Beyond Form-Meaning: Investigating the Potential and Limits of Captioned Video in Building Declarative and Automatized Vocabulary Knowledge

Kazuya Saito[1,2]
Xinran Fan[1]
Ana Pellicer-Sánchez[1]
Takumi Uchihara[2]

[1]University College London
[2]Tohoku University

Correspondence concerning this article should be addressed to Kazuya Saito, University College London, 20 Bedford Way, London, WC1H0AL, United Kingdom Email: k.saito@ucl.ac.uk

Acknowledgement

This research was supported by a Leverhulme Trust Grant (RPG-2024-391) and a UK-ISPF Grant (1185702223), awarded to Kazuya Saito. We thank the anonymous reviewers and Editor María del Pilar García Mayo for their insightful comments on an earlier version of the manuscript.

Abstract

The present study tested the hypothesis that phonological vocabulary knowledge consists of declarative and automatized dimensions, which develop at different rates. Chinese English-as-a-Foreign-Language students completed three tasks (recognition, recall, and lexicosemantic judgments) to assess learning gains of 18 multiword expressions before and after short enhanced audiovisual training. The results showed that (a) task effects were minor at the outset due to participants' limited prior knowledge of target items, (b) training facilitated greater gains in the declarative than in the automatized dimension, and (c) clearer task effects (declarative > automatized knowledge) emerged following training. These findings provide longitudinal support for the developmental trajectories proposed by the declarative-automatized model of phonological vocabulary acquisition.

Key words: Vocabulary, multiword expressions, automatization, audiovisual training, listening

According to Nation's (2013) influential framework of second language (L2) word knowledge for successful listening comprehension, vocabulary competence involves not only the strength of form–meaning mapping (i.e., knowing what a word sounds like and what it means) but also learners' ability to access that knowledge in relation to surrounding words in context—i.e., the use aspect of L2 word knowledge. Building on the skill acquisition theoretical paradigm, recent cross-sectional studies have conceptualized and measured L2 vocabulary knowledge as two interrelated but distinct constructs: declarative and automatized knowledge (Saito et al., 2025; Uchihara et al., 2025). To extend this framework longitudinally, the present study focused on the acquisition of multiword units through a well-established method of L2 vocabulary training: *enhanced* audiovisual input. Previous research has shown that substantial vocabulary learning occurs through viewing, particularly when the learning experience is enhanced with captions (Montero Perez, 2022) and when learners are guided to attend to target items beforehand (Majuddin et al., 2024). In this study, we examined Chinese learners of English who engaged in such an enhanced audiovisual training programme and tested the following hypotheses:

1. Learners' declarative and automatized knowledge of the target items would be comparable *prior to* training.
2. Greater vocabulary learning gains would occur for declarative than for automatized knowledge *during* training.
3. A clearer hierarchical pattern (declarative > automatized) would emerge *after* training.

Automatization, by definition, develops gradually through extended and contextualized practice. Accordingly, the present study does not aim to demonstrate complete automatization but rather to examine whether the declarative and automatized dimensions—as conceptualized in skill acquisition theory—respond differently to short-term enhanced audiovisual training, with learning gains expected to be more evident in the declarative than in the automatized dimension.

Background Literature

Enhanced Audiovisual Training

Given the growing popularity of audiovisual materials, both researchers and practitioners have shown increasing interest in how learners acquire vocabulary while viewing video clips containing target lexical items (for an overview, see Montero Perez, 2022). Over the past decade, extensive empirical evidence has demonstrated that while mere exposure to videos can lead to incidental learning (Peters & Webb, 2018), learning outcomes improve substantially when pedagogical enhancements are introduced to help learners attend to target items during viewing—an approach termed *enhanced audiovisual training*.

One widely used enhancement involves captions. A large body of research has shown that learning gains are greater when videos are accompanied by captions (e.g., Peters, 2019; Teng, 2025; Winke et al., 2010; see Montero Perez et al., 2013, and Kurokawa et al., 2024, for meta-analyses). Kurokawa et al.'s (2024) meta-analysis revealed that captioned viewing led to an average relative gain of 19.70%, compared to 15.15% for uncaptioned viewing—a roughly 5% advantage. According to the subtitle principle (Mayer et al., 2020), this benefit arises because captions act as a redundant channel that aids word segmentation, recognition, and retention. This is particularly advantageous for L2 learners, who often struggle to parse continuous speech due to unclear word boundaries and rapid speech rates. By providing concurrent orthographic and auditory input, captions facilitate the mapping of phonological forms to orthographic representations and semantic meanings. This dual-modality input supports both bottom-up decoding and top-down comprehension, enhancing learners' ability to notice and integrate new lexical items (Vanderplank, 2016).

Empirical studies (e.g., Montero Perez et al., 2020; Teng, 2025) consistently report that captioned videos yield greater gains in meaning recognition and recall than uncaptioned videos, especially among beginner- and intermediate-level learners. Captioned input also tends to outperform other modalities such as listening-only, reading, or uncaptioned viewing, a finding attributed to dual-channel encoding and the ability of captions to direct learners' attention to key lexical targets. Moreover, captions can reduce comprehension difficulties and maintain engagement, particularly when learners encounter unfamiliar topics or accents (Montero Perez, 2022).

Another enhancement involves explicit vocabulary guidance prior to viewing. Many earlier studies emphasizing incidental learning have deliberately avoided announcing vocabulary tests beforehand, aiming to ensure that participants focused solely on message comprehension (e.g., Majuddin et al., 2021). However, for pedagogical purposes, other researchers have examined whether providing explicit vocabulary guidance before viewing can facilitate learning. Such guidance typically informs learners that vocabulary tests will follow, encouraging greater attention to lexical items during viewing (i.e., the notion of combined intentional and incidental learning; Laufer & Hulstijin, 2001). For example, Peters et al. (2009) found that announcing vocabulary tests significantly increased learning gains from reading activities. In the audiovisual domain, Majuddin et al. (2024) likewise demonstrated that explicit pre-viewing guidance substantially enhanced vocabulary learning from audiovisual training, even after a single exposure, whereas minimal gains were observed when the same training was delivered without such guidance (cf. Majuddin et al., 2021).

Building on these findings, the present study adopted this well-established enhanced/explicit audiovisual training method to investigate how learning gains emerge across two dimensions of word knowledge: declarative and automatized. Specifically, we tested the hypothesis that enhanced audiovisual training would lead to clear gains in declarative knowledge but more limited improvements in automatized knowledge.

Skill Acquisition Framework of L2 Phonological Vocabulary

Many scholars have highlighted that developing robust phonological vocabulary knowledge (including the knowledge of single-word items and multiword expressions) is a cornerstone of L2 speech proficiency. Indeed, extensive research has identified such knowledge as a primary contributor to the attainment of global listening (e.g., Cheng et al., 2023; McLean., 2015; Wallace, 2022) and speaking skills (e.g., De Jong et al., 2012; Takizawa et al., 2026; Tavakoli & Uchihara, 2020). Nation's (2013) influential framework suggests that phonological vocabulary knowledge can be characterized by two core aspects: *form-meaning mapping* and *use-in-context*. The former refers to the ability to associate the phonological form of a word with its meaning, while the latter entails deploying this knowledge in semantically, collocationally, and grammatically appropriate ways within broader linguistic contexts. Form-meaning mapping typically involves recognising individual target words (or phrases) in isolation, whereas use-in-context reflects the capacity to understand and integrate those words (or phrases) within larger clause- or sentence-level units. In this sense, L2 listening proficiency is generally assumed to progress from basic form–meaning mapping toward more advanced use-in-context skills.

In Schmitt's (2019) methodological synthesis of L2 vocabulary research, it was noted that the vast majority of studies have concentrated almost exclusively on form-meaning mapping—typically assessed through meaning recognition and recall tasks (for methodological foundations and distinctions between recognition and recall, see González-Fernández & Schmitt, 2020). However, Schmitt underscored that surprisingly few have directly examined the use-in-context dimension, despite its arguably greater ecological validity in capturing the lexical demands of real-world listening and speaking.

