granada, Granada, Spain and [3]Faculty of Health Sciences, International University of La Rioja, Logroño, Spain

Research Article

Cite this article: Bellegarda, M., García-Gámez, A.B. and Macizo, P. (2026). Seeing, hearing, learning: Using pictorial and auditory stimuli to enhance second language acquisition. *Bilingualism: Language and Cognition* 1–16. https://doi.org/10.1017/S1366728926101692

Received: 29 January 2026 UTC
Revised: 22 May 2026 UTC
Accepted: 22 June 2026 UTC

Keywords:
bilingualism; vocabulary learning; adult foreign language acquisition; multimodal learning; environmental sounds

Corresponding author:
Melodie Bellegarda;
Email: mbellegarda@ugr.es

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

Environmental sounds (e.g., a dog barking) represent an underexplored modality for multisensory support in foreign language (FL) learning. Past research proposes that FL learning contexts enriched by relevant sensory stimuli (e.g., pictorial, gestural, auditory) strengthen memory representations. This study investigated whether combining auditory and pictorial modalities facilitates FL vocabulary acquisition. In an intraparticipant design, native Spanish speakers learned FL words under three auditory conditions: environmental sounds, neutral tones or silence. During learning, we aimed to foster the activation of conceptual representations through dual input: a relevant picture alongside the auditory condition. Vocabulary acquisition was assessed using four tasks: forward and backward translation, picture naming and lexical decision. Accuracy and reaction times were recorded. Results revealed that both semantically relevant and neutral sounds improved learning relative to silence. The findings suggest that auditory cues provide a benefit to FL word learning in adults, and this advantage operates independently of semantic congruency.

Highlights

- Multisensory strategies were applied to foreign language (FL) word learning.
- FL concepts were activated via pictures and three sound conditions.
- Sound presence improved performance on semantically mediated tasks over silence.
- Environmental sounds did not outperform neutral tones.
- Lexically mediated tasks were insensitive to sound conditions.

1. Introduction

If a picture is worth a thousand words, what is a sound worth? Environmental sounds, or everyday noises that we encounter in our surroundings, such as the rhythmic patter of rain against a windowpane, or the sharp ring of an alarm clock, are not merely registered in the brain as isolated auditory events or background noise: they convey meaning. Much like pictures, sounds function as powerful sensory cues that enrich the environment, evoking mental representations of associated events (e.g., rainfall; an alarm clock). The simultaneous engagement of perception and meaning plays a critical role in subsequent memory and learning, which in turn can play a role in actions (e.g., bringing an umbrella; getting up or hitting snooze). It is, therefore, worth considering how such sensory stimuli might be leveraged for practical applications. In this paper, we will explore how environmental sounds could support foreign language (FL) vocabulary learning.

In today's globalized world, multilingualism is a highly valued skill, offering both social and professional benefits (Adsera & Pytlikova, 2015; Chiswick & Miller, 2002; Fieles-Ahmad & Huber, 2022; Stöhr, 2015). Despite clear personal and societal benefits of FL skills, linguistic distance between languages and the cognitive, temporal and economic costs associated with learning continue to pose significant barriers to acquisition (Sprenger, 2024). Not all adults had the social and economic opportunity to learn languages as children, when language is acquired intuitively through immersive, multisensory experiences. For adults, acquiring a new language often presents formidable challenges: mastering unfamiliar vocabulary, grammatical rules, phonemic repertoires or even new writing conventions. It is possible that conventional, memorization-heavy pedagogical models of language learning (e.g., two-column vocabulary lists and written materials) lack the sensory engagement which would be enriching and lead to experientially driven learning (for reviews on adult FL learning, see Plonsky, 2011; Rasouli & Jafari, 2016).

Addressing these challenges calls for new, cognitively grounded approaches to adult language learning – approaches that align more closely with how humans naturally acquire meaning

© 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 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

through perception and sensory experience. There is a substantial body of research indicating that FL learning is improved when relevant information is delivered through multiple sensory channels rather than relying on just one modality (Li & Deng, 2023; Rahmanu & Molnár, 2024). This has been examined for FL vocabulary learning across various input channels including visual (Altarriba & Knickerbocker, 2011; Bates & Son, 2020; Yu & Liu, 2022), gestural (García-Gámez & Macizo, 2019, 2020, 2022; Sweller et al., 2020), olfactory (Honda et al., 2025; Xia et al., 2024) and auditory (Bellegarda et al., 2026; Kaplan-Rakowski & Loranc-Paszylk, 2019) modalities. A multimodal, or multisensory, learning environment may be particularly effective because it mirrors how real-world experiences naturally integrate diverse sensory inputs (Ellis, 2019; Shams & Seitz, 2008).

Several theoretical frameworks help explain why multisensory learning can strengthen memory and support vocabulary acquisition. According to Dual Coding Theory (Paivio, 1969), encoding information in both verbal and nonverbal (i.e., sensory) formats creates parallel memory traces, increasing the likelihood of retention and retrieval of said information. Closely related, Multimedia Learning Theory (Mayer, 2012) posits that learning is enhanced when information is presented across multiple modalities (e.g., visual and auditory), as learners can process information through partially independent channels. Specifically, according to the principle of *coherence* (see Mayer, 2024, for a review), information that aligns with the task objective (FL learning, in our case) would facilitate knowledge acquisition (images and sounds that share the same meaning), whereas learning would be hindered by the presence of information irrelevant to the task goals (no irrelevant semantic contents were added in our study). Other *redundant information*, such as descriptive text accompanying learning materials (see Mayer & Johnson, 2008), would hinder learning. In this study, we prioritize coherent information within the framework of Multimedia Learning Theory, by the inclusion of images and sounds that convey the same meaning, while avoiding redundant information, such as text or additional explanations during FL learning.

From another view, the Scaffolding Assumption (Eitel et al., 2013) further suggests that being exposed to a sensory stimulus (such as a picture) before corresponding linguistic input can preactivate global conceptual structures that serve as mental scaffolds for later comprehension. While these original formulations emphasized visual imagery, their underlying principles extend to other sensory modalities, which can likewise foster meaningful connections. Complementary to these perspectives, embodied cognition (Barsalou, 2008; Smith & Gasser, 2005) posits that cognition is grounded in sensorimotor experiences, such that understanding arises from interacting with the environment. From this standpoint, meaning is not merely conceived as abstract representation (i.e., the semantic category a concept belongs to); instead, semantic knowledge inherently involves sensory components (i.e., visual, auditory, olfactory, etc.) derived from individuals' interactions with the world. Thus, understanding a concept such as "dog" involves integrating sensory-perceptual features such as its appearance, movements and barking. By engaging multiple senses during learning, learners can trigger these perceptual-conceptual systems by mentally simulating real-world contexts, producing richer and more durable memory representations. Together, these frameworks buttress the idea that multimodal approaches do more than present information in varied formats: they actively link linguistic forms to conceptual and experiential knowledge, enhancing comprehension and long-term learning.

To understand how multisensory language learning fits within existing theoretical frameworks for FL learning, we can turn to the Bilingual Interactive Developmental Model (BIA-d, Grainger et al., 2010) which builds on its earlier iterations including the BIA (Grainger & Dijkstra, 1992), the BIA+ (Dijkstra & van Heuven, 2002), as well as the Revised Hierarchical Model (RHM, Kroll & Stewart, 1994). According to this framework, beginner learners initially depend heavily on their first language (L1) to mediate the understanding of new words in the FL. With continued exposure, however, learners begin to form direct associations between FL lexical forms and their meanings, gradually reducing their reliance on L1 translations. These emerging direct links are reinforced through Hebbian learning, the neural process summarized by the phrase "neurons that fire together wire together" (Hebb, 2005). In other words, each coactivation (i.e., "fire together") of an FL form and its associated conceptual representation strengthens the connection between the two (i.e., "wire together"). Within this language learning framework, multisensory inputs could facilitate learning by providing additional conceptual activation. Following Hebbian learning principles, each coactivation of an FL word and its conceptual referent strengthens their reciprocal mapping, with additional reinforcement when the concept is supported by sensory input (e.g., sounds or images). This process may accelerate the formation of direct FL-concept connections, reducing reliance on L1 mediation.

Among the cited sensory modalities, environmental sounds are a promising yet underexplored avenue for FL vocabulary learning. As Vanderveer (1979) notes, environmental sounds are the opposite of arbitrary: they directly indicate the events or objects that produce them, automatically evoking mental representations of those referents. The strong, direct associations make environmental sounds a uniquely rich source of sensory information that could support FL vocabulary acquisition (Ballas & Howard, 1987). Although environmental sounds are nonlinguistic, they nonetheless carry semantic content, as demonstrated in multiple experiments. For instance, they facilitate the recognition of both spoken and written words related to the sound stimulus (Frey et al., 2014; Orgs et al., 2006, 2007, 2008; Van Petten & Rheinfelder, 1995). Moreover, environmental sounds engage similar cognitive and neural mechanisms as words: event-related potential (ERP) studies show that meaningful sounds elicit N400 responses comparable to those elicited by words, whereas nonmeaningful sounds do not (Cummings et al., 2006, 2008; Dick et al., 2007). Functional magnetic resonance imaging (fMRI) and lesion studies further indicate that the semantic processing of environmental sounds and words relies on largely overlapping neural networks (Beauchamp et al., 2004; Saygin et al., 2003, 2005). Importantly, environmental sounds have also been shown to contribute to embodied cognition by linking perception and meaning: certain action-related sounds (e.g., claps or whistles) integrate at the neural level with action words, suggesting that hearing these sounds can activate corresponding motor representations (Grisoni et al., 2016, 2019).

To date and to our knowledge, there exist only two studies that directly examine how sounds affect FL vocabulary acquisition. Kaplan-Rakowski and Loranc-Paszylk (2019) researched how Polish university students learned low-frequency onomatopoeic English nouns (e.g., *lash*), which phonetically resemble the sounds they represent (e.g., the crack of a whip). Onomatopoeic sounds can be thought of as a subset of environmental sounds (e.g., rain would be an environmental sound, but patter would be an onomatopoeic sound because when said out loud, the pronunciation of the word resembles the meaning). In a classroom setting, participants learned the new words under four sound conditions: onomatopoeic sound only, spoken pronunciation of the new word only, both and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

no-audio. During learning, a Microsoft PowerPoint slide was displayed for 15 seconds and contained five components: the written FL word, its Polish translation, a definition in English, a corresponding image and the corresponding audio condition. The results indicated that presenting the onomatopoeic sound alone during the learning improved both the immediate and the delayed (seven-day) recall compared to the no-audio condition. In contrast, the spoken pronunciation condition and the condition in which both sounds were present (onomatopoeic and pronunciation) did not yield advantages over the no-audio condition. The authors proposed that the simultaneous presentation of both sound inputs may have exceeded the learners' working memory capacity, thereby limiting potential benefits from the dual auditory input. They explained from the perspective of the Cognitive Load Theory (Sweller, 1994) that the dual input could have introduced competing information, which divided attention across signals rather than integrate it. Nonetheless, the findings highlight that sounds that are contextually relevant (such as onomatopoeic sounds) can act as effective learning cues that support encoding and retrieval of word meanings.

More recently, Bellegarda et al. (2026) extended this line of inquiry by examining how environmental sounds affect FL word learning in Spanish-speaking university students. In their study, participants learned high-frequency Spanish-FL word pairs under three auditory conditions: environmental sounds (e.g., the sound of rain paired with the word *rain*), neutral tones or silence. Participants were exposed solely to the written word pairs and the sound condition, with no additional inputs. Subsequent testing revealed task-dependent effects. For tasks that relied heavily on semantic knowledge, both sound conditions (environmental sounds and neutral tones) tended to show higher accuracy than the silence condition. However, for tasks where form-based recognition was most important, the opposite pattern was observed. The absence of differences between the environmental sounds and neutral tones suggests that the observed benefit was driven by the general presence of auditory input that supported semantic access, rather than by deeper semantic enrichment. From an embodied cognition perspective, this pattern may indicate that the environmental sounds used in the study did not sufficiently strengthen the multisensory memory trace to active the perceptual-semantic representations fully.