To advance theoretical and methodological understanding in this area, recent studies (Saito et al., 2025; Uchihara et al., 2025) have reconceptualized the two-stage development of phonological vocabulary knowledge—i.e., from form-meaning mapping to use-in-context—within the framework of L2 phonological vocabulary development. Traditionally applied to rule-based morphosyntactic learning in instructed contexts, skill acquisition theory distinguishes between declarative, procedural, and automatized knowledge (DeKeyser & Suzuki, 2025). According to this model, second language learning progresses sequentially from declarative knowledge ('knowing *what*'), to procedural knowledge ('knowing *how*' in controlled settings), and ultimately to automatized knowledge ('knowing how' in *real-life* and *global* contexts). Drawing on conceptual parallels between Nation's (2013) framework and skill acquisition theory, Saito et al. (2025) and Uchihara et al. (2025) aligned form–meaning mapping with declarative/procedural knowledge and use-in-context with automatized knowledge.

Under this view, form-meaning mapping involves the explicit association between a word's sound and meaning, typically assessed through recognition and recall tasks. At this stage, learners build declarative representations of lexical items through explicit, form-focused instruction, reflecting changes in *symbolic* representations of knowledge. By contrast, use-in-context corresponds to automatized knowledge, defined as the ability to access and integrate lexical items quickly, accurately, and consistently within broader sentential contexts. After developing declarative and procedural knowledge, repeated exposure to authentic L2 aural input (e.g., TV shows, films, conversations with native and non-native speakers) enables learners to integrate lexical items into surrounding discourse as cohesive units at the sentence level. This, in turn, leads to faster, more accurate, and more stable processing with reduced attentional demands (Ellis, 2006), reflecting qualitative improvements in *subsymbolic* representations of knowledge.

In his more recent work, DeKeyser has revised and expanded skill acquisition theory to encompass a wider range of instructed L2 learning domains, with the aim of offering a more nuanced account of L2 development in classroom settings and clearer pedagogical implications (DeKeyser & Suzuki, 2025). Building on this, Suzuki and DeKeyser (in press) have explicitly applied the framework to the development of phonological vocabulary knowledge in the context of real-life listening and speaking (cf. see Saito & Plonsky, 2019 for the application of skill acquisition theory to pronunciation teaching). Echoing Saito et al. (2025) and Uchihara et al. (2025), Suzuki and DeKeyser argue that "initial declarative knowledge, such as that called upon in form-meaning recognition and recall, must be automatized through repeated contextual use to become truly employable…Achieving automaticity enables accurate and efficient lexical processing needed to support fluent comprehension and production. This is not just about knowing what a word means, but being able to access and integrate that meaning effortlessly during real-time language use."

According to skill acquisition theory in instructed L2 learning (DeKeyser & Suzuki, 2025; Suzuki & DeKeyser, in press), a key distinction lies between the non-automatized (i.e., declarative and procedural) and automatized dimensions of knowledge. On the one hand, declarative knowledge is generally assessed through single-modal measures that allow learners to focus their attention on specific knowledge components under relatively controlled conditions. Within the domain of phonological vocabulary, this dimension has typically been operationalised through meaning recognition tasks (McLean et al., 2015) or meaning recall tasks (Cheng et al., 2023), both of which require learners to explicitly retrieve form-meaning associations.

On the other hand, automatized knowledge is examined using dual-modal tasks that test learners' ability to access target linguistic representations while simultaneously engaging in meaningful communication. Such tasks emulate real-life communicative contexts, where accurate and rapid use of language must occur largely subconsciously while multiple linguistic subsystems are activated in parallel. These conditions reveal the extent to which learners can deploy knowledge automatically in the service of fluent communication. Despite its theoretical and pedagogical importance, however, relatively little empirical work has addressed how best to capture the automatized dimension of phonological vocabulary knowledge (Suzuki & Elgort, 2025).

In applied linguistics, automatized L2 knowledge—defined as the accurate, prompt, and stable application of acquired linguistic information—is commonly evaluated through acceptability judgment tasks (Suzuki & Elgort, 2025, for a comprehensive overview). For example, automatized morphosyntactic knowledge is often assessed with grammaticality judgment tasks (GJTs; Plonsky et al., 2020), where L2 learners evaluate the grammatical correctness of aurally or visually presented sentences. Such tasks include both well-formed and error-containing sentences, thereby probing learners' morphosyntactic knowledge at the sentence level rather than in isolation.

Although relatively few studies have adopted acceptability judgment tasks to assess vocabulary, some pioneering work has done so. Ellis et al. (2008) investigated how L1 and L2 speakers judged the grammaticality of word sequences, contrasting well-formed expressions (e.g., *by the way*) with ill-formed ones (e.g., *by way the*). Their findings showed that L1 speakers' judgments were primarily driven by collocational strength, whereas L2 speakers relied more on word frequency. Similarly, Foster et al. (2014) asked advanced L2 learners to detect non-nativelike collocations embedded in narrative passages (e.g., "he replied *by a shrug*" instead of "he replied *with a shrug*"). Their results indicated that learners with higher proficiency and an earlier age of arrival were more accurate in identifying non-standard combinations. Notably, both studies presented stimuli in written form.

More recently, researchers have developed and validated the Lexicosemantic Judgment Task (LJT) as a measure of the automatized dimension of L2 phonological vocabulary knowledge (Saito et al., 2025; Uchihara et al., 2025). In an LJT, learners evaluate whether a target word is used appropriately in a given sentence—for example, a correct usage (e.g., *I work hard for promotion*) versus an incorrect one (e.g., *I ate a promotion last night*). Drawing on dual-task paradigms, the LJT captures the complexities of lexical processing during holistic listening comprehension, requiring learners to integrate lexical, phonological, morphosyntactic, and pragmatic cues in real time. Unlike traditional recognition-based tests that isolate specific lexical items, the LJT engages multiple linguistic subsystems simultaneously, thereby better reflecting the automaticity required for fluent comprehension.

Validation studies with Japanese learners of English have provided empirical support for the LJT. Administering the LJT alongside recognition and recall tasks as well as the standardized TOEIC listening test, these studies found that recognition and recall measures clustered together, while the LJT loaded onto a distinct factor. This pattern suggests that recognition and recall primarily index declarative knowledge, whereas the LJT more effectively taps automatized knowledge. Furthermore, LJT performance showed stronger correlations with TOEIC listening scores (r = .60-.70) compared with recognition and recall measures (r = .40-.50), underscoring the LJT's closer alignment with real-time global comprehension skills (see also Saito et al., 2026 for replication and extension studies).

Motivation for Current Study

Although previous research has distinguished between what meaning recognition and recall measure (i.e., form-meaning mapping) and what the LJT measures (i.e., use-in-context), these distinctions have thus far been examined in cross-sectional studies. Such studies have shown that L2 learners' automatized vocabulary knowledge (measured via the LJT) correlates more strongly with global comprehension skills than their declarative vocabulary knowledge (measured via meaning recognition and recall). To extend this line of inquiry, the present study adopted a longitudinal perspective to examine the development of these distinct dimensions of phonological vocabulary knowledge. Importantly, the declarative-automatized model of L2 knowledge has been developed, elaborated, and refined as a dynamic framework for explaining how L2 knowledge evolves in classroom settings, rather than as a stable trait assessed at a single time point (DeKeyser & Suzuki, 2025). Using the enhanced audiovisual training paradigm (Majuddin et al., 2021, 2024), we designed an intervention study with a pretest–posttest design to investigate whether, to what extent, and how L2 learners differentially develop declarative and automatized dimensions of phonological vocabulary knowledge when exposed to and learning new L2 word expressions (see below).

As articulated in skill acquisition theory for instructed L2 learning (DeKeyser & Suzuki, 2025; Suzuki & DeKeyser, in press), the initial stage of L2 development involves establishing and strengthening form-meaning associations with explicit and conscious attention (proceduralization of declarative knowledge). With repeated exposure to target items in sentential contexts, later stages of development involve retrieving target words more quickly, automatically, and semantically appropriately in relation to surrounding words (automatization). This developmental perspective also aligns with the instructional approach suggested by Schmitt (2008) that an explicit approach (i.e., building a form-meaning link) needs to be followed by repeated contextual exposure to the word so that its knowledge is *consolidated* (i.e., a more robust form-meaning link) and *enhanced* (i.e., acquisition of other types of word knowledge).

Building on this developmental paradigm, the current study focused on the learning of multiword expressions, defined according to the corpus-based measures of frequency (e.g., phrase frequency and mutual information) as combinations of two or more words that frequently occur in natural language use. Our target items therefore included a wide range of expression types such as collocations (e.g., *slippery slope*), idioms (e.g., *on the same page*), phrasal verbs (e.g., *chip in*), and so forth (Siyanova-Chanturia & Van Lancker Sidtis, 2019). The rationale for the choice of multiword expressions as target items was that participants were assumed to be generally familiar with the forms of constituent words (i.e., 96% of the words were within the range of the high-frequency words—i.e., most frequent 3,000 word families). Such increased familiarity was expected to create a favorable condition for learning where encoding of novel word forms would not be necessary. Thus, learners would be able to expend greater attentional resources for encoding the semantic and contextual information about the multiword expressions.

Following the methodological paradigm reviewed earlier (Majuddin et al., 2021, 2024), we employed enhanced audiovisual training to investigate the acquisition of multiword expressions. We regarded this context as an ideal testing ground for examining the declarative and automatized dimensions of L2 vocabulary knowledge. Although the training was brief (20 minutes of audiovisual input), previous research has shown that such enhanced audiovisual exposure reliably facilitates L2 vocabulary learning and constitutes an explicit instructional intervention that primarily targets the *initial stage* of L2 lexical acquisition. Accordingly, we anticipated that learning gains would be most evident in the declarative dimension—reflecting early, explicit learning of form–meaning connections—whereas evidence of automatization would likely be weaker or less clearly observable at this early stage. Given the exploratory nature of this study (i.e., the first longitudinal investigation of its kind) and the limited training duration (20 minutes), it was crucial to design conditions that maximised the learnability of target items and allowed us to trace developmental trajectories ranging from the establishment of form–meaning mappings (declarative knowledge) to the emergence of context-based, automatic retrieval (automatized knowledge):

1. Prior to exposure to target multiword expressions, differences in learners' declarative and automatized knowledge would be minimal, reflecting limited experience with these expressions.
2. During training, learners would show larger improvements in the declarative dimension, as evidenced by greater gains on meaning recognition and recall tasks, whereas gains in automatized knowledge (assessed by the LJT) would remain limited.
3. Following training, the distinction between declarative and automatized knowledge would become more evident, with declarative knowledge showing earlier and stronger development than automatized knowledge.

Method

Setup

The experiment was conducted individually using an online experiment platform (Gorilla) in combination with a videoconferencing tool (Zoom). The project was advertised as an English vocabulary learning study through an electronic flyer distributed across a university in China. Interested participants contacted an investigator (a native speaker of Mandarin Chinese) and arranged a convenient time to meet via Zoom. At the beginning of each session, the researcher conducted a sound check and verified participants' equipment (headphones, computer setup, and internet stability). Participants then completed a short version of a vocabulary test in Gorilla (5 minutes). Once the researcher confirmed their eligibility, they proceeded with the pre-tests (10 minutes), the audiovisual vocabulary training (20 minutes), and the immediate post-tests (5 minutes). Throughout the session, the researcher remained virtually present, and participants could contact them at any point with questions.

Participants

A total of 34 Chinese learners of English as a Foreign Language (EFL) participated in this experiment. All were native speakers of Mandarin, aged between 21 and 25 ($M = 23.0$, $SD = 0.4$). To ensure adequate comprehension of the audiovisual training materials, participants were required to have sufficient lexical knowledge. Recruitment criteria were therefore: (a) successful completion of the College English Test Band 6 (CET-6; CEFR B2 to C1), and (b) receptive knowledge of at least 2,000 of the most frequent English word families, given the lexical demand of the video materials used in this study (Durbahn, Rodgers, & Peters, 2020; Webb & Rodgers, 2009). Participants' vocabulary knowledge of the most frequent 3,000 word families was verified using a modified version of the Vocabulary Size Test (Nation & Beglar, 2007). On average, they scored 24.5 out of 30 ($SD = 2.4$), indicating receptive knowledge of approximately 2,500–2,900 of the most frequent English word families.

Power Analysis of Sample Size

A power analysis was conducted using *G*Power* (Faul, Erdfelder, Lang, & Buchner, 2007) to evaluate whether the sample size ($n = 34$) provided sufficient statistical power. In the original studies by Majuddin et al. (2021, $n = 23$; 2024, $n = 24$), participants who completed the same audiovisual training procedure as in the present study showed significant improvements in vocabulary knowledge—measured through form recognition and recall—with medium-to-large effect sizes (Cohen's $d = 0.80$–0.90). These findings are consistent with Kurokawa et al.'s (2024) meta-analysis, which reported a medium effect size (Hedges' $g = 0.56$) for the relationship between captioned video viewing and vocabulary learning. Based on these prior results, we set the expected effect size at $f = 0.35$ (medium-to-large) and determined that a minimum sample size of $N = 29$ would be required to achieve a power level of .95. Accordingly, the current sample ($n = 34$) was considered adequate for detecting the hypothesized within-participant effects.

Although the sample size meets conventional power requirements for medium-to-large effects, we acknowledge that it remains modest for drawing population-level generalizations. The present findings should therefore be interpreted as evidence of short-term learning patterns within a specific intentional captioned-viewing paradigm. To address potential small-sample bias, we reported 95% confidence intervals for all key estimates and conducted sensitivity analyses using alternative random-effects structures and jackknife resampling across participants. All robustness checks converged on the same pattern (Recognition > Recall > LJT), indicating that the observed effects are unlikely to be artifacts of a particular sample composition. Future research with larger, multi-site samples and extended training durations will be necessary to examine whether the same declarative–automatized hierarchy persists over longer learning trajectories and across learner populations.

Audiovisual Vocabulary Training

Materials. We adopted training materials that had already been shown to be effective in promoting vocabulary learning, particularly among university-level EFL students (i.e., the target population of this study). Specifically, we used the audiovisual materials from Majuddin et al. (2021, 2024), as they met two key conditions: (1) at least 95% of the lexical profile of the materials consisted of frequent word families, ensuring that Chinese EFL learners could comprehend the video content (Van Zeeland & Schmitt, 2013); and (2) the materials included a sufficient number of potentially unfamiliar items (i.e., 18 target multiword expressions; see below).

In Majuddin et al. (2021, 2024), an episode of the American comedy series *Fresh Off the Boat* was selected because its engaging content sustained learners' attention over a 20-minute viewing period. The lexical coverage of the episode was analyzed using the RANGE software (Nation & Heatley, 2002), which indicated that the most frequent 1,000, 2,000, and 3,000 word families accounted for 91.16%, 94.87%, and 96.37% of the video transcript, respectively. These results align with prior research on the lexical demands of L2 video input. Webb and Rodgers (2009) estimated that a vocabulary size of approximately 2,000–3,000 word families is required for adequate comprehension of video content, depending on the genres of video materials. More recent work (Durbahn, Rodgers, & Peters, 2020) suggests that comprehension can be achieved with knowledge of about 90% of the words in a video. Based on these benchmarks, the current study set a receptive vocabulary of at least 2,000 of the most frequent word families as the minimum recruitment criterion for participants.

In Majuddin et al.'s (2021, 2024) study, participants who viewed the video clip once demonstrated relatively high levels of comprehension (M = 81.2%, SD = 10.7). To ensure the suitability of the material for the present study, we conducted a pilot test with a comparable population of Chinese EFL students (n = 54), which yielded similar comprehension levels (M = 85.7%, SD = 9.5). Because all three outcome measures had been validated and used in previous research (Majuddin et al., 2021; Uchihara et al., 2025; Saito et al., 2025), no additional pilot testing of item difficulty or discrimination was undertaken.