The present study directly addresses this issue. To determine whether the advantage observed by Bellegarda et al. (2026) was purely due to auditory input or whether richer sensory information could further support semantic learning, we will introduce a critical methodological change: the inclusion of a visual image as an additional, novel source of conceptual input. The present design combines pictures and sounds to provide dual-sensory conceptual activation. This richer, semantically aligned input will allow us to test whether engaging multiple sensory modalities effectively supports embodied semantic systems and strengthens vocabulary learning.

1.1. The current study

Taken together, this emerging evidence points to the promise yet complexity of integrating environmental sounds into multimodal FL learning. For the present study, we designed an experiment to examine whether conceptual information could be activated solely through sensory input, while minimizing reliance on L1 mediation. Specifically, we asked whether visual and auditory cues (in the form of pictures and environmental sounds) presented before an FL word can jointly enhance conceptual representations and thereby facilitate FL word learning. Bellegarda et al. (2026) employed pairs of L1 and FL words in FL vocabulary learning, accompanied by various auditory stimuli. In this study, we replaced the L1 words with images representing the meanings of the FL words. This change was made for two reasons. Firstly, to promote semantic processing, and secondly, to reduce the influence of L1 during FL word learning. As an anonymous reviewer of an earlier version of this article noted, processing a picture can trigger associated lexical information (e.g., phonology and orthography; see the Independent Network Model, Caramazza, 1997). However, if we compare learning using L1–FL word pairs (Bellegarda et al., 2026) with learning using pictures–FL words (the current study), semantic processing would be greater and L1 lexical processing would be minimized in this study. Support for these predictions comes from classical studies. For example, Glaser and Glaser (1989) concluded through six experiments that words have privileged access to the lexicon, whereas pictures have privileged access to the semantic network.

Prior work has shown that pictures can serve as scaffolds that preactivate relevant schemas and improve subsequent comprehension (García-Gámez & Macizo, 2020; Yu & Liu, 2022), and that environmental sounds can reinforce links between referents and labels in some cases (Bellegarda et al., 2026; Kaplan-Rakowski & Loranc-Paszylk, 2019). Yet it remains unclear (a) whether FL-concept associations can be formed effectively when learning relies primarily on nonverbal (sensory) input, and (b) to what extent the combination of relevant environmental sounds and visual cues leads to stronger FL-concept associations.

To address these questions, we adopted a novel, dual-sensory conceptual activation paradigm tailored to FL learning and designed to amplify the multisensory memory trace. We selected high-frequency concrete nouns typical of early FL instruction and paired each item with a written FL[1] target, a picture and a sound. During learning, each concept was activated via a picture and one of three sound conditions: a relevant environmental sound "congruent" with the picture, a neutral sound or silence. In the congruent condition, the environmental sound (e.g., a bee buzzing) was semantically related to the picture (e.g., a bee), providing coherent dual-sensory activation. The neutral condition used a single-frequency tone that was acoustically salient but semantically irrelevant; this condition controlled for the general contribution of auditory presence and tested whether non-semantic auditory input could support learning by promoting increased learner engagement during encoding (self-involvement theory, Helstrup, 1987). Our third condition, the silence condition, served as a baseline for FL learning with only visual activation. We intentionally did not include incongruent pairings because our aim was to model naturalistic learning contexts (where mismatches are uncommon) and to test facilitation as our primary goal, rather than interference.

Participants performed four evaluation tasks designed to assess different aspects of vocabulary learning: forward and backward translation tasks, a picture naming task, and a lexical decision task. While all four tasks were completed after the entire learning phase, the translation tasks were additionally completed halfway through the learning. This allowed us to track how learning progressed over time (see García-Gámez & Macizo, 2023), and specifically, to examine how lexical and semantic representations evolved under different sound conditions. We selected the four tasks because they are widely used in FL learning research and provide evidence for formation of semantic and lexical links between new FL words, L1

[1]FL words were created artificially, to eliminate any language familiarity bias. Additionally, the items were unrelated to the L1 words and each other, resembling how new words are typically added to lexicon (Leach & Samuel, 2007).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

words and conceptual knowledge (Grainger et al., 2010; Lindsay & Gaskell, 2013; Mestres-Missé et al., 2007).

Additionally, using these four tasks enables us to eliminate the possibility that differences in participants' performance are due to the specific demands of each task rather than the fact that two of them are semantically mediated (forward translation and picture naming) and two are lexically mediated (backward translation and lexical decision).

The forward translation and picture naming tasks required participants to engage in explicit semantic processing and are thus considered conceptually mediated (Kroll & Stewart, 1994; Sholl et al., 1995). In the forward translation task, participants saw a Spanish item and produced its FL equivalent, while in the picture naming task they named pictures representing the learned concepts in the FL. Successful performance on both tasks relied on access to the semantic network: the forward translation task via activation of the conceptual representation associated with the L1 word and the picture naming task through conceptual activation from the image (de Groot & van Hell, 2005; Vigliocco et al., 2002). We predicted that the enriched congruent condition (learning with a semantically relevant visual and auditory cues) would enhance performance on these tasks, as the meaningful inputs should strengthen conceptual representations, rendering them more vivid and accessible.

In contrast, the backward translation and lexical decision tasks do not rely on explicit semantic access, but can still reveal the organization and consolidation of the FL lexicon (Goldinger, 1996; Jusezyk & Luce, 2002; Strauß et al., 2022; Vitevitch, 2003). In the backward translation task, participants saw a word in the FL and provided its Spanish equivalent, while in the lexical decision task they decided whether a visually presented FL word had been learned or was novel. Prior findings indicate that learning methods emphasizing semantic mediation not only support the development of conceptual links but also facilitate lexical consolidation (García-Gámez & Macizo, 2019; García-Gámez et al., 2024). Therefore, it might be possible that the congruent sound condition would also promote more robust lexical organization, leading to better performance on these lexically mediated tasks.

With regard to the other conditions, we anticipated a better performance for words learned with a neutral tone relative to silence. In both these conditions, conceptual information was provided largely by the picture. However, although the tone lacked semantic meaning, its presence may still enrich memory representations by enhancing engagement and learner self-involvement during encoding (Helstrup, 1987).

We are aware that there is evidence that previous language experience (LEX) (particularly multilingualism) and socioeconomic status (SES) are both factors that could play a role in language learning. LEX has been linked to enhanced cognitive flexibility and metalinguistic awareness with regard to language tasks (Bonnet & Siemund, 2018; Cenoz, 2013; Hirosh & Degani, 2018; Kaiser et al., 2015; Kaushanskaya & Marian, 2009). Similarly, SES, a multidimensional construct that includes parental education, income and access to resources, has been shown to influence both cognitive and linguistic development. Although its effects are more pronounced in childhood (Abo Hamza et al., 2024), SES-related disparities can persist into adulthood (Dong, 2024). Therefore, we propose measuring LEX and SES with the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) and the MacArthur Scale of Subjective Social Status (Adler et al., 2000), respectively, to account for their potential impact on FL acquisition. As part of an exploratory analysis, these variables will be included to examine how they may modulate learning outcomes, although no specific predictions were made.

2. Methods

2.1. Participants

A total of 40 native Spanish speakers (25 women; M_{age} = 21.25, SD_{age} = 3.32) from the University of Granada took part in the study in return for course credit. None of the participants reported visual, auditory, neurological or language-related impairments. Written informed consent was obtained before participation. The study procedures followed national and institutional guidelines, and were consistent with the principles on human experimentation outlined in the Helsinki Declaration (1964; latest revision 2024). The experimental protocol was reviewed and approved by the University of Granada Ethics Committee (approval number: 957/CEIH/2019).

The sample size required for this study was estimated using G*Power (Faul et al., 2009) based on a repeated-measures ANOVA with one within-subject factor (sound condition, 3 levels). This approach was used as a pragmatic and conservative approximation, as simulation-based power estimation for mixed-effects models requires a priori specification of variance components and random-effects structures, which are often unavailable before data collection (Kumle et al., 2021). A total of N = 36 participants was needed to achieve 95% statistical power with α = .05 and a medium effect size (Cohen's f = 0.275, η^2p = .07).

Before beginning the experiment, participants completed two questionnaires to provide demographic information. SES was assessed with the MacArthur's Scale of Subjective Social Status (Adler et al., 2000), in which individuals placed themselves on a ladder with 10 rungs, representing their relative standing with regards to income, education and occupation. The participants' mean subjective rating for SES was M = 6.15, SD = 1.33, range 3–8.

Language background and experience (LEX) were evaluated using a modified version of the LEAP-Q (Marian et al., 2007). Participants reported the languages they knew, ordered by acquisition, fluency and frequency of exposure. They also indicated what percentage of time they would typically read or speak in each language and provided self-ratings of reading, writing, listening and speaking abilities (1–10 scale) for all languages in which they had experience. In addition, participants indicated the cultural groups they self-identified with. To quantify LEX, we calculated average second language (L2) proficiency across the four skill domains (reading, writing, speaking, listening), since all participants reported knowledge of at least two languages. This was relevant given evidence that prior LEX can influence subsequent language learning (Hirosh & Degani, 2018). On average, participants reported speaking 2.68 languages (SD = 0.92, range 2–6). Participants self-reported their L2 proficiency skills as M = 7.23, SD = 1.33, range 3.5–10. Summary statistics for these language measures are presented in Table 1.

2.2. Design and materials

The data, analyses, study materials and supplementary files supporting the findings of this research are openly available in the Open Science Framework (OSF) repository at: https://doi.org/10.17605/OSF.IO/6QAZ2. In this study, participants learned 42 FL words. Each word was presented visually within a sequence that began with a picture representing the word, which was accompanied either by silence, a congruent sound or a neutral tone. Using a within-subject design, 14 words were assigned to each condition, totaling 42 new words. Previous research suggests that this number of items can be learned comfortably in one session with a comparable number of repetitions (Bellegarda et al., 2026; García-Gámez & Macizo, 2019, 2023;

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

Table 1. Language use and proficiency by order of acquisition as well as cultural identities self-reported by participants

Language use				
Language	% Daily exposure	% Reading	% Speaking	
L1	80.95 (11.07)	79.23 (15.45)	77.43 (27.57)	
L2	17.20 (9.87)	19.65 (14.43)	20.73 (26.44)	
L3	3.44 (3.50)	2.78 (6.28)	4.00 (8.67)	
L4	2.43 (1.99)	0.71 (1.50)	0.86 (1.57)	
L5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
L6	5.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
Language proficiency				
Language	Reading	Writing	Speaking	Listening
L1	9.75 (0.49)	9.68 (0.53)	9.65 (0.66)	9.85 (0.36)
L2	7.60 (1.24)	6.83 (1.48)	6.98 (1.75)	7.50 (1.77)
L3	5.28 (2.44)	3.83 (1.95)	4.17 (2.26)	4.94 (2.26)
L4	3.00 (3.03)	2.33 (1.97)	2.67 (1.75)	4.00 (3.58)
L5	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
L6	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Cultural identity				
Cultural group	% Identifying	Strength of identification		
Spanish (national)	94.87	6.97 (2.03)		
Spanish (regional or local)[a]	84.62	7.86 (2.32)		
Anglophone[b]	25.64	4.20 (2.94)		
Mediterranean	15.38	8.17 (1.94)		
Non-Iberian European[c]	10.26	3.25 (1.50)		
Latin American[d]	7.69	6.83 (3.55)		
Arab/Moroccan	7.69	5.00 (4.36)		

Note: Language use measures reported mean daily exposure (i.e., percentage of time participants reported exposure to each language) and preferences for reading and speaking (i.e., percentage of time participants reported choosing to read or speak in each language). Participants reported experience with 2–6 languages: there were 22 bilinguals, 11 trilinguals and 7 multilinguals with experience with 4 or more languages. Language proficiency reports mean proficiency across 4 dimensions (reading, speaking, writing, listening), on a scale of 1–10. Cultural identity measures the percentage of participants identifying with each cultural group, along with the mean strength of identification (scale of 1–10). Participants could indicate multiple cultural identities. Standard deviations are shown in parentheses.

[a]Includes Andalusian, Asturian, Canary Islander, Castilian, Catalonian, Extremaduran, and Granadan.

[b]Includes English.

[c]Includes French, German, and Italian.

[d]Includes Chilean, Cuban, and Mexican.

García-Gámez et al., 2024). The assignment of words to the different sound conditions was counterbalanced across participants.

2.3. Materials

2.3.1. L1 and FL words

A total of 42 Spanish nouns were chosen, with half representing natural entities (e.g., "gato," or "cat" in English) and half representing man-made entities (e.g., "alarma," or "alarm" in English). All nouns were familiar words (M = 5.83, SD = 0.51, scale 1–7 with 7 indicating maximum familiarity) and had an average lexical frequency of 22.12 (SD = 51.46) occurrences per million words (SUBTLEX-ES corpus, Cuetos et al., 2011). Eighty-four FL words were created conforming to Spanish phonological and orthographical patterns. The FL items had few Spanish orthographic neighbors (M = 0.13, SD = 0.46). Half of the FL words were used in the learning task, while the remaining half of the FL word served as pseudowords ("no" responses) in the lexical decision task, as they were never presented to the learners. In the learning task, an FL word was randomly paired with one of the Spanish nouns (although creating pairings that began with the same initial phoneme was avoided). The 42-word pairs were grouped into 3 sets of 14 words, which were matched on multiple lexical-semantic variables for the Spanish words (lexical frequency, number of syllables, orthographic length, number of orthographic neighbors, number of shared graphemes with the FL pair, familiarity, imageability and concreteness) and lexical variables for the FL words (orthographic length, number of orthographic[2] neighbors and number of syllables). Data on number of orthographic neighbors, familiarity, imageability and concreteness were consulted in EsPal database

[2]Because existing databases do not provide orthographic neighborhood counts for nonwords, we developed a custom script that compared each FL word with a list of 100,880 Spanish words to calculate its orthographic neighborhood size.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

(Duchon et al., 2013). For the lexical decision task, "learned" (FL real words) and "new" items (nonpracticed FL words that functioned as pseudowords) were matched in terms of number of syllables, number of orthographic neighbors and orthographic length. The full set of linguistic materials is available in the Supplementary Materials, Table S1, with descriptive statistics for the learning and lexical decision tasks provided in Supplementary Tables S2 and S3, respectively.

2.3.2. Sounds

Forty-two environmental sounds corresponding to the 42 selected Spanish nouns were adapted from Bellegarda et al. (2026). These sounds belonged to various semantic categories: alarms/alerts (n = 4, e.g., alarm clock ringing), household sounds (n = 6, e.g., someone pulling a curtain), sports (n = 1, e.g., someone bowling), tools (n = 7, e.g., someone hammering), transport (n = 3, e.g., a car starting), animal vocalizations (n = 11, e.g., horse whinnying), environmental phenomena (n = 8, e.g., wind whistling), and insect sounds (n = 2, e.g., bee buzzing). The neutral sound condition consisted of 14 monochromatic unique sinusoidal tones ranging from 525 to 1000 Hz, created using Praat software (Boersma & Weenink, 2020). Both the environmental sounds and the tones lasted 4 seconds, were digitized at 48 kHz with a 16-bit sampling rate, and were normalized to an RMS level of 70 dB SPL (Dittinger et al., 2016).

2.3.3. Pictures

Forty-two colored pictures representing the meaning of the Spanish-FL word pairs were selected from the MultiPic database (Duñabeitia et al., 2018). For items not available in that database, additional stimuli were sourced from the PicPsy database (Martínez et al., 2020) and other databases with public domain licenses. These pictures were used in a picture familiarization task, the learning task and the picture naming task.

To verify that the selected sounds and pictures accurately represented the meaning of the Spanish nouns and were easily recognizable, we conducted a norming study with 23 participants who were not involved in the main experiment. Participants completed three rating tasks. In the first task, they heard an auditory stimulus while a Spanish word appeared on the screen and rated how well the word's meaning matched the sound on a 7-point Likert scale (1 = "high mismatch," 7 = "high match"). In the second and third tasks, the procedure was the same, but participants judged either picture-word or picture-sound pairings. To prevent fatigue and avoid excessive repetition (since the 42 original stimuli would otherwise only appear in congruent combinations), each stimulus was shown twice per task: once in a congruent pairing and once in an incongruent one. In incongruent trials, the mismatched item always came from the same semantic category (e.g., an oinking sound was paired with "marrano" [hog in English] in the congruent condition, and with "caballo" [horse in English] in the incongruent condition). In total, participants rated 252 stimuli pairs (84 per task) in a session lasting roughly 20 minutes. Results confirmed that participants consistently distinguished congruency across all modalities. Sound-word pairs received higher ratings when congruent (M = 6.46, SD = 1.30) than when incongruent (M = 1.62, SD = 1.45), $t(22)$ = 34.55, p < .001. Picture-word pairs were also rated higher in the congruent (M = 6.84, SD = 0.82) than the incongruent (M = 1.34, SD = 1.10) condition, $t(22)$ = 45.89, p < .001. Finally, picture-sound pairs showed the same pattern (congruent: M = 6.63, SD = 1.23; incongruent: M = 1.38, SD = 1.23), $t(22)$ = 45.12, p < .001.

2.4. Familiarization task

Participants first completed an initial familiarization phase, which was included to ensure that the pictures were consistently associated with the intended conceptual meanings of the words to be learned. Participants viewed 42 pictures along with their corresponding Spanish labels (as described in *Materials*). Each label corresponded to the noun most commonly associated with the depicted concept.[3] The pictures were presented one by one in random order, and participants were instructed to take their time to study the images and words, advancing at their own pace by pressing the spacebar.

2.5. Learning task

The three sets of Spanish-FL word pairs described in the Materials section were each assigned to one of three sound conditions (silence, congruent sound or neutral tone). The assignment of sets to conditions was counterbalanced across the participants. Learning took place in a blocked manner, with one block per sound condition, see Figure 1. Blocked presentation, rather than interleaving sound trials, was used to avoid introducing additional attentional or carry-over effects between trials. One round of learning included all 3 blocks, providing exposure to all 42 FL words and all 3 sound conditions. Participants completed 10 rounds of learning in total, resulting in 10 exposures to each item (420 trials in all). They were given a short break after every two rounds (approximately every 14 minutes). Each trial began with a central fixation point (500 ms), followed by the display of a picture accompanied by the sound condition (silence, congruent sound or neutral tone) (4000 ms). The corresponding written FL word then appeared centrally (4000 ms), followed by an intertrial interval (500 ms). Participants were told that each picture represented the meaning of the upcoming FL word and were instructed to focus on learning the word meanings. Within each sound block, the order of the FL word-picture pairs was randomized.

2.6. Evaluation tasks

2.6.1. Forward translation task

Participants completed an oral forward translation task where the written target was presented in Spanish and an oral response was given in FL. This task was administered at two timepoints: after 5 exposures to the items in the learning task and after 10 exposures. At the first timepoint, participants translated half of the words (21 items) and at the second timepoint, participants translated the remaining items (21 words).

2.6.2. Backward translation task

Participants also completed an oral backward translation task where the written target was presented in FL and an oral response was given in Spanish. Similarly, this task was administered at two timepoints during the experiment: after 5 exposures to the items in the learning task and again after 10 exposures. At the first timepoint, participants translated half of the words (21 items) and at the

[3]Items drawn from standardized databases (MultiPic and PicPsy) showed high levels of picture-name agreement (M = 96%, SD = 7%). Items not available in these databases were validated in a separate naming study with 95 Spanish-speaking participants, confirming high picture-name agreement for the selected stimuli (M = 82%, SD = 16%). In some cases, we used close synonyms (e.g., "ganso" [wild goose] vs. "oca" [domesticated goose]) of the modal response to ensure consistency in linguistic properties across our selected items.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

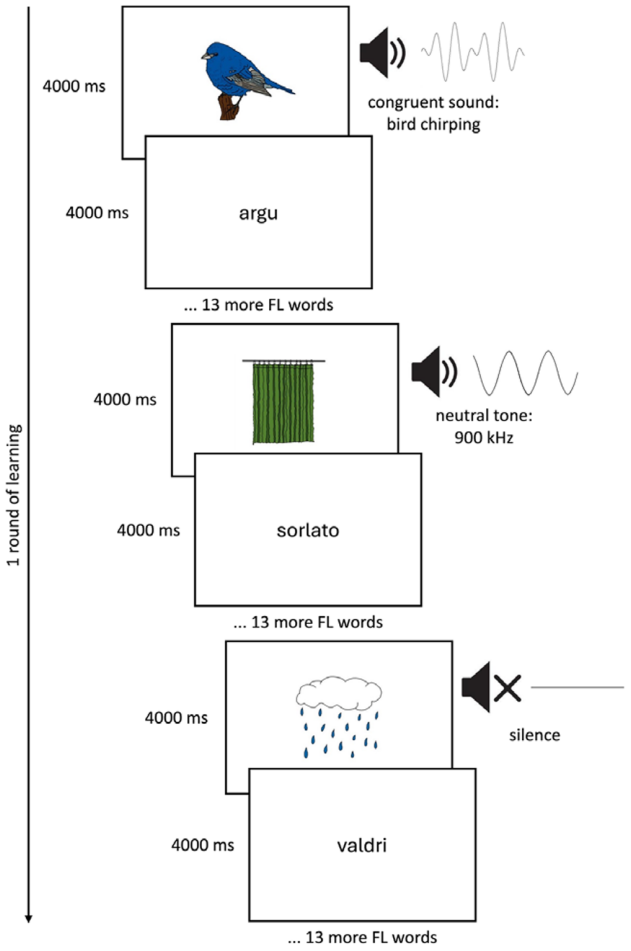


Figure 1. Description of one round of foreign language (FL) word learning. Participants learned 42 items in a blocked design, with 14 items in each block, and each block associated to a sound condition (congruent environmental sound, neutral tone or silence).

second timepoint, participants translated the remaining items (21 words).

The order of the performance of the translation task (forward or backward) at each of the two timepoints was counterbalanced across the participants. The assignment of items to Timepoint 1 or Timepoint 2 was counterbalanced across participants, so that each word appeared equally often at each timepoint across the sample. Across both timepoints, all 84 items (42 Spanish and 42 FL) were translated once, ensuring that no item was repeated across tasks. Within each translation task, items were presented in a random order.

Both translation tasks had similar trial events. Each trial began with a centrally presented fixation point (500 ms), followed by the written display of a Spanish or FL stimulus (500 ms). A question mark (?) then appeared and remained on the screen until the participants provided their oral response, which triggered the voice-activated key. The trial then concluded with a blank screen (450 ms). Participants were instructed to respond loudly and clearly, focusing on both speed and accuracy and to avoid stuttering.