Target Items. Following Majuddin et al. (2021, 2024), the audiovisual materials were carefully examined to identify multiword expressions (MWEs) likely to be unfamiliar to participants, which were then used as the target items for vocabulary training. Building on prior evidence of the effectiveness of audiovisual input for single-word learning (Peters et al., 2019), Majuddin et al. demonstrated that such effects can also extend to the acquisition of MWEs. As the present input was drawn from an American comedy, each candidate expression was cross-referenced with the Corpus of Contemporary American English (COCA) to ensure its authenticity. To be selected, MWEs had to satisfy at least one of the following criteria: (1) a minimum of 100 occurrences in COCA, or (2) a Mutual Information (MI) score above 3.0, a commonly used indicator of collocational strength (Majuddin et al., 2021, 2024). A total of 18 target items were identified on this basis. Frequency of occurrence within the input was also considered, as repetition has been shown to affect vocabulary learning (Webb, 2007). Items appearing only once in the material were therefore given special consideration.

Because pretests may raise learners' awareness of specific items and encourage greater attention to them during subsequent tasks—a phenomenon known as *test effects*—an additional set of 10 distractor items was incorporated into the pretest. These distractors were drawn from the documentary series *Fry's Planet Word*, also used in Puimège and Peters (2020). Lexical profiling of this resource indicated that 91% of its vocabulary falls within the top 2,000 most frequent word families in the British National Corpus and COCA, and 93.76% within the top 3,000 families, closely resembling the lexical profile of the present study's input. Distractor items were chosen to be comparable to the target items, following the same selection criteria (≥100 COCA occurrences and/or MI > 3.0). For a list of 18 target items and its corpus frequency and MI score, see Supporting Information-S1.

Procedure. Participants first received explicit vocabulary guidance, followed by pre-tests, 20 minutes of viewing with standard L2 captions, and post-tests. Given that the main goal of the study related to the testing of the declarative-automatized distinction within the already well-researched method of L2 vocabulary training, we followed the enhanced and explicit audiovisual training paradigm set by Majuddin et al. (2021, 2024). Specifically, at the outset of the projects, participants were first explicitly made aware of the following before they proceeded with pre-tests and they watched the video:

- The purpose of the activity: The viewing activity served the purpose of learning MWEs.
- The anticipated assessment: They were told that a test would come before and after viewing.
- The focus of the test: The test would specifically focus on the MWEs encountered in the video.

Following Majuddin et al., participants were *not* provided with a list of the 18 target items in advance, in order to prevent an excessive focus on linguistic forms. Although this training can be characterised as explicit and intentional in nature, the primary goal of the activity was comprehension and enjoyment of the L2 video content, with vocabulary learning framed as a secondary, by-product outcome. This combination of intentional and incidental learning aligns with the framework proposed by Laufer and Hulstijn (2001). The approach has been shown to enhance engagement with unfamiliar items during reading activities (Peters et al., 2009) and, more recently, during audiovisual viewing (Majuddin et al., 2021, 2024).

After the explicit vocabulary guidance, participants completed three tasks to assess their pre-existing knowledge of the target items: (1) the Lexicosemantic Judgment Task, (2) a meaning recall test, and (3) a meaning recognition test, in this order (see below for details). All the participants then watched the 20-minute video clip under the same condition. Following the standard captioned viewing design used in Majuddin et al. (2021, 2024), the video was presented with *standard* L2 captions displayed throughout. However, to prevent excessive attention to specific lexical forms and to avoid potential negative effects on overall comprehension, *enhanced* captions were not used (see Majuddin et al., 2021, 2024, for evidence that enhanced captions, while beneficial for vocabulary learning, can hinder comprehension).To ensure that participants engaged with the video content at the discourse level rather than focusing solely on isolated lexical items, they also answered ten comprehension questions. These comprehension questions were included solely to maintain engagement with the storyline and were not analysed as part of the study's outcome measures.

Following Majuddin et al. (2021, 2024), the present study focused exclusively on the standard-caption condition because prior research has consistently shown that vocabulary gains occur only when captions are available, whereas uncaptioned viewing produces negligible improvement. This approach allowed us to examine how a validated audiovisual training method influences different dimensions of lexical knowledge (declarative vs. automatized) rather than re-establishing the baseline captioning effect. Indeed, Majuddin et al. found that watching the same video without captions did not lead to significant vocabulary gains, regardless of whether participants received explicit vocabulary guidance. Similar findings have been reported in Saito et al. (2024), where explicit vocabulary training yielded robust improvements in both declarative and automatized knowledge, while a comparison group who merely completed the same tests twice showed no learning gains. These converging results indicate that short-term vocabulary gains under captioned or explicit training cannot be attributed to test–retest or familiarity effects.

Outcome Measures

To assess the impact of audiovisual training on both declarative and automatized dimensions of vocabulary knowledge, participants completed three tasks in the following order: (1) the LJT, (2) meaning recall, and (3) meaning recognition. Following Uchihara et al.'s (2025) methodological framework, these tasks were assumed to capture complementary dimensions of L2 phonological vocabulary knowledge. The task order (LJT → meaning recall → meaning recognition) was chosen to minimise participants' excessive focus on target lexical forms, particularly during the LJT, which required processing not only the target items but also surrounding words and sentence-level meaning. Meaning recall was administered before meaning recognition to avoid revealing answers from L1 options provided through the meaning-recognition test because the former is generally more demanding—it requires not only knowledge of form–meaning mappings but also lexical retrieval abilities (González-Fernández & Schmitt, 2020). Conducting the recall task first reduced the risk that participants would focus too narrowly on lexical forms in the subsequent tasks. In contrast, the meaning recognition task primarily assessed whether participants recognised the target expressions without requiring retrieval or sentence-level processing. Following Majuddin et al. (2021, 2024), although the order of tasks was fixed, the test items within each task were presented in randomised order to minimise potential order and practice effects.

As outlined in skill acquisition theory for instructed L2 learning (DeKeyser & Suzuki, 2025; Suzuki & DeKeyser, in press), dual-modality considerations were applied in operationalizing these constructs. For declarative knowledge, measured through meaning recognition and recall, participants focused on target lexical items in isolation (i.e., single modality), albeit at different levels of processing (Chen et al., 2023). Meaning recognition assessed learners' ability to identify the correct meaning of a word from a list of options (e.g., multiple-choice format), while meaning recall tested their ability to produce a word's meaning without prompts (e.g., providing a definition or translation).

For automatized knowledge, measured through the LJT, learners' vocabulary knowledge was assessed within sentential contexts that simultaneously engaged grammar, pragmatics, and discourse processing (i.e., multi-modality). The LJT required participants to evaluate the appropriateness of target lexical items in sentence-level contexts, thereby capturing the extent to which their lexical knowledge could be rapidly and accurately integrated during real-time comprehension.

Validation evidence supports this conceptual distinction. Uchihara et al. (2025) found that the three tasks loaded onto two separate latent factors, with the LJT reflecting automatized knowledge and recall/recognition reflecting declarative knowledge. Moreover, these two constructs demonstrated differential associations with learners' global listening proficiency, with automatized knowledge (LJT) showing stronger predictive power than declarative knowledge (recall and recognition).

Lexicosemantic Judgements. Participants listened to 36 sentences (produced by a male native speaker of American English) and judged whether each was semantically appropriate. All sentences were grammatically accurate and simple (i.e., no subordination), and all words were drawn from the 2,000 most frequent word families. The sentences were divided into two types: semantically appropriate (*n* = 18) and semantically inappropriate (n = 18). In the appropriate sentences, a target lexical expression (e.g., "*on the same page*") was used in a contextually correct manner (e.g., "*We need to be <u>on the same page</u> before we move forward with the project*"). In the inappropriate sentences, the same expression was used in an ill-formed context (e.g., "*Please put the box <u>on the same page</u> as the book*"). Another example was "*turn a profit*," which was used appropriately in "*The company expects to <u>turn a profit</u> in the next year*" and inappropriately in "*She decided to <u>turn a profit</u> on the floor*." For the test stimuli, see Supporting Information-S3.