2.6.3. Picture naming task

Participants saw the 42 pictures from the learning task and were instructed to name each aloud using the FL vocabulary they had learned previously, focusing on speed and accuracy. Each trial began with a centrally displayed fixation point (300 ms), followed by an image (e.g., the picture of a door, 10 × 10 cm) appearing centrally (500 ms). A question mark (?) signaled that the participant could provide his or her spoken response. This prompt remained on screen until the response was made and the voice-key triggered, after which a blank screen was shown (450 ms). The pictures were presented in random order.

2.6.4. Lexical decision task

Participants were asked to indicate, using the M and Z keys, whether a presented item was a learned, real FL word or an FL pseudoword. The assignment of the keys to "yes" or "no" responses was counterbalanced across participants. Each trial began with a centrally presented fixation point (300 ms), followed by the target word presented visually, which remained on screen until a response was made. After the response, a blank screen was displayed (450 ms) before the next trial. The task included a total of 84 trials: 42 trials with previously learned words, and 42 new, never-seen words. All stimuli were presented in random order.

2.7. Procedure

Before coming to the laboratory, participants completed online questionnaires (LEAP-Q and SES) via a link provided on the recruitment platform. In the laboratory, they performed the learning and evaluation tasks in a single session lasting approximately 110 minutes, though the duration varied depending on the individual performance. The session took place in a quiet room to minimize distractions, and participants wore a Trust Mauro USB headset to listen to the binaural stimuli. Initially, participants completed 48 trials of a visual Stroop task (Stroop, 1935), which lasted about 5 minutes; the materials were from Bellegarda and Macizo (2021). Data from this task were collected for a different purpose and are not analyzed in this paper. Next, they completed the familiarization task (2 minutes), followed by the first half of the learning task (35 minutes), the translation tasks (5 minutes), then the second half of the learning task (35 minutes) and again the translation tasks (5 minutes). After a short break, participants performed the picture naming task (10 minutes) and the lexical decision task (10 minutes). The task order was fixed, with picture naming preceding the lexical decision task. The lexical decision task was intentionally placed last to avoid giving only a subset of participants extra exposure to the target FL words (counterbalancing would have resulted in unequal input across groups). All stimuli were presented using E-Prime 2.0 (Schneider et al., 2012) on a 19-inch Captiva E1903 monitor. Phonation initiation during the translation and picture naming tasks was recorded using an ATR20 Audio-Technica Cardioid Low Impedance microphone connected to a PST serial response box (Psychology Software Tools Inc.), with responses also captured on a Sony ICD PX-440 voice recorder.

2.8. Data analysis

The primary index of FL word learning was task accuracy (translations, picture naming and lexical decision) and discrimination ability (d' scores), as successful performance presupposes FL learning. However, reaction times (RTs) were also examined.

Statistical analyses were conducted using linear mixed-effects models for continuous data and general linear mixed-effects models for binomial data using the *lme4* package (Bates, Maechler, et al., 2015) in R version 4.2.1 (R Core Team, 2022). Statistical

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

models included random intercepts for participants and items. Random slopes were initially considered but could not be retained in the final models due to convergence issues; therefore, parsimonious random structures were adopted in line with recommendations to capture data variability without overfitting (Bates, Kliegl, et al., 2015). For the RT analyses, only correct responses were considered. Univariate outliers exceeding ±3 SD from each participant's mean were removed, and RTs were log-transformed (natural log). The significance of fixed-effects and interactions was evaluated using likelihood ratio tests, and estimated marginal means were obtained via *emmeans* package (Lenth, 2023). For binomial mixed-effects models, pairwise comparisons of estimated marginal means were exponentiated to obtain odds ratios (ORs), which quantify the differences in the odds of a correct response between conditions. ORs > 1 indicate higher odds in the first condition of the contrast relative to the second, whereas ORs < 1 indicate lower odds. Ninety-five percent confidence intervals (CIs) are also reported on the OR scale. Tukey's correction (Tukey, 1953) was applied to adjust for multiple comparisons in *post hoc* analyses. Full model structures are provided in Table 2.

For the binary coding of verbal response data in the translation and picture naming tasks, responses were scored as 1 (correct) if they met the criteria for correctness and given a 0 (incorrect) otherwise. Minor errors were tolerated depending on the target word's length: (a) for monosyllabic words, a vowel substitution; (b) for disyllabic words, a vowel or consonant substitution, but not both; (c) for words with three or more syllables, a vowel-consonant inversion or the replacement of a vowel or consonant. These criteria follow those used in prior research employing verbal response data in FL word learning (García-Gámez & Macizo, 2019; García-Gámez et al., 2024). Semantically equivalent responses (e.g., responding "cerdo" instead of "marrano," comparable to "pig" instead of "hog") were also treated as correct (score = 1). For RT analyses, data points were excluded if: (a) the voice key was triggered by nonverbal sounds or (b) the participant stuttered or hesitated during the response.

Table 2. Summary of mixed-effects models by task and dependent variable (DV)

Task	DV	Model formula
Semantically mediated		
Forward translation	ACC	glmer(ACC ~ sound * timepoint + (1|participant) + (1|item)
	RT	lmer(logRT ~ sound * timepoint + (1|participant) + (1|item)
Picture naming	ACC	glmer(ACC ~ sound + (1|participant) + (1|item)
	RT	lmer(logRT ~ sound + (1|participant) + (1|item)
Lexically mediated		
Backward translation	ACC	glmer(ACC ~ sound * timepoint + (1|participant) + (1|item)
	RT	lmer(logRT ~ sound * timepoint + (1|participant) + (1|item)
Lexical decision	ACC	glmer(ACC ~ sound + (1|participant) + (1|item)
	d'	lmer(dprime ~ sound + (1|participant) + (1|item)
	RT	lmer(logRT ~ sound + (1|participant) + (1|item)

Note: ACC = accuracy; RT = response time (ms); logRT = natural logarithm of response time (ms); glmer = generalized linear mixed-effects model; lmer = linear mixed-effects model. Fixed-effect predictors included Timepoint (levels: 1, 2) and/or Sound (levels: congruent, silence, tone), depending on task. All models included random intercepts for participants and items.

2.8.1. Questionnaires

In an exploratory analysis, we extended our primary model to include two additional fixed-effect predictors (SES and LEX), given previous findings suggesting their potential influence on learning outcomes. Both predictors were scaled and centered before analysis. For each task and each dependent variable (accuracy, RTs, d'), two additional models were tested, one including SES and one including LEX. To reduce the likelihood of false positives due to multiple comparisons, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was applied to control the false discovery rate (FDR) across all exploratory analyses.

2.9. Semantically mediated tasks

2.9.1. Forward translation task

A generalized linear mixed-effects model was fitted to the binomial accuracy data (correct and incorrect responses), and a linear mixed-effects model was fitted to the RT data. The models included Timepoint (1, 2) and Sound (congruent, silence, neutral) as fixed-effects predictors.

2.9.2. Picture naming task

A generalized linear mixed-effects model was used to assess accuracy (correct and incorrect responses), and a linear mixed-effects model was used for the RTs. In both models, Sound (congruent, silence, neutral) was entered as the fixed-effect predictor.

2.10. Lexically mediated tasks

2.10.1. Backward translation task

A generalized linear mixed-effects model was fitted to the binomial accuracy data (correct and incorrect responses) and a linear mixed-effects model was fitted to the RT data. The models included Timepoint (1, 2) and Sound (congruent, silence, neutral) as fixed-effects predictors.

2.10.2. Lexical decision task

Binomial accuracy data (correct and incorrect responses) were analyzed with a generalized linear mixed-effects model. To account for the two types of errors, misses (or, responding "new" to an old item) and false alarms (or, responding "old" to a new item), we calculated d' values by z-transforming hit and false alarm rates for each participant (MacMillan & Creelman, 2005). These d' values were analyzed with a linear mixed-effects model. An additional linear mixed-effects model was fitted to the RT data. Sound (congruent, silence, neutral) served as the fixed-effect predictor in all models.

3. Results

Participants whose accuracy on any evaluation task was less than 10% after completing all 10 exposures to the FL items were excluded from the analyses for that specific task, in line with the criteria established by García-Gámez and Macizo (2020). Accordingly, one participant was removed from the translation task analyses at the second timepoint, and two participants were removed from the picture naming task analysis. The results[4] presented below

[4]Following a reviewer's suggestion, we conducted additional analyses including the factor Type (natural vs. man-made) and its interaction with Sound (congruent, tone, silence). No significant Sound x Type interactions were observed (all $ps > .14$), and the overall pattern of results remained unchanged when Type was included in the models. These findings indicate that the effects of auditory input were consistent across item categories. Full results are available on OSF.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

Table 3. Mean accuracy (%) and response times (RT in ms and logRT) by evaluation task and sound condition

		Accuracy		RT in ms		logRT	
Task	Condition	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI	*M* (*SD*)	95% CI
Forward translation, T1	Congruent	47.86 (50.04)	[42.00, 53.72]	3188 (2604)	[2721, 3656]	7.79 (0.74)	[7.66, 7.92]
	Silence	37.86 (48.59)	[32.17, 43.55]	3638 (2950)	[3051, 4224]	7.92 (0.74)	[7.77, 8.07]
	Tone	47.50 (50.03)	[41.64, 53.36]	2862 (2528)	[2402, 3322]	7.70 (0.69)	[7.57, 7.82]
Forward translation, T2	Congruent	76.92 (42.21)	[71.92, 81.93]	2856 (2856)	[2527, 3185]	7.68 (0.75)	[7.58, 7.79]
	Silence	70.33 (45.77)	[64.90, 75.76]	2751 (2751)	[2457, 3046]	7.69 (0.68)	[7.59, 7.78]
	Tone	73.99 (43.95)	[68.78, 79.21]	2891 (2891)	[2550, 3233]	7.70 (0.71)	[7.60, 7.80]
Backward translation, T1	Congruent	69.60 (46.08)	[64.13, 75.06]	2439 (2024)	[2134, 2743]	7.52 (0.73)	[7.41, 7.63]
	Silence	58.97 (49.28)	[53.13, 64.82]	2363 (1953)	[2044, 2682]	7.52 (0.70)	[7.41, 7.63]
	Tone	61.90 (48.65)	[56.13, 67.68]	2770 (2536)	[2370, 3169]	7.65 (0.70)	[7.54, 7.76]
Backward translation, T2	Congruent	84.25 (36.49)	[79.92, 88.58]	2184 (1822)	[1942, 2425]	7.43 (0.69)	[7.34, 7.52]
	Silence	84.25 (36.49)	[79.92, 88.58]	2160 (1764)	[1928, 2393]	7.44 (0.66)	[7.36, 7.53]
	Tone	87.55 (33.08)	[83.62, 91.47]	2295 (2481)	[1976, 2615]	7.43 (0.73)	[7.33, 7.52]
Picture naming	Congruent	76.88 (42.20)	[73.29, 80.47]	2113 (1819)	[1930, 2296]	7.38 (0.72)	[7.31, 7.45]
	Silence	73.31 (44.28)	[69.54, 77.07]	2320 (2364)	[2079, 2562]	7.42 (0.77)	[7.34, 7.50]
	Tone	78.57 (41.07)	[75.08, 82.06]	2199 (2074)	[1994, 2404]	7.37 (0.78)	[7.29, 7.45]
Lexical decision	Congruent	94.82 (22.18)	[92.98, 96.66]	1029 (643)	[973, 1084]	6.82 (0.43)	[6.78, 6.86]
	Silence	93.04 (25.48)	[90.93, 95.15]	1049 (712)	[987, 1110]	6.84 (0.43)	[6.80, 6.87]
	Tone	94.11 (23.57)	[92.15, 96.06]	1016 (532)	[970, 1062]	6.83 (0.41)	[6.79, 6.86]

Note: Values are presented as means (*M*), with standard deviations (*SD*) in parentheses and 95% confidence intervals (CI) indicated in brackets with lower and upper limits. Accuracy values are reported as percentages. RT = response time (ms); logRT = natural logarithm of response time (ms); T1 = Timepoint 1; T2 = Timepoint 2.

exclude these participants in the indicated tasks. Task means are provided in Table 3.