Target lexical items appeared in varied positions within the sentences. After the development of 18 target items and their 36 corresponding test sentences, an expert review was conducted to ensure the reliability and validity of the stimuli. This process involved collaboration with two experts in second language acquisition, each with over two decades of experience in the field: one a native English speaker and the other a near-native English speaker. To avoid test-training overlap, none of the items were embedded in lexical contexts that had appeared in the video clip. Following Uchihara et al. (2025), a strict scoring criterion was applied: one point was awarded only when participants both accepted the semantically appropriate sentence and rejected the corresponding inappropriate one. This method ensures that scores reflect accurate lexical-semantic integration, not partial recognition, and has been shown to yield high reliability in validation studies.

The LJT was administered in an untimed auditory format. Participants were instructed to respond as quickly and accurately as possible, but no time limit was imposed. This approach aligns with prior validation research showing that untimed auditory LJTs strongly predict L2 listening proficiency (Uchihara et al., 2025) and that participants' reaction times (operationalized as coefficients of variation) are unrelated to L2 listening proficiency (Saito et al., 2025), both of which indicate automatized lexical–semantic integration. Because auditory stimuli unfold once and at a fixed pace, additional time constraints were deemed unnecessary and could disadvantage learners with slower lexical decoding speeds (Maie & Godfroid, 2022; see also Hulstijn et al., 2009 for their critical discussion on reaction time, performance variability, and automaticity). Further evidence from Saito et al. (2026) likewise shows comparable predictive validity for timed and untimed formats, supporting the present design.

Meaning Recall. The meaning recall test was designed to measure the productive knowledge of the connection between the aural form of multiword expressions and the corresponding meanings. Participants first listened to a target multiword expression (produced by a male native speaker of American English) and were asked to indicate whether they recognised each one. If they answered "Yes," they were then prompted to demonstrate their understanding by providing a translation, an explanatory definition, or a synonymous expression in their L1 Mandarin Chinese. When participants did not recognise an expression, they were instructed to enter "I do not know," which served to reduce guessing and enhance the reliability of the data.

Meaning Recognition. The meaning recognition test was designed to assess participants' receptive knowledge of the connection between the aural form of multiword expressions and the corresponding meanings. The multiple choice format was adopted in which each test item presented a target expression alongside four Chinese translation options. These options were directly adapted from Majuddin et al. (2021, 2024) (for test stimuli, see Supporting Information S2). Participants were instructed to select the option they believed most accurately captured the meaning of the target expression. This format enabled a controlled assessment of recognition-level understanding, allowing us to capture partial gains of target words.

Results

Descriptive statistics for participants' vocabulary test scores across different tasks and time conditions are reported in Table 1 and illustrated in Figure 1. Relatively high pretest scores (with accuracy rates of about 70% or higher across all tasks) were both expected and desirable, as the aim of this study was to examine not only the initial stage of learning (i.e., declarative form–meaning knowledge) but also the subsequent stage of lexical acquisition, which involves encoding use-in-context properties (i.e., automatized knowledge). The learners' relatively high familiarity with the constituent words, along with their partial knowledge of the phrasal meanings (likely derived from inferring the whole meaning from the parts), was expected to enable them to allocate sufficient cognitive resources to encoding the semantic, collocational, and grammatical properties of the target items.

Results of the Shapiro-Wilk normality tests indicated no statistically significant deviation from a normal distribution ($p > .05$). Participants' vocabulary scores were analysed using linear mixed-effects models implemented in the lme4 package (Bates et al., 2015) in R (R Core Team, 2025). The dependent variable was participants' percentage accuracy on the vocabulary tests. Time (pre-test vs. post-test) and Task (recognition, recall, LJT) were included as fixed effects, with their interaction also specified, and Participant ID was entered as a random effect.

To identify the optimal random-effects structure, we compared alternative models varying in random slopes for Time and Task using maximum-likelihood estimation (AIC-based model comparison). We then selected the most parsimonious non-singular model with uncorrelated random slopes. The final model included random intercepts and random slopes for Time and Task by participant, DV ~ Time * Task + (1 + Time + Task || ID). This model provided a good fit ($R^2_{marginal} = .729$; $R^2_{conditional} = .895$). A sensitivity analysis using a simpler random-effects structure ((1 + Time || ID)) yielded the same fixed-effect pattern and planned contrasts (Recognition > Recall > LJT), with a lower $R^2_{conditional}$. A sensitivity analysis using a simpler random-effects structure ((1 + Time || ID)) produced comparable fixed-effect estimates and planned contrasts (Recognition > Recall > LJT) but yielded a smaller proportion of variance explained ($R^2_{\text{marginal}} = .298$; $R^2_{\text{conditional}} = .748$), confirming the robustness of the main findings.

TABLE 1. Descriptive statistics of participants' vocabulary scores (%) at pretest and posttest.

Task conditions	Time	*M* (%)	*SE*	95% *CI* Lower	95% *CI* Upper
Recognition	Pre	70.5	1.68	67.1	74
	Post	83.5	1.98	79.5	87.5
Recall	Pre	71.2	1.56	68.0	74.4
	Post	80.1	1.89	76.3	84
LJT	Pre	68.9	1.84	65.1	72.6
	Post	76.8	2.23	72.2	81.3

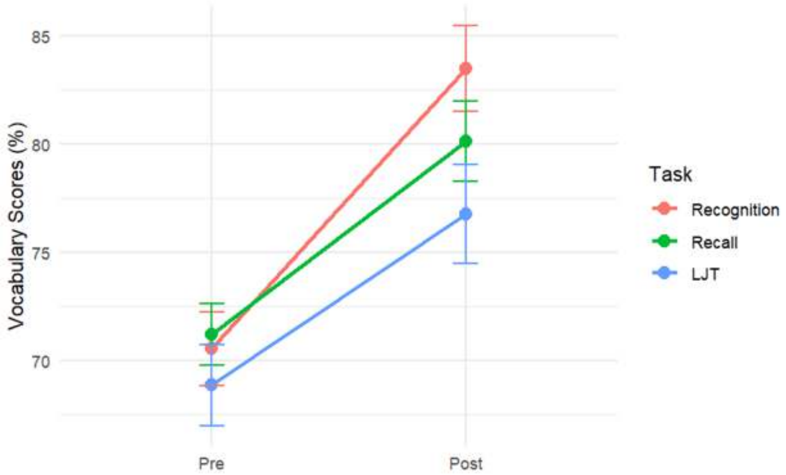

FIGURE 1. Visual summary of participants' vocabulary training gains under three task conditions. Improvements differed across tasks in the following order: Recognition > Recall > LJT. The task effect was more evident at the posttest session (relative to the pretest session).

In the final model (summarized in Table 2), there was a significant main effect of Time, $F(1, 32) = 29.11$, $p < .001$, indicating that post-test scores were higher than pre-test scores. Crucially, a significant Time × Task interaction emerged, $F(2, 64) = 28.10$, $p < .001$, demonstrating that the size of the improvement differed across tasks. To interpret this interaction, we calculated estimated gains (post – pre) for each task using the model coefficients. The overall gain for Recognition was $b = 12.96$, while the additional interaction terms indicated smaller gains for Recall ($b = 12.96 + [–4.04] = 8.92$) and LJT ($b = 12.96 + [–5.05] = 7.91$). Thus, learning gains followed the order Recognition > Recall > LJT, consistent with the descriptive results (Table 1, Figure 1). The standardised coefficients (β) confirmed a large effect for Time ($\beta = 1.09$) and substantial Time × Task effects, whereas the between-task differences at pre-test were negligible.

TABLE 2. Summary of the final linear mixed-effects model predicting vocabulary test scores.

Predictor	*b*	*β*	*SE*	*t*	*p*
Intercept (Recognition, Pre)	70.54	–0.39	1.68	41.99	< .001*
Time (Post vs. Pre)	12.96	1.09	1.89	6.87	< .001*
Task: Recall vs. Recognition	0.67	0.06	0.87	0.77	0.443
Task: LJT vs. Recognition	–1.68	–0.14	1.36	–1.24	0.224
Time × Recall	–4.04	–0.34	0.71	–5.67	< .001*
Time × LJT	–5.05	–0.42	0.71	–7.08	< .001*
Type III ANOVA	*F*	*p*			
Time	29.11	< .001*			
Task	5.43	.009*			
Time × Task	28.10	< .001*			

Notes. b = unstandardized fixed-effect estimate; β = standardized coefficient (DV scaled to SD units). Reference condition = Recognition at Pre-test. Random-effects structure: (1 + Time + Task || ID) (uncorrelated slopes). Model fit: $R^2_{marginal} = .729$; $R^2_{conditional} = .895$.