3.1. Questionnaires

For the additional models performed, we found that neither LEX (all uncorrected *p*s > .02) nor SES (all uncorrected *p*s > .12) provided a significant effect on performance for any task or measure when *p*-correction was performed (FDR-corrected *p*s > .31).

3.2. Semantically mediated tasks

3.2.1. Forward translation task

Accuracy. Participants showed a general accuracy rate of $M = 59\%$, $SE = 1.21$. There was a main effect of Timepoint, (χ^2) (1) = 184.19, $p < .001$, with accuracy increasing from 44% ($SE = 1.72$) at Timepoint 1 to 74% ($SE = 1.54$) at Timepoint 2. There was also a main effect of Sound, (χ^2) (2) = 15.97, $p < .001$. The Timepoint × Sound interaction was not significant, (χ^2) (2) = 1.59, $p = .45$. The *post hoc* for Sound detected differences between the tone ($M = 61\%$, $SE = 2.08$) and silence conditions ($M = 54\%$, $SE = 2.12$) (OR = 1.68, 95% CI [1.13, 2.48], $p = .006$), between congruent ($M = 62\%$, $SE = 2.06$) and silence conditions (OR = 1.84, 95% CI [1.24, 2.74], $p < .001$), but not between tone and congruent conditions, $p = .84$. Given the absence of a Timepoint × Sound interaction, accuracy is shown collapsed across timepoints in Figure 2; however, descriptive statistics by timepoint are available in Table 3.

Reaction times. The percentage of data eliminated as outliers was 2.07%. There was a main effect of Timepoint, with participants responding faster at Timepoint 2 ($M_{log} = 7.69$, $SE_{log} = 0.03$) in relation to the first ($M_{log} = 7.80$, $SE_{log} = 0.04$), (χ^2) (1) = 16.11, $p < .001$. There was no Sound main effect, (χ^2) (2) = 1.48, $p = .48$ or Sound × Timepoint interaction, (χ^2) (2) = 3.32, $p = .19$.

3.2.2. Picture naming task

Accuracy. Participants achieved an overall accuracy of $M = 76\%$ ($SE = 1.07$). The model revealed a main effect of Sound, (χ^2) (2) = 7.15, $p = .028$. *Post hoc* tests indicated that participants were more accurate in the tone condition ($M = 79\%$, $SE = 1.78$) than in the silence condition ($M = 73\%$, $SE = 1.92$) (OR = 1.61, 95% CI [1.06, 2.45], $p = .022$), see Figure 2. No significant differences were found between the congruent sound ($M = 77\%$, $SE = 1.83$) and tone conditions, $p = .53$, or between the congruent and silence conditions, $p = .26$.

Reaction times. Outlier removal excluded 2.22% of the data. Results from our model showed that Sound had no effect, (χ^2) (2) = 1.02, $p = .60$.

3.3. Lexically mediated tasks

3.3.1. Backward translation task

Accuracy. Participants showed a general accuracy rate of $M = 74\%$, $SE = 1.08$. There was a main effect of Timepoint, (χ^2) (1) = 110.11, $p < .001$, with accuracy increasing from 63% ($SE = 1.68$) at Timepoint 1 to 85% ($SE = 1.24$) at Timepoint 2. There was also a main effect of Sound, (χ^2) (2) = 6.62, $p = .037$. The Timepoint × Sound interaction was significant, (χ^2) (2) = 6.94, $p = .031$. The *post hoc* of the interaction showed that differences were detected at Timepoint 1 between congruent ($M = 70\%$, $SE = 2.79$) and silence ($M = 59\%$, $SE = 2.98$) conditions (OR = 2.07, 95% CI [1.23, 3.49], $p = .003$), and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

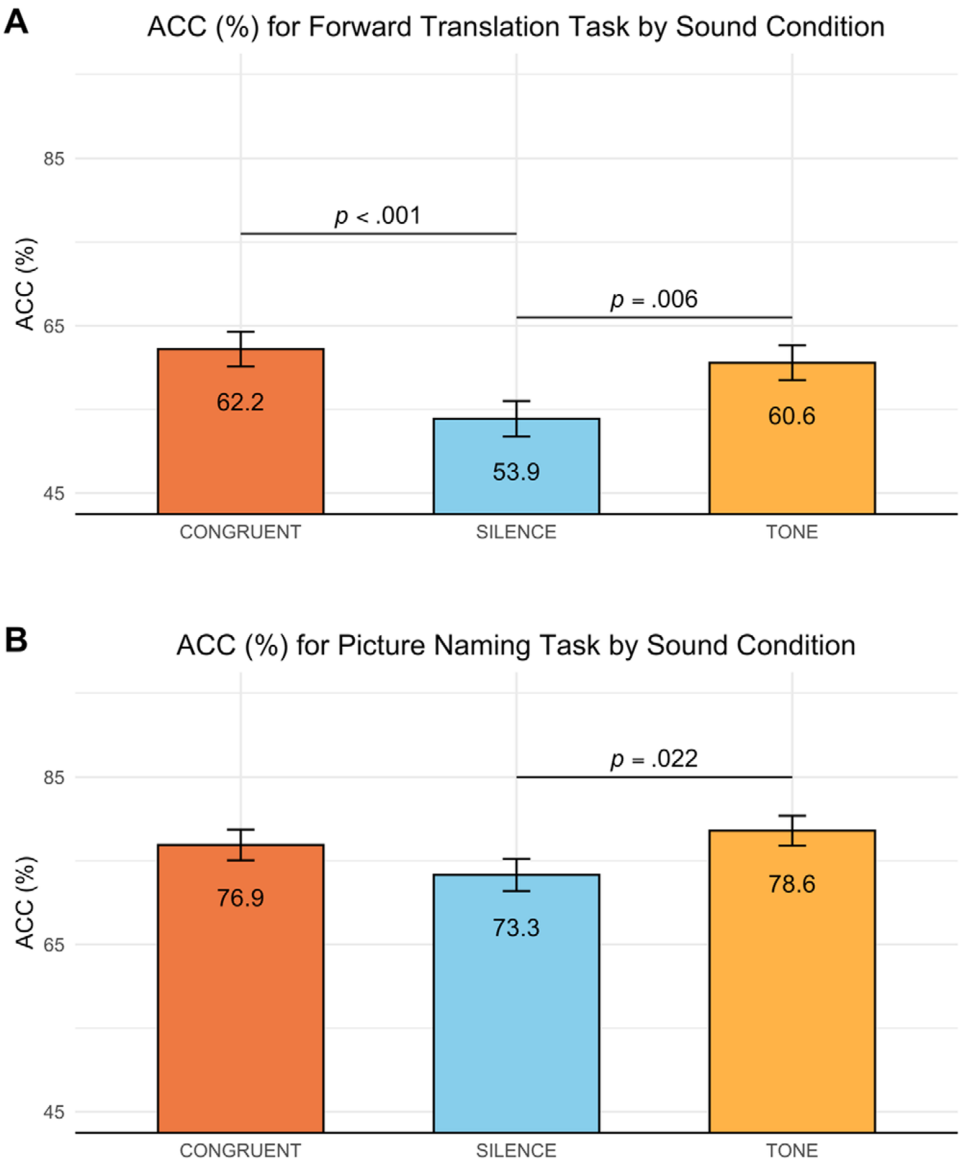


Figure 2. Observed mean accuracy (%) in both semantically mediated tasks: (A) Forward translation task and (B) Picture naming task by sound conditions (congruent sound, silence, neutral tone). Vertical lines indicate standard error, and mean accuracy is indicated for each condition.

between congruent and tone (M = 62%, SE = 2.94) conditions (OR = 1.69, 95% CI [1.00, 2.85], p = .049). The difference at Timepoint 1 between silence and tone conditions was not significant, p = .61, nor were any of the comparisons at Timepoint 2, ps > .35.

Reaction times. The percentage of data eliminated as outliers was 2.06%. There was a main effect of Timepoint, with participants responding faster at Timepoint 2 (M_{log} = 7.43, SE_{log} = 0.03) in relation to the first (M_{log} = 7.56, SE_{log} = 0.03), (χ^2) (1) = 38.01, p < .001. There was no Sound main effect, (χ^2) (2) = 1.55, p = .46 or Sound × Timepoint interaction, (χ^2) (2) = 4.49, p = .11.

3.3.2. Lexical decision task

Accuracy. Participants were highly accurate (M = 95%, SE = 0.37). Statistical analysis revealed that Sound had no effect, (χ^2) (2) = 1.83, p = .40, see Figure 3.

D′. The mean sensitivity score was M = 9.43 (SE = 0.49). All d' values were above 0, showing that participants' performance was better than chance. Higher d' values reflect greater sensitivity and ability to discriminate items. No significant main effect of Sound was found for this measure, (χ^2) (2) = 2.72, p = .26.

Reaction times. A small proportion of trials (2.34%) were excluded as outliers. Analysis showed that RTs were not influenced by Sound (χ^2) (2) = 0.57, p = .75.

4. Discussion

This study was the first to investigate dual-sensory input applied to the realm of FL learning by examining whether the combination of semantically meaningful environmental sounds and visual cues (pictures) could enhance recall. Previous literature has established the general utility of environmental sounds for enhancing learning (Heikkilä et al., 2015) and specifically, for FL acquisition (Bellegarda et al., 2026; Kaplan-Rakowski & Loranc-Paszylk, 2019), and it is well established that pictures are beneficial for FL learning (Altarriba & Knickerbocker, 2011; Bates & Son, 2020; Yu & Liu, 2022). Importantly, unlike previous studies that presented written linguistic cues alongside