With respect to estimated marginal means (see Table 1), vocabulary scores increased significantly from pre- to post-test across all tasks: recognition (M = 70.5% → 83.5%), recall

(*M* = 71.2% → 80.1%), and LJT (*M* = 68.9% → 76.8%). The corresponding contrasts confirmed that post-test scores were significantly higher than pre-test scores for recognition, $t = –6.87$, $p < .001$, $d = 1.16$, recall, $t = –4.73$, $p < .001$, $d = 1.10$ and LJT, $t = –4.19$, $p < .001$, $d = 0.64$. Pairwise comparisons among tasks indicated that at pre-test there were no significant differences between recognition, recall, and LJT (all *ps* > .05). At post-test, however, recognition scores were significantly higher than recall, $t = 3.87$, $p = .001$, $d = 0.67$, and LJT, $t = 4.95$, $p < .001$, $d = 0.84$; recall scores were also significantly higher than LJT, $t = 2.47$, $p = .047$, $d = 0.38$.

Taken together, these results suggest that while participants entered the study with broadly comparable performance across tasks, recognition improved the most following training, followed by recall, with LJT showing the smallest relative gains. However, the effect size results added that the difference between meaning recognition and the LJT was large, whereas the relative difficulty of meaning recall (compared to LJT) appeared to be small.

Discussion

By integrating skill acquisition theory for instructed L2 learning (DeKeyser & Suzuki, 2025; Suzuki & DeKeyser, in press) into Nation's (2013) oft-cited framework of vocabulary knowledge, recent research has increasingly discussed phonological vocabulary knowledge as comprising two distinct dimensions—declarative and automatized—along with their corresponding measures, such as meaning recognition and recall (declarative knowledge) and lexicosemantic judgments (automatized knowledge; e.g., Uchihara et al., 2025). As skill acquisition theory accounts for the processes and products of learning, the declarative–automatized distinction suggests a two-stage model of phonological vocabulary development: form–meaning mapping to use-in-context. In the first stage, L2 learners focus on form–meaning mapping under relatively controlled conditions, where they can process and retrieve target lexical items without significant time pressure (i.e., proceduralization of declarative knowledge). In the second stage, learners begin to use these items more promptly and appropriately in collocational, grammatical, and discourse contexts at the sentence level, approximating real-life comprehension.

Given that existing studies have largely provided only cross-sectional evidence, the present study adopted a longitudinal, pre-post intervention design with audiovisual input to examine whether, and to what extent, L2 learners differentially develop the declarative and automatized dimensions of phonological vocabulary knowledge through training. Specifically, our predictions were threefold: (a) task effects would be minimal prior to training, given participants' lack of prior exposure to the target lexical items; (b) with only a single training exposure, participants would show stronger gains in declarative knowledge—captured by meaning recognition and recall—than in automatized knowledge—captured by the LJT; and (c) after training, the differences between declarative and automatized dimensions would become more apparent, with clearer task effects reflecting the early stages of proceduralization but limited automatization.

In the context of Chinese EFL students' acquisition of 18 multiword expressions during 20 minutes of captioned video exposure, three main findings emerged. First, participants' baseline knowledge of the target lexical items was comparable across the three task conditions. Second, training gains differed by task, with the largest improvements observed in the following order: Meaning Recognition > Meaning Recall > LJT. Third, after training, task effects became more clearly observed, with participants demonstrating the highest levels of knowledge when assessed via meaning recognition, followed by meaning recall and then the LJT. Taken together, these findings broadly supported our hypotheses, providing empirical evidence for the declarative–automatized distinction. Specifically, they suggest that L2 phonological vocabulary development proceeds along a trajectory from form-meaning mapping to use-in-context, and that this process can be systematically captured by different outcome measures (e.g., meaning recognition as an index of proceduralization, and the LJT as an index of automatization).

A more tentative finding concerned the meaning recall task. In Uchihara et al.'s (2025) cross-sectional investigation, meaning recall and recognition were found to cluster together, distinct from the LJT. In the present study, however, gains on meaning recall appeared somewhat closer to those on the LJT (small effects), compared to the substantial difference between meaning recognition and LJT (large effects). This raises the possibility that recall may under certain conditions tap into processing beyond form-meaning mapping. This interpretation aligns with the suggestions in the literature that recall tasks can sometimes index deeper, more integrative processing of lexical items (Chen et al., 2023).

At the same time, this finding should be interpreted cautiously, given the current study's design: participants encountered the target items only once, albeit with captions and pre-training instructions that encouraged them to attend to these items. Such limited exposure may have encouraged partial learning beyond form–meaning mapping, but was unlikely to yield fully automatized knowledge. Since automatization of phonological vocabulary knowledge—arguably what the LJT is designed to measure—requires repeated encounters with target items in communicatively authentic contexts, future studies should examine whether, and to what extent, meaning recall and the LJT converge or diverge when learners are given richer and more frequent exposure. Furthermore, it is important to note that long-term retention of the declarative and automatized knowledge was not examined; thus, lexical gains observed in this study should be considered a fragment of the overall lexical development. Future studies should investigate whether more extensive training (e.g., extensive viewing over one year) leads to not only the initial learning of form-meaning and use-in-context knowledge but also the retention of the learning over time (see Saito & Uchihara, 2024 for the experiential and perceptual correlates of *long-term* development of declarative vs. automatized knowledge).

Implications and Future Directions

Given the cross-sectional evidence (e.g., Uchihara et al., 2025) and the longitudinal findings of the present study, several pedagogical/methodological implications and future research directions can be proposed. First and foremost, vocabulary learning should be conceptualised and measured *beyond* the form-meaning level. This claim does not seem to be new, given that the importance of assessing different types of word knowledge (e.g., collocation, association, grammatical function) with varying degrees of sensitivity (e.g., partial to complete knowledge) was emphasised quite a while ago in instructional L2 vocabulary research (Waring & Takaki, 2003; Webb, 2007; see also Schmitt, 2008). This proposal can be viewed as a significant shift in the paradigm of L2 vocabulary learning research, encouraging scholars to use multiple tests to measure vocabulary acquisition. However, the way to elicit responses from learners, whether recognition or production, has thus far remained limited to the single-task procedure, where each of the different aspects (or sensitivities) of word knowledge is measured separately through a word-in-isolation task (e.g., translation and multiple-choice). This issue echoes one of the major problems identified by Schmitt (2019, p. 269):

> What we really want in vocabulary measurement is the ability to infer what learners can DO with the target words. (Nobody interprets test scores as simply words that learners can answer on a vocabulary test!) … That is, receptive/productive knowledge of vocabulary is usage-based, and should presumably be measured with skill-based instruments. However, it is hardly ever measured this way.

In response to Schmitt's call, our longitudinal study provided initial evidence suggesting that the Lexicosemantic Judgement Task may function as a "skill-based instrument" (targeting listening as the focal skill) capable of capturing lexical development beyond declarative knowledge of form and meaning. We argue that future studies, by incorporating a skill-based instrument (e.g., the LJT) into the traditional test battery, may enable researchers to assess not only how many words (or phrases) are learned through instruction (i.e., declarative knowledge), but also the extent to which these items progress toward becoming employable in real-life communication (i.e., automatized knowledge).

From a pedagogical perspective, the current study highlights the importance of supporting learners' development beyond initial mapping, towards automatization of vocabulary knowledge. Future research should explore instructional designs and input conditions that not only promote form–meaning mapping but also facilitate the gradual automatization of lexical knowledge over time (cf. Saito et al., 2024 for meaning recognition training for declarative knowledge development and lexicosemantic judgement training for automatized knowledge development). Not all types of vocabulary activities are necessarily assumed to facilitate the development of automatized lexical knowledge. A key factor in promoting automatization is the encoding of use-in-context properties associated with individual words (Uchihara et al., 2025). Accordingly, in line with Nation's (2007) four strands, promising instructional approaches worthy of future investigation are those that emphasize opportunities for meaning-focused input and fluency development.