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

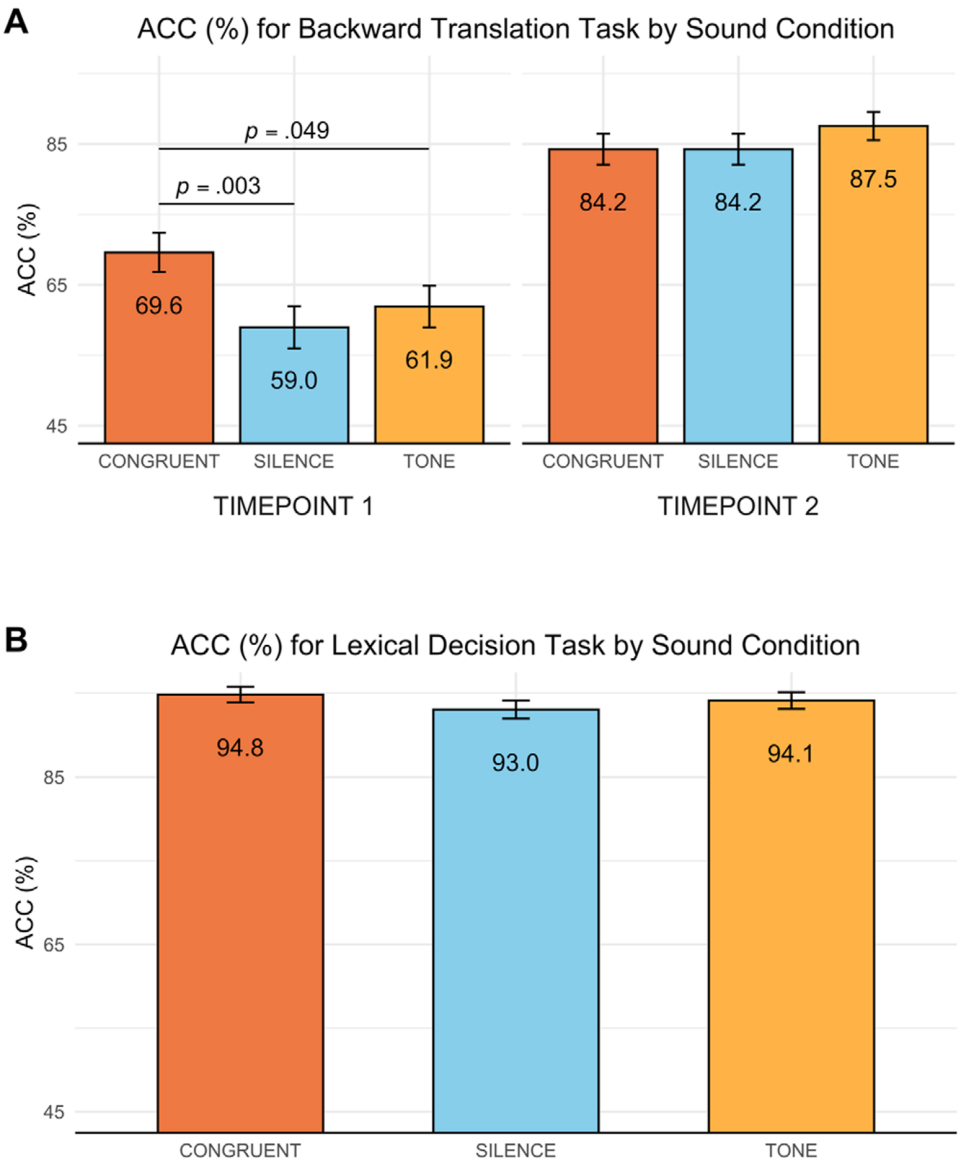


Figure 3. Observed mean accuracy (%) in both lexically mediated tasks: (A) Backward translation task and (B) Lexical decision task by sound conditions (congruent sound, silence, neutral tone). Vertical lines indicate standard error, and mean accuracy is indicated for each condition.

new words, our paradigm relied on conceptual activation of the target. This design aimed to strengthen the conceptual node and the FL-concept connection by engaging learners' semantic networks through sensory stimulation while minimizing L1 mediation. To assess how the amount and quality of sensory input influenced learning, we manipulated the number of sensory inputs (single visual input or dual visual/auditory input) and the characteristic of the auditory input. Participants were exposed to three auditory conditions: semantically meaningful environmental sound, neutral tone and silence. Across tasks, a coherent pattern emerged in accuracy performance depending on if tasks were semantically (forward translation and picture naming) or lexically (backward translation and lexical decision) mediated. Accuracy results for the forward translation task indicated that the presence of auditory input in general – regardless of semantic relevance – was associated with better learning relative to the silence condition. However, the environmental sounds did not appear to provide additional benefits beyond the presence of sound alone. In contrast, lexically mediated tasks showed no difference between sound conditions once learning had taken place. RTs showed no significant differences across conditions. This pattern may reflect the intensive repetition during the learning phase, which could have reduced variability in retrieval speed across conditions. This also highlights accuracy as the most reliable indicator of learning in our paradigm. The implications of the findings, and in particular a possible redundancy due to the dual-sensory input, are discussed below.

Contrary to our initial hypothesis, which predicted that two channels of semantically relevant input (e.g., pictures and environmental sounds) would enhance semantic processing and lead to superior learning outcomes, the environmental sound condition did not outperform the other neutral sound condition. Instead, both types of sounds yielded similar performance during testing and outperformed the silence condition. This pattern can be understood through several complementary, though not mutually exclusive, mechanisms. One possibility is that the presence of auditory

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

input in general, independent of semantic content or task relevance, may have enriched the encoding process and supported the formation of more elaborate semantic memory representations. This would be consistent with the Dual Coding Theory, which posits that an additional sensory trace enhances the probability of later correct retrieval (Paivio, 1969). In addition, both types of sounds may have increased participants' involvement or engagement in the task, in line with the self-involvement theory (Helstrup, 1987). Consistent with our pattern, Bellegarda et al. (2026) also found that both environmental sounds and neutral tones performed similarly in an FL learning paradigm. Thus, even though the present study enriched the sensory trace through two modalities (visual and auditory) instead of one (auditory only, Bellegarda et al., 2026), this enrichment did not yield additional semantic benefits for the congruent condition. Instead, both studies converge on the finding that semantic relevance of the auditory input did not affect the learning outcomes, suggesting that the mere presence of sound, rather than its informational content, may be the critical factor shaping encoding.

A second explanation derives from the redundancy or semantic overload effect. In multisensory learning, redundancy occurs when multiple modalities provide overlapping information, diminishing the additive value of each, while semantic overload may occur when multiple inputs activate conceptual information simultaneously, exceeding optimal processing capacity (Sweller, 1994). According to Multimedia Learning Theory (Mayer, 2012), additional information through separate input channels does not necessarily enhance learning unless it contributes uniquely to the representation. In these situations, learners may inefficiently allocate cognitive resources to processing the duplicated information across different inputs. In the present design, methodological steps were taken to reduce redundancy by using a sequential presentation, thereby minimizing disruption to orthographic encoding. At the same time, the materials were designed to maintain coherence, ensuring that auditory and visual information were not extraneous or distracting. Nevertheless, during the simultaneous multisensory presentation, because pictures already provided rich conceptual access, additional semantic information from environmental sounds may have offered limited incremental benefit. Under these conditions, both environmental sounds and neutral tones may have supported learning to a similar extent, as neither provided uniquely diagnostic information beyond what was already available from visual input to foster further integration of the FL word into the mental conceptual model created.

This interpretation would align with evidence suggesting that visual stimuli may present certain advantages for conceptual activation compared to environmental sounds. Pictures offer immediate perceptual correspondence through structural and idiosyncratic features (i.e., shape, spatial configuration), allowing fast access to stored conceptual knowledge (Paivio, 1969). This view is further supported by evidence from picture-word interference paradigms showing strong semantic involvement in picture naming (Bajo et al., 2003; Macizo & Bajo, 2004). In contrast, environmental sounds may provide less specific conceptual information and can be more ambiguous with respect to their source. For example, a picture of a horse directly specifies the concept "horse," whereas a whinny requires listeners to infer the sound source before retrieving the associated concept. So, although environmental sounds can activate semantic representations, pictures may provide more precise and diagnostically rich conceptual cues. Altogether, the pattern in our data can be interpreted as the conceptual activation being largely driven during the learning phase through the visual modality. Nevertheless, the presence of both environmental sound and neutral tone conditions still yielded an additional benefit, likely through general engagement-related mechanisms.

From a broader theoretical perspective, these findings align with the BIA-d framework (Grainger et al., 2010), which proposes that early FL vocabulary learning is driven by repeated coactivation of the FL form and its conceptual representation. In our paradigm, the picture was intended to activate the conceptual node, and the presence of sound (regardless of semantic content) likely boosted this activation by increasing engagement. Strong conceptual activation at the moment the FL word appeared would therefore strengthen the emerging FL-concept link, accounting for the advantage of any sound over silence. At the same time, the absence of an advantage for environmental sounds over neutral tones is consistent with the possibility of a ceiling or boundary condition on conceptual activation when visual cues are already highly informative: once the concept was fully activated by the picture, the possibility of congruent sounds providing additional semantic value (viz., conceptual activation) was limited.

Given the BIA-d assumption that conceptual support primarily benefits meaning-based tasks, the absence of sound effects in the two lexically mediated tasks is not surprising. Visual lexical decision relies largely on word form recognition and does not require accessing the conceptual system, making it less sensitive to the conceptual boosting that occurred during learning. Similarly, backward translation requires mapping newly learned FL forms onto highly dominant L1 lexical representations, such that performance does not hinge on explicit semantic access (Kroll & Stewart, 1994). These processing characteristics also make both tasks less demanding compared to the semantically mediated tasks that required active item retrieval, form-to-meaning mapping and the resolution of conflict arising from competing L1 representations, which are dominant and well-established in the lexicon. In the case of lexical decision in particular, the task required only a binary keypress (indicating whether a word was learned or new). Consistent with this, overall accuracy in both lexically mediated tasks approached ceiling (95% in lexical decision and 85% in backward translation at the second timepoint), suggesting that participants were able to recognize or produce almost all studied words regardless of the auditory condition; this ceiling effect likely obscured any subtle differences that might otherwise have emerged. Indeed, even participants who struggled in the more demanding meaning-based tasks performed well above chance in lexical decision.

An interesting observation is that in backward translation at the first timepoint, congruent sounds provided a modest advantage over silence and neutral tones, suggesting that early in learning, auditory semantic cues may transiently facilitate FL form-meaning mapping. By the second timepoint, however, this advantage disappeared. This pattern aligns with behavioral and neurophysiological evidence showing that even very novice learners can access semantic representations of newly learned FL words (Comesaña et al., 2009; García-Gámez & Macizo, 2022; Pu et al., 2016). As learning progressed and the FL forms became more established, participants appeared to rely primarily on direct FL–L1 lexical connections to perform lexically mediated tasks, reducing the contribution of explicit semantic mediation.

An important note of comparison is that Bellegarda et al. (2026) reported reduced performance from both congruent and neutral sounds in their lexical decision task, whereas in the present study we observed a null effect. Part of this difference may reflect methodological variations: Bellegarda et al. used more items and provided written Spanish forms, while our design relied exclusively on

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

conceptual activation through pictures, which may have promoted especially stable memory traces (Comesaña et al., 2009). This divergence can also be interpreted in light of the different sensory configurations: their study presented sounds alone, without visual input, and the observed decrease in performance was attributed to increased cognitive load hindering lexical access. In contrast, in the present study, beneficial effects from pictures and potential difficulty introduced by concurrent sounds may have acted in opposite directions, effectively canceling each other out and producing the null effect observed here. These findings reinforce that multisensory contributions are not inherently additive: its effects can interact nonlinearly, particularly in tasks driven by rapid form-based processing such as lexical decision and backward translation.

It is also interesting to understand how previous LEX or SES could impact FL vocabulary acquisition. Although both have been linked to differences in linguistic development, our null result in our exploratory analysis suggests that their influence may be limited in the context of short-term FL vocabulary learning. One possibility is that the associative nature of the task relied primarily on domain-general memory mechanisms rather than on preexisting linguistic systems or linguistic transfer processes. Additionally, multilingualism effects often appear when learning involves interference between known and new languages (e.g., cognates, grammatical overlap). In this experiment, participants learned isolated FL nouns and there were neither cognates nor grammatical elements. Regarding SES, given that our sample primarily consisted of university students, this lack of variability may have attenuated differences. Future research may probe these variables under conditions involving greater linguistic interference and greater SES variability.

4.1. Future directions

Environmental sounds and neutral tones differ acoustically and have been proposed to engage partially different processing routes. In acoustic terms, neutral (or pure) tones are stimuli with predictable spectral and temporal structures. In contrast, environmental sounds are acoustically dense signals. They feature rapid, asymmetric modulations and an intrinsic statistical structure that leads to a different processing route compared to neutral tones (Theunissen & Elie, 2014). However, we did not find behavioral evidence for this distinction. This finding is informative in itself, as it suggests that the contribution of auditory input to FL learning may depend on a specific boundary condition that remains to be specified in future research.