Meaning-focused input activities should expose learners to L2 vocabulary repeatedly and across varied contexts. For beginners, repeated engagement with the same materials (e.g., repeated viewing; Majuddin et al., 2021, 2024) can be effective, followed by topic-related but varied materials (e.g., narrow viewing, Rodgers & Webb, 2011; narrow reading while listening, Chang, 2019), and ultimately, extensive exposure to authentic materials over time (e.g., extensive viewing, Webb, 2015). Remember that the consolidation of the form-meaning mapping is prerequisite for subsequent enhancement of the knowledge (Schmitt, 2008; Saito et al., 2024). Thus, the effectiveness of such comprehension-based activities may be maximized when target vocabulary is taught explicitly in advance (e.g., pre-teaching before viewing, Pujadas & Muñoz, 2019) or when learners' attention is drawn to word forms while they engage with meaning-focused activities (e.g., L1 explanations, Zhang & Graham, 2020; dictogloss, Yu et al., 2025).

Finally, it is important to note the relationship between the LJT paradigm and existing work on lexical automaticity. Although limited, several attempts have been made to measure the extent to which lexical access and recognition occur rapidly and without conscious effort. One such approach is the lexical decision task (LDT), which captures both accuracy (reflecting lexical knowledge) and reaction times (reflecting the efficiency and automaticity of lexical retrieval). For the latter, Segalowitz and Segalowitz (1993) proposed analysing both mean reaction times and their coefficients of variation to distinguish between general speed-up effects and genuine automatization (but see Hulstijn et al., 2009, Maie & Godfroid, 2023, and Saito et al., 2026 for a critical discussion and empirical evidence on the roles of reaction time and variability in automaticity). Building on this foundation, L2 studies have employed LDTs to trace the development of lexical fluency and automaticity in L2 learners. For instance, Elgort (2011) investigated how learners can automatize their knowledge of pseudowords through intentional training involving meaning recognition and feedback. The study used lexical decision tasks in which participants were shown letter strings and asked to make intuitive judgements about whether each represented a word or a non-word. The results showed that participants recognised target words more quickly and accurately after exposure to similar forms with related semantic meanings. These findings highlight the effectiveness of deliberate learning strategies in enhancing automatic lexical processing.

Furthermore, Hui and Godfroid (2020) demonstrated that L2 listeners' vocabulary processing speed and stability (operationalized via coefficients of variation) during LDTs correlate with listening proficiency. Following Hui and Godfroid, an intriguing future direction concerns the examination of different stages of automatization. The first stage involves prompt and stable access to words in isolation (as indexed by LDT performance), whereas the second stage concerns lexical automaticity in sentence-level processing (as indexed by LJT performance). Both cross-sectional studies (examining the relative strength of association between LDT/LJT measures and global listening proficiency) and training studies (involving sequential LDT and LJT practice) could offer more nuanced insights into a hierarchical model of lexical automaticity.

Beyond behavioural measures, future research could further advance the understanding of lexical automaticity by incorporating eye-tracking and pupillometry techniques. Recent work has demonstrated that task-evoked pupil dilation serves as a sensitive indicator of cognitive effort and attentional engagement during word recognition and retrieval (e.g., McLaughlin, Zink, et al., 2022; for a scope review on the use of the technique in L2 and bilingual research, see Rojas, Vega-Rodríguez, 2024). Specifically, larger and later pupil dilations have been associated with lower proficiency, increased lexical competition, and greater processing difficulty, whereas smaller and earlier dilation responses signal more automatized lexical access. By time-locking pupil responses to the onset of target words—either in isolation (as in LDT) or embedded in sentential contexts (as in LJT)—researchers can capture the temporal dynamics of cognitive effort across different stages of lexical processing. Integrating pupillometric indices (e.g., peak amplitude, latency, and area under the curve) with traditional behavioural measures such as accuracy, reaction time, and response variability could therefore provide a multimodal window into the hierarchical development of lexical automaticity. Such an approach would offer richer insights into how L2 learners transition from effortful to automatic lexical access, linking behavioural performance with underlying neurocognitive mechanisms.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Chang, A. C. (2019). Effects of narrow reading and listening on L2 vocabulary learning: Multiple dimensions. *Studies in Second Language Acquisition*, *41*(4), 769-794.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5-34.

DeKeyser, R. M., & Suzuki, Y. (2025). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (4th ed., pp. 157-182). Routledge.

Cheng, J., Matthews, J., Lange, K., & McLean, S. (2023). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, *57*(1), 213-241.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, 102166.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*, 367-413.

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375-396.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481-505.

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089-1115.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? Applied Psycholinguistics, 30, 555–582.

Kurokawa, S., Hein, A. M., & Uchihara, T. (2024). Incidental vocabulary acquisition through captioned viewing: A meta-analysis. *Language Learning*.

Majuddin, E., Boers, F., & Siyanova-Chanturia, A. (2024). The effects of enhancing L2 multiword items in captions: An approximate replication of Majuddin, Siyanova-Chanturia, and Boers (2021). *Language Teaching*, 1-18.

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition*, *43*(5), 985-1008.

Mayer, R. E., Fiorella, L., & Stull, A. (2020). Five ways to increase the effectiveness of instructional video. *Educational Technology Research and Development*, *68*(3), 837–852.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760.

McLaughlin, D. J., Zink, M. E., Gaunt, L., Spehar, B., Van Engen, K. J., Sommers, M. S., & Peelle, J. E. (2022). Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults. *Psychonomic Bulletin & Review*, *29*(1), 268-280.

Nation, I. S. P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 2-13.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*, 9-13.

Nation, I. S., & Heatley, A. (2002). *Range: A program for the analysis of vocabulary in texts.* Retrieved from http://www.vuw.ac.nz/lals/staff/paulnation/nation.aspx.

Perez, M. M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, *42*(4), 749-773.

Perez, M. M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, *55*(2), 163-192.

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*(4), 1008-1032.

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, *59*(1), 113-151.

Puimège, E., & Peters, E. (2020). Learning formulaic sequences through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *42*(3), 525-549.

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, *47*(4), 479-496.

Rodgers, M. P., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*(4), 689-717.

Rojas, C., Vega-Rodríguez, Y. E., Lagos, G., Cabrera-Miguieles, M. G., Sandoval, Y., & Crisosto-Alarcón, J. (2024). Applicability and usefulness of pupillometry in the study of lexical access. A scoping review of primary research. *Frontiers in Psychology*, *15*, 1372912.

Saito, K., Hosaka, I., Suzukida, Y., Takizawa, K., & Uchihara, T. (2026). Timed vs. untimed lexicosemantic judgement task for measuring automatized phonological vocabulary knowledge. *Second Language Research*.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652-708.

Saito, K., & Uchihara, T. (2025). Experiential, perceptual, and cognitive individual differences in the development of declarative and automatized phonological vocabulary knowledge. *Bilingualism: Language and Cognition, 28,* 427-443.

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2024). Declarative and automatized phonological vocabulary knowledge in L2 listening proficiency: A training study. *Applied psycholinguistics*, *45*(6), 1187-1218.

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2025). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*, *47*(1), 26-52.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329-363.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274.

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice and the differentiation of speedup from automatization effects. *Applied Linguistics, 14*, 369–385.

Suzuki, Y., & DeKeyser, R. M. (in press). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), The Routledge handbook of instructed second language acquisition (2nd ed.). Routledge.

Suzuki, Y., & Elgort, I. (2023). Measuring automaticity in a second language: A methodological synthesis of experimental tasks over three decades (1990-2021). In Y. Suzuki (Ed.), *Practice and automatization in second language research: Perspectives from skill acquisition theory and cognitive psychology* (pp. 206-234). New York, NY: Routledge.

Siyanova-Chanturia, A, & Van Lancker Sidtis, D. (2019). What online processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 31-61). Routledge.

Takizawa, K., Saito, K., & Suzukida, S., Kurokawa, A., & Uchihara, T. (2026). Automatized knowledge to automaticity in speech: Examining the contribution of automatized phonological vocabulary knowledge to L2 utterance fluency. *Applied Linguistics*. https://doi.org/10.1093/applin/amaf042

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning, 70*(2), 506–547.