Additionally, to better understand the processing of environmental sounds vs. neutral sounds, neuroimaging techniques could clarify how these cues recruit semantic networks differently at a neural level, even when behavioral performance may appear unchanged. Word meaning is not stored in a single, encapsulated module, but is instead supported by distributed neuronal assemblies spanning multiple cortical regions (Pulvermüller, 2005). From this perspective, environmental sounds may function as sensory "anchors" that facilitate the automatic recruitment of topographically specific semantic networks. Neurophysiological methods such as electroencephalography (EEG) or magnetoencephalography (MEG) are particularly well suited to testing this possibility. By examining the spatiotemporal dynamics of brain activity, future studies could determine whether environmental sounds promote stronger grounding of lexical representations in action and perception systems compared to neutral cues, potentially leading to more durable memory traces through the engagement of distributed neurocognitive networks.

Another important direction for future work concerns consolidation over time. Because our study tested participants' learning on the same day, it remains possible that the benefits of environmental sounds may emerge more clearly over time. Semantically richer encoding could produce more durable memory traces, which have greater resistance to forgetting. Indeed, Kaplan-Rakowski and Loranc-Paszylk (2019) observed sustained advantages for sound-enhanced words on a seven-day delayed recall test when compared to the silence control condition. Notably, this advantage was not present for their pronunciation or sound and pronunciation conditions. Future research incorporating delayed assessments would therefore be valuable for examining whether environmental sounds support longer-term retention.

Finally, expanding the lexical scope of training materials to include a greater diversity of lexical items (i.e., verbs) may further clarify how multisensory input supports vocabulary acquisition. Expanding the learning set could also help determine whether the effects generalize across a broader range of linguistic contexts. Verbs, in particular, are often more "embodied" than nouns, as they engage sensorimotor systems during learning and retrieval (Grisoni et al., 2016, 2019).

5. Conclusions

To conclude, this study examined whether relevant multisensory stimuli enhance adult FL vocabulary learning. The results showed that the presence of sound during learning (regardless of semantic relevance) led to improved performance compared to silence, particularly in tasks that required semantic mediation. Although environmental sounds did not provide an advantage over neutral tones for learning, the overall benefit of sound suggests that sensory engagement plays a meaningful role during early FL learning. Importantly, the absence of sound effects in lexically mediated tasks highlights that multisensory contributions are not inherently additive: combined influences may interact in task-dependent ways, sometimes resulting in null behavioral effects. Rather than indicating that environmental sounds lack pedagogical value, our results point to a possible boundary condition under which multisensory enrichment is most effective. When visual input already provides strong conceptual activation, additional semantic cues may yield diminishing returns. Understanding when auditory meaning contributes uniquely and when it becomes redundant remains an important direction for future research.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926101692.

Data availability statement. The data, analyses, study materials and supplementary files supporting the findings of this research are openly available in the Open Science Framework (OSF) repository at: https://doi.org/10.17605/OSF.IO/6QAZ2

Acknowledgments. This research was supported by the Spanish Ministry of Science and Innovation through an FPU University Teacher Training grant (grant number FPU19/04616) awarded to Melodie Bellegarda and the project (reference number PID2019-111359GB-I00) awarded to Pedro Macizo. Additional funding was provided to the Mind, Brain and Behavior Research Center (CIMCYC), University of Granada through CEX2023-001312-M/MICIU/AEI/10.13039/501100011033 and UCE-PP2023-11/UGR. Funding for open access charge: University of Granada / CBUA. We are grateful to Daniel Navarrete Sancho for his valuable assistance in data collection. All procedures performed in this study involving human participants were in accordance with the ethical standards of the research ethical committee at the University of Granada (number issued by the Ethical Committee: 957/CEIH/2019) as well as APA ethical standards. We are grateful to two anonymous reviewers for their help in the review process.

Competing interests. The authors declare none.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

References

Abo Hamza, E., **Tindle, R.**, **Pawlak, S.**, **Bedewy, D.**, & **Moustafa, A. A.** (2024). The impact of poverty and socioeconomic status on brain, behaviour, and development: A unified framework. *Reviews in the Neurosciences*, **35**(6), 597–617. https://doi.org/10.1515/revneuro-2023-0163.

Adler, N. E., **Epel, E. S.**, **Castellazzo, G.**, & **Ickovics, J. R.** (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy white women. *Health Psychology*, **19**(6), 586–592. https://doi.org/10.1037/0278-6133.19.6.586.

Adsera, A., & **Pytlikova, M.** (2015). The role of language in shaping international migration. *The Economic Journal*, **125**(586), F49–F81. https://doi.org/10.1111/ecoj.12231.

Altarriba, J., & **Knickerbocker, H.** (2011). Chapter 2. Acquiring second language vocabulary through the use of images and words. In *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics* (pp. 21–48). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.30.06alt.

Bajo, M. T., **Puerta-Melguizo, M. C.**, & **Macizo, P.** (2003). The locus of semantic interference in picture naming. *Psicológica*, **24**(1), 31–55.

Ballas, J. A., & **Howard, J. H.** (1987). Interpreting the language of environmental sounds. *Environment and Behavior*, **19**(1), 91–114. https://doi.org/10.1177/0013916587191005.

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

Bates, D., **Kliegl, R.**, **Vasishth, S.**, & **Baayen, H.** (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., **Maechler, M.**, **Bolker, B.**, & **Walker, S.** (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1–10. http://CRAN.R-project.org/package=lme4.

Bates, J., & **Son, J. B.** (2020). English vocabulary learning with simplified pictures. *TESL-EJ: The Electronic Journal for English as a Second Language*, **24**(3) 1–20. http://www.tesl-ej.org/wordpress/issues/volume24/ej95/ej95a12

Beauchamp, M. S., **Lee, K. E.**, **Argall, B. D.**, & **Martin, A.** (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, **41**(5), 809–823. https://doi.org/10.1016/s0896-6273(04)00070-4.

Bellegarda, M., **García-Gámez, A. B.**, & **Macizo, P.** (2026). The impact of relevant environmental sounds on foreign language word learning. *Bilingualism: Language and Cognition*, **29**(3), 773–783. https://doi.org/10.1017/S1366728925100394.

Bellegarda, M., & **Macizo, P.** (2021). Cognitive control and bilingualism: The bilingual advantage through the lens of dimensional overlap. *Frontiers in Psychology*, **12**, 614849. https://doi.org/10.3389/fpsyg.2021.614849.

Benjamini, Y., & **Hochberg, Y.** (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300. http://www.jstor.org/stable/2346101; https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boersma, P., & **Weenink, D** (2020). Praat: Doing phonetics by computer [Computer program]. Version 6.1.33. http://www.praat.org/

Bonnet, A., & **Siemund, P.** (2018). Introduction. In *Foreign language education in multilingual classrooms* (pp. 1–29). John Benjamins Publishing Company. https://doi.org/10.1075/hsld.7.01bon.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, **14**(1), 177–208. https://doi.org/10.1080/026432997381664.

Cenoz, J. (2013). The influence of bilingualism on third language acquisition: Focus on multilingualism. *Language Teaching*, **46**(1), 71–86. https://doi.org/10.1017/S0261444811000218.

Chiswick, B. R., & **Miller, P. W.** (2002). Immigrant earnings: Language skills, linguistic concentrations and the business cycle. *Journal of Population Economics*, **15**(1), 31–57. https://doi.org/10.1007/PL00003838.

Comesaña, M., **Perea, M.**, **Piñeiro, A.**, & **Fraga, I.** (2009). Vocabulary teaching strategies and conceptual representations of words in L2 in children: Evidence with novice learners. *Journal of Experimental Child Psychology*, **104**(1), 22–33. https://doi.org/10.1016/j.jecp.2008.10.004.

Cuetos, F., **González-Nosti, M.**, **Barbón, A.**, & **Brysbaert, M.** (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, **32**, 133–143.

Cummings, A., **Ceponiene, R.**, **Dick, F.**, **Saygin, A. P.**, & **Townsend, J.** (2008). A developmental ERP study of verbal and non-verbal semantic processing. *Brain Research*, **1208**, 137–149. https://doi.org/10.1016/j.brainres.2008.02.015.

Cummings, A., **Ceponiene, R.**, **Koyama, A.**, **Saygin, A. P.**, **Townsend, J.**, & **Dick, F.** (2006). Auditory semantic networks for words and natural sounds. *Brain Research*, **1115**(1), 92–107. https://doi.org/10.1016/j.brainres.2006.07.050.

de Groot, A. M., & **van Hell, J. G.** (2005). *The learning of foreign language vocabulary* (pp. 9–29). Oxford University Press.

Dick, F., **Saygin, A. P.**, **Galati, G.**, **Pitzalis, S.**, **Bentrovato, S.**, **D'Amico, S.**, **Wilson, S.**, **Bates, E.**, & **Pizzamiglio, L.** (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: Evidence from functional magnetic resonance imaging and lesion data. *Journal of Cognitive Neuroscience*, **19**(5), 799–816. https://doi.org/10.1162/jocn.2007.19.5.799.

Dijkstra, T., & **Van Heuven, W. J.** (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, **5**(3), 175–197. https://doi.org/10.1017/S1366728902003012.

Dittinger, E., **Barbaroux, M.**, **D'Imperio, M.**, **Jäncke, L.**, **Elmer, S.**, & **Besson, M.** (2016). Professional music training and novel word learning: From faster semantic encoding to longer-lasting word representations. *Journal of Cognitive Neuroscience*, **28**(10), 1584–1602. https://doi.org/10.1162/jocn_a_00997.

Dong, L. (2024). Examining the relationship between socioeconomic status, self-regulated learning strategies, and writing proficiency in English as a second language learning context. *Journal of Educational Psychology*, **116**(5), 686–702. https://doi.org/10.1037/edu0000834.

Duchon, A., **Perea, M.**, **Sebastián-Gallés, N.**, **Martí, A.**, & **Carreiras, M.** (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, **45**(4), 1246–1258. https://doi.org/10.3758/s13428-013-0326-1.

Duñabeitia, J. A., **Crepaldi, D.**, **Meyer, A. S.**, **New, B.**, **Pliatsikas, C.**, **Smolka, E.**, & **Brysbaert, M.** (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, **71**(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261.

Eitel, A., **Scheiter, K.**, & **Schüler, A.** (2013). How inspecting a picture affects processing of text in multimedia learning. *Applied Cognitive Psychology*, **27**(4), 451–461. https://doi.org/10.1002/acp.2922.

Ellis, N. C. (2019). Essentials of a theory of language cognition. *The Modern Language Journal*, **103**(S1), 39–60. https://doi.org/10.1111/modl.12532.

Faul, F., **Erdfelder, E.**, **Buchner, A.**, & **Lang, A. G.** (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, **41**(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Fieles-Ahmad, O. M., & **Huber, M.** (2022). Learn German, buy German? Language-learning opportunities abroad and exports. *The World Economy*, **45**(10), 3031–3058. https://doi.org/10.1111/twec.13280.

Frey, A., **Aramaki, M.**, & **Besson, M.** (2014). Conceptual priming for realistic auditory scenes and for auditory words. *Brain and Cognition*, **84**(1), 141–152. https://doi.org/10.1016/j.bandc.2013.11.013.

García-Gámez, A. B., **Cervilla, Ó.**, **Casado, A.**, & **Macizo, P.** (2024). Seeing or acting? The effect of performing gestures on foreign language vocabulary learning. *Language Teaching Research*, **28**(3), 1055–1086. https://doi.org/10.1177/13621688211024364.