Teng, M. F. (2025). Modality of input and factors affecting incidental vocabulary learning: Reading, listening, and viewing with captions. *Applied Linguistics Review*, *16*(4), 1607-1635.

Uchihara, T., Saito, K., Kurokawa, S., Takizawa, K., & Suzukida, Y. (2025). Declarative and automatized phonological vocabulary knowledge: Recognition, recall, lexicosemantic judgment, and listening-focused employability of second language words. *Language Learning*, *75*(2), 458-492.

Vanderplank, R. (2016). 'Effects of' and 'effects with' captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, *49*(2), 235-250.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied linguistics*, *34*(4), 457-479.

Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language learning*, *72*(1), 5-44.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 1-27.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*, 45-65.

Webb, S. (2015). Extensive viewing: Language learning through watching television. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 159-168). Routledge.

Webb, S., & Rodgers, M. P. (2009). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407-427.

Yu, X., Boers, F., & Tremblay, P. (2025). Learning multiword items through dictation and dictogloss: How task performance predicts learning outcomes. *Language Teaching Research*, *29*(6), 2658-2678.

Zhang, P., & Graham, S. (2020). Vocabulary learning through listening: Comparing L2 explanations, teacher codeswitching, contrastive focus-on-form and incidental learning. *Language Teaching Research*, *24*(6), 765-784.

Supporting Information-S1: Target Lexical Items with Corpus Frequency and MI Score

MWEs	COCA	MI score
(Somebody’s) hands are tied	406	5.55
Going through a rough patch	63	14.51
Put (somebody) on the spot	108	5.04
Tighten up	461	5.43
Whisked away	338	6.38
Let up on (someone)	138	1.3
Chip in	1016	1.54
on (someone’s) hands	4,071	4.52
Work (something) out with (someone)	25	8.53
On the same page	2044	9.93
Look out for (someone)	2784	3.18
Turn a profit	491	10.48
Talk some sense into (someone)	258	5.98
Beg to differ	623	10.98
Bear with someone	1069	0.35
Slippery slope	1433	13.87
Root for	1516	2.48
Kill (someone) with kindness	28	11.76

Supporting Information-S2: Test Stimuli for Meaning Recogtniion Task

1. Have something on one’s hands
(a) 有一个人或一件事必须要处理 (c)对某个人的死亡负责
(b) 收到了一个被偷的东西 (d) 在某人的安全保护下

2. Tighten up
(a) 更严格或严肃地对待某件事 (c) 拧紧盖子或覆盖物
(b) 使一个关系变得更加紧密或牢固 (d) 保守秘密

3. Look out for someone
(a) 和某人在一起时要小心 (c) 检查某人
(b) 监视某人 (d) 照顾或关心某人

4. Someone’s hands are tied
(a) 参与到其他人的事物中 (c) 被抓到做违法的事情
(b) 不能够提供帮助 (d) 不放弃或抛弃某人或某事

5. Kill someone with kindness
(a) 仁慈或慷慨地对待某人 (c) 温柔地杀死某人
(b) 笑里藏刀 (d) 捧杀

6. On the same page
(a) 按照其他人的方式做某件事 (c) 有相似的思考或理解方式
(b) 原地踏步，没有任何进展 (d) 读好几次

7. Rooting for someone
(a) 寻找并除掉制造麻烦的人 (c) 把某人从家乡里调走
(b) 鼓励某人因为你想他们实现某件事 (d) 从小就教给某人他们的文化

8. Going through a rough patch
(a) 使用可利用的资源 (c) 经历低谷期
(b) 在崎岖不平的路上开车 (d) 质量低劣的作品

9. Turn a profit
(a) 拒绝收钱 (c) 公司倒闭或停止运营
(b) 扣除成本后赚取利润 (d) 抢别人的生意

10. Let up on someone
(a) 告诉某人一个秘密 (c) 不再指望某人会进步
(b) 对某人失望 (d) 减轻对某人的压力或要求

11. Slippery slope
(a) 一种可能会发展成为极坏的习惯 (c) 又湿又危险的小路
(b) 一个很难抓到的罪犯 (d) 一个聪明又不诚实的人（狡猾）

12. Put someone on the spot

(a) 迫使某人回答一个难题 (c) 在正确的时间和地点遇到某人
(b) 把某人和其他人分开 (d) 观看某人在舞台上表演

13. Bear with someone
(a) 因某人所犯的错误而受苦 (c) 耐心地等待某人做某事
(b) 紧紧拥抱某人 (d) 和一个不友善的人在一起

14. Whisked away
(a) 过分地或极端地做某事 (c) 对某事或某物采取冷淡或随意的态度来忽视
(b) 从争论中脱身而出来避免进一步的争 (d) 把某人迅速地从某处带到另一处

15. Talk some sense into someone
(a) 说服某人以合理的方式行事 (c) 详细地解释某事
(b) 说服某人承担责任 (d) 鼓励某人尝试一个新的活动

16. Beg to differ
(a) 改变某人的观点 (c) 请求一个短暂的休息时间
(b) 将某物储存在不同的容器中 (d) 坚决反对某人

17. Chip in
(a) 不管结果如何都让某事发生 (c) 折断或折成小段
(b) 捐钱购买或支付某物 (d) 在正餐之前吃点零食

18. Work something out with someone
(a) 以妥协来达成协议 (c) 找出数学题的答案
(b) 引导某物通过障碍 (d) 在非常短的时间内准备某事

Supporting Information-S3: Test Stimuli for Lexicosemantic Judgement Task

1. On the same page

- Appropriate: We need to be on the same page before we move forward with the project.
- Inappropriate: Please put the box on the same page as the book.

2. Turn a profit

- Appropriate: The company expects to turn a profit in the next year.
- Inappropriate: She decided to turn a profit on the floor.

3. Have something on someone's hands

- Appropriate: I can't come to the party tonight because I have a lot of work on my hands.
- Inappropriate: I can help you because I have a lot of work on my hands.

4. Tighten up

- Appropriate: We need to tighten up security to protect ourselves.
- Inappropriate: He wanted to tighten up the song.

5. Kill someone with kindness

- Appropriate: She decided to kill him with kindness by always being nice.
- Inappropriate: The police were permitted to kill the criminal with kindness.

6. Root for

- Appropriate: My brother was rooting for his sister to win.
- Inappropriate: She tried to root for the book.

7. Someone's hands are tied

- Appropriate: I'd love to help, but my hands are tied.
- Inappropriate: She said her hands were tied, so she could help.

8. Slippery slope

- Appropriate: Doing that could lead us down a slippery slope.
- Inappropriate: The plan was on a slippery slope as it became better and better.

9. Put someone on the spot

- Appropriate: I hate when my teacher puts me on the spot in front of my friends.
- Inappropriate: The camera put him on the spot because it was too small.

10. Talk some sense into someone

- Appropriate: I tried to talk some sense into him, but he wouldn't listen.

- Inappropriate: The book talked some sense into her and she changed her mind.

11. Beg to differ

- Appropriate: I beg to differ, as I believe your argument is wrong.
- Inappropriate: She decided to beg to differ because she ran out of money.

12. Whisked away

- Appropriate: The couple was whisked away on a nice weekend.
- Inappropriate: The tree was whisked away from the bed and placed on the street.

13. Chip in

- Appropriate: They all decided to chip in to help pay for the food.
- Inappropriate: The dog decided to chip in.

14. Bear with someone

- Appropriate: Please bear with me while I try to work this problem out.
- Inappropriate: She decided to bear with the cake until it was finished.

15. Go through a rough patch

- Appropriate: They went through a rough patch in their relationship but finally worked things out.
- Inappropriate: Our friendship is going through a wonderful rough patch.

16. Let up on someone

- Appropriate: The teacher decided to let up on the students after their hard work in practice.
- Inappropriate: My mother always lets up on me by asking me to try harder.

17. Look out for someone

- Appropriate: Can you please look out for my little brother while I'm at work?
- Inappropriate: I always look out for my coffee in the morning.

18. Work something out

- Appropriate: They were able to work this problem out.
- Inappropriate: I'm trying to work my forehead out.