García-Gámez, A. B., & **Macizo, P.** (2019). Learning nouns and verbs in a foreign language: The role of gestures. *Applied Psycholinguistics*, **40**(2), 473–507. https://doi.org/10.1017/S0142716418000656.

García-Gámez, A. B., & **Macizo, P.** (2020). The way in which foreign words are learned determines their use in isolation and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **46**(2), 364–379. https://doi.org/10.1037/xlm0000721.

García-Gámez, A. B., & **Macizo, P.** (2022). Lexical and semantic training to acquire words in a foreign language: An electrophysiological study. *Bilingualism: Language and Cognition*, **25**(5), 768–785. https://doi.org/10.1017/S1366728921000456.

García-Gámez, A. B., & **Macizo, P.** (2023). Gestures as scaffolding to learn vocabulary in a foreign language. *Brain Sciences*, **13**(12), 1712. https://doi.org/10.3390/brainsci13121712.

Glaser, W. R., & **Glaser, M. O.** (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General*, **118**(1), 13–42. https://doi.org/10.1037/0096-3445.118.1.13.

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, **11**(6), 559–568. https://doi.org/10.1080/016909696386944.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

Grainger, J., & **Dijkstra, T.** (1992). On the representation and use of language information in bilinguals. In R. J. Harris (Ed.), *Advances in psychology* (Vol. **83**, pp. 207–220). https://doi.org/10.1016/S0166-4115(08)61496-X.

Grainger, J., **Midgley, K.**, & **Holcomb, P. J.** (2010). Chapter 14. Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). In M. Kail & M. Hickmann (Eds.), *Language acquisition across linguistic and cognitive systems* (Vol. **52**, pp. 267–283). John Benjamins Publishing. https://doi.org/10.1075/lald.52.18gra.

Grisoni, L., **Dreyer, F. R.**, & **Pulvermüller, F.** (2016). Somatotopic semantic priming and prediction in the motor system. *Cerebral Cortex*, **26**(5), 2353–2366. https://doi.org/10.1093/cercor/bhw026.

Grisoni, L., **Mohr, B.**, & **Pulvermüller, F.** (2019). Prediction mechanisms in motor and auditory areas and their role in sound perception and language understanding. *NeuroImage*, **199**, 206–216. https://doi.org/10.1016/j.neuroimage.2019.05.071.

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press. https://doi.org/10.4324/9781410612403.

Heikkilä, J., **Alho, K.**, **Hyvönen, H.**, & **Tiippana, K.** (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology*, **62**(2), 123–130. https://doi.org/10.1027/1618-3169/a000279.

Helstrup, T. (1987). One, two, or three memories? A problem-solving approach to memory for performed acts. *Acta Psychologica*, **66**(1), 37–68. https://doi.org/10.1016/0001-6918(87)90017-5.

Hirosh, Z., & **Degani, T.** (2018). Direct and indirect effects of multilingualism on novel language learning: An integrative review. *Psychonomic Bulletin Review*, **25**(3), 892–916. https://doi.org/10.3758/s13423-017-1315-7.

Honda, Y., **Tanizawa, K.**, **Masai, K.**, **Nakamura, Y.**, **Choi, H.**, & **Fukushima, S.** (2025). Generating sensory vocabulary cues for describing scents in olfactory-enhanced second language vocabulary learning. In *International conference on human-computer interaction* (pp. 167–184). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93746-0_13.

Jusezyk, P. W., & **Luce, P. A.** (2002). Speech perception and spoken word recognition: Past and present. *Ear and Hearing*, **23**(1), 2–40. https://doi.org/10.1097/00003446-200202000-00002.

Kaiser, A., **Eppenberger, L. S.**, **Smieskova, R.**, **Borgwardt, S.**, **Kuenzli, E.**, **Radue, E. W.**, **Nitsch, C.**, & **Bendfeldt, K.** (2015). Age of second language acquisition in multilinguals has an impact on gray matter volume in language-associated brain areas. *Frontiers in Psychology*, **6**, 638. https://doi.org/10.3389/fpsyg.2015.00638.

Kaplan-Rakowski, R., & **Loranc-Paszylk, B.** (2019). The impact of verbal and nonverbal auditory resources on explicit foreign language vocabulary learning. *System*, **85**, 102114. https://doi.org/10.1016/j.system.2019.102114.

Kaushanskaya, M., & **Marian, V.** (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, **16**(4), 705–710. https://doi.org/10.3758/PBR.16.4.705.

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.

Kumle, L., **Võ, M. L.**, & **Draschkow, D.** (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, **53**(6), 2528-2543. https://doi.org/10.3758/s13428-021-01546-0

Leach, L., & **Samuel, A.** (2007). Lexical configuration and lexical engagement: When adults learn new words. *Journal of Cognitive Psychology*, **55**(4), 306–353. https://doi.org/10.1016/j.cogpsych.2007.01.001.

Lenth, R. (2023). *Emmeans: Estimated marginal means, aka least-squares means*. R package version 1.8.5. https://CRAN.R-project.org/package=emmeans.

Li, J., & **Deng, S. W.** (2023). Facilitation and interference effects of the multisensory context on learning: A systematic review and meta-analysis. *Psychological Research*, **87**(5), 1334–1352. https://doi.org/10.1007/s00426-022-01733-4.

Lindsay, S., & **Gaskell, M. G.** (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **39**(2), 608–622. https://doi.org/10.1037/a0029243.

Macizo, P., & **Bajo, M. T.** (2004). Semantic facilitation and lexical competition in picture naming. *Psicologica: International Journal of Methodology and Experimental Psychology*, **25**(1), 1–22.

MacMillan, N. A., & **Creelman, C. D.** (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Marian, V., **Blumenfeld, H. K.**, & **Kaushanskaya, M.** (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research*, **50**(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067).

Martínez, N., **Matute, H.**, & **Goikoetxea, E.** (2020). PicPsy: A new bank of 106 photographs and line drawings with written naming norms for Spanish-speaking children and adults. *PLoS one*, **15**(9), e0238976. https://doi.org/10.1371/journal.pone.0238976.

Mayer, R. E. (2012). *Multimedia learning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511811678.

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, **36**(1), 8. https://doi.org/10.1007/s10648-023-09842-1.

Mayer, R. E., & **Johnson, C. I.** (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology*, **100**(2), 380–386. https://doi.org/10.1037/0022-0663.100.2.380.

Mestres-Missé, A., **Rodriguez-Fornells, A.**, & **Münte, T. F.** (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, **17**(8), 1858–1866. https://doi.org/10.1093/cercor/bhl094.

Orgs, G., **Lange, K.**, **Dombrowski, J.**, & **Heil, M.** (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition*, **62**(3), 267–272. https://doi.org/10.1016/j.bandc.2006.05.003.

Orgs, G., **Lange, K.**, **Dombrowski, J.**, & **Heil, M.** (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology*, **65**(2), 162–166. https://doi.org/10.1016/j.ijpsycho.2007.03.003.

Orgs, G., **Lange, K.**, **Dombrowski, J. H.**, & **Heil, M.** (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. *Neuroscience Letters*, **436**(2), 133–137. https://doi.org/10.1016/j.neulet.2008.03.005.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, **76**(3), 241–263. https://doi.org/10.1037/h0027272.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, **61**(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x.

Pu, H., **Holcomb, P. J.**, & **Midgley, K. J.** (2016). Neural changes underlying early stages of L2 vocabulary acquisition. *Journal of Neurolinguistics*, **40**, 55–65. https://doi.org/10.1016/j.jneuroling.2016.05.002.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, **6**(7), 576–582. https://doi.org/10.1038/nrn1706.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rahmanu, I. W. E. D., & **Molnár, G.** (2024). Multimodal immersion in English language learning in higher education: A systematic review. *Heliyon*, **10**(19), e38357. https://doi.org/10.1016/j.heliyon.2024.e38357.

Rasouli, F., & **Jafari, K.** (2016). A deeper understanding of L2 vocabulary learning and teaching: A review study. *International Journal of Language and Linguistics*, **4**(1), 40–46. https://doi.org/10.11648/j.ijll.20160401.16.

Saygin, A. P., **Dick, F.**, & **Bates, E.** (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods*, **37**(1), 99–110. https://doi.org/10.3758/bf03206403.

Saygin, A. P., **Dick, F.**, **Wilson, S. W.**, **Dronkers, N. F.**, & **Bates, E.** (2003). Neural resources for processing language and environmental sounds: Evidence from aphasia. *Brain: A Journal of Neurology*, **126**(4), 928–945. https://doi.org/10.1093/brain/awg082.

Schneider, W., **Eschman, A.**, & **Zuccolotto, A.** (2012). *E-prime user's guide*. Psychology Software Tools, Inc.

Shams, L., & **Seitz, A.** (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, **12**(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006.

Sholl, A., **Sankaranarayanan, A.**, & **Kroll, J. F.** (1995). Transfer between picture naming and translation: A test of asymmetries in bilingual memory. *Psychological Science*, **6**(1), 45–49. https://doi.org/10.1111/j.1467-9280.1995.tb00303.x.

Smith, L., & **Gasser, M.** (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, **11**(1–2), 13–29. https://doi.org/10.1162/1064546053278973.

Sprenger, E. (2024). What makes us move, what makes us stay: The role of language and culture in intra-EU mobility. *Journal of International Migration*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692

and Integration, **25**(4), 1825–1855. https://doi.org/10.1007/s12134-024-01134-z.

Stöhr, T. (2015). The returns to occupational foreign language use: Evidence from Germany. *Labour Economics*, **32**, 86–98. https://doi.org/10.1016/j.labeco.2015.01.004.

Strauß, A., **Wu, T.**, **McQueen, J. M.**, **Scharenborg, O.**, & **Hintz, F.** (2022). The differential roles of lexical and sublexical processing during spoken-word recognition in clear and in noise. *Cortex*, **151**, 70–88. https://doi.org/10.1016/j.cortex.2022.02.011.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, **18**(6), 643–662. http://doi.org/10.1037/h0054651.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, **4**(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5.

Sweller, N., **Shinooka-Phelan, A.**, & **Austin, E.** (2020). The effects of observing and producing gestures on Japanese word learning. *Acta Psychologica*, **207**, 103079. https://doi.org/10.1016/j.actpsy.2020.103079.

Theunissen, F. E., & **Elie, J. E.** (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, **15**(6), 355–366. https://doi.org/10.1038/nrn3731.

Tukey, J. W. (1953). The problem of multiple comparisons. *Technometrics*, **5**(2), 99–114. https://doi.org/10.1080/00401706.1953.10405634.

Van Petten, C., & **Rheinfelder, H.** (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, **33**(4), 485–508. https://doi.org/10.1016/0028-3932(94)00133-A.

Vanderveer, N. J. (1979). *Ecological acoustics: Human perception of environmental sounds* (Order No. 8004002). [Doctoral Dissertation, Cornell University]. ProQuest Dissertations & Theses Global.

Vigliocco, G., **Vinson, D. P.**, **Damian, M. F.**, & **Levelt, W.** (2002). Semantic distance effects on object and action naming. *Cognition*, **85**(3), B61–B69. https://doi.org/10.1016/S0010-0277(02)00107-5.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, **17**(6), 487–499. https://doi.org/10.1080/0269920031000107541.

Xia, L., **Qin, Y.**, & **Fan, J.** (2024). Amplifying language learning effects with olfactory-enhanced virtual reality: An empirical study. In *International conference on human-computer interaction* (pp. 190–204). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61691-4_13.

Yu, J., & **Liu, X.** (2022). Text first or picture first? Evaluating two modes of multimodal input for EFL vocabulary meaning acquisition. *SAGE Open*, **12**(3), 21582440221119469. https://doi.org/10.1177/21582440221119469.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 18:07:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101692