9325 words

Declarative and Automatized Phonological Vocabulary Knowledge in L2 Listening Proficiency: A Training Study

Kazuya Saito,[1] Takumi Uchihara,[2] Kotaro Takizawa,[3] and Yui Suzukida[1]

[1]University College London
[2]Tohoku University
[3]Waseda 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

We extend our gratitude to all participants for their involvement and to the *Applied Psycholinguistics* reviewers and associate editor, Ethan Kutlu, for their constructive feedback. The project was supported by the Leverhulme Trust Research Grant (RPG-2024-391) and the UK-ISPF Research Grant (1185702223), both awarded to the corresponding author (KS).

Abstract

By adopting a pre- and post-test design, the current study longitudinally examined the complex relationship between two different dimensions of phonological vocabulary knowledge (declarative vs. automatized) and their ultimate impacts on global L2 listening proficiency among 133 Japanese EFL students. The declarative group focused solely on what target words sound like and mean via meaning recognition tasks. The automatization group worked not only on such form-meaning mappings but also on prompt access to the target words in a semantically, collocationally, and grammatically appropriate manner via lexicosemantic judgement tasks. Compared to the declarative group, the automatization group showed relatively robust learning in both declarative and automatized dimensions of target words. Although neither training approach showed clear superiority, the results suggest that relative gains in automatized, rather than declarative, dimensions are associated with enhanced L2 listening proficiency. The distinction between declarative and automatized dimensions of phonological vocabulary knowledge, along with the absence of a direct link between training type and improved listening proficiency, offers valuable insights for future extension studies.

Key words: *Second Language Vocabulary, Listening, Automatization, Instructed Second Language Learning*

Introduction

Few disagree that attaining advanced listening comprehension skills plays a critical role in successful L2 communication in academic, personal, and business settings in a globalized society (Vandergrift, 2007). To date, researchers have extensively examined the factors that explain the incidence of L2 learners with such advanced L2 listening proficiency and which factors should be prioritized in syllabi to help learners achieve this goal optimally. While those who score higher in general L2 listening proficiency tests likely access a range of listening strategies (Vandergrift & Goh, 2012) and possess greater cognitive abilities (e.g., Linck et al., 2013 for working and long-term memory), there is substantial research evidence showing that the primary determining factor is learners' phonological vocabulary knowledge. This explains the largest amount of variance in L2 listening test outcomes (for meta-analysis, see Zhang & Zhang, 2022).

Importantly, existing literature has exclusively focused on the *declarative* dimensions of phonological vocabulary knowledge (i.e., mapping what a word sounds like and means), typically measured via translation or multiple-choice tasks such as meaning recognition tests (MRT). In line with the skill acquisition theory for instructed L2 learning (DeKeyser, 2017; Suzuki, 2023), there is growing attention towards defining, assessing, and distinguishing the *automatized* dimensions of phonological vocabulary knowledge (i.e., accurate, fast, and stable access to words in context), possibly measured via lexicosemantic judgement tests (LJT). Cross-sectional evidence supports the view that declarative and automatized phonological vocabulary could be overlapping but essentially distinct constructs (Saito et al., 2023; Uchihara et al., 2024) and that different experience and cognitive variables anchor the development of declarative and automatized phonological vocabulary (e.g., working memory and the length of L2 learning for declarative dimensions vs. study-abroad and extracurricular activities for automatization; Saito & Uchihara, 2024).

To take a longitudinal look at this topic, the current investigation explores how a total of 133 Japanese English-as-a-Foreign-Language students develop declarative and automatized phonological vocabulary knowledge (measured via MRT and LJT) when they receive two different types of training (MR tasks with and without LJ tasks). We predict that a combination of MR and LJ training can help learners enhance both declarative and automatized phonological vocabulary knowledge, and subsequently improve their global L2 listening proficiency. However, the effectiveness of MR training alone may be limited to the declarative dimensions of phonological vocabulary knowledge.

Background Literature

Roles of Vocabulary Knowledge in Real-Life L2 Listening Proficiency

L2 listening is a complex and demanding task that requires a combination of linguistic and non-linguistic knowledge. Unlike L1 listening, where listeners can rely on automatic, subconscious, and effortless access to their native language system, L2 listening involves the use of L2 representations and processing skills that learners are gradually, partially, and concurrently developing (Vandergrift, 2007). For bottom-up processing, learners first need to detect prosodic structures in order to segment the speech stream into words (Cutler, Dahan, & Van Donselaar, 1997). They further rely on segmental details of words to distinguish phonologically similar words, such as those with higher lexical density (e.g., Bradlow & Pisoni, 1999) and minimal pairs (Munro & Derwing, 2006), and attend to the morphological attributes of each word to comprehend the grammatical patterns in a sentence. Additionally, L2 learners need to understand a speaker's intention in accordance with conversational, societal, and cultural backgrounds (Taguchi, 2011) and extralinguistic factors, such as vocal tone, facial cues, and bodily gestures (Kamiya, 2022).

Regarding top-down processing, certain listeners can not only use a range of strategies, such as attending to the main points of L2 listening discourse, evaluating what they have understood versus what they have not, and inferring meaning from context, but also have high-level awareness of their own listening abilities and available strategies (Vandergrift & Goh, 2012). Successful L2 comprehension is greatly related to more precise auditory processing for better capturing of acoustic details of speech (Vandergrift & Baker, 2015) and greater working memory for holding and sustaining linguistic information for further linguistic analyses (Linck et al., 2013).

While all the abilities above are essential for the successful comprehension of L2 speech, ample studies have examined what factors account for the variances in global L2 listening proficiency, typically measured via high-stakes L2 listening proficiency tests (e.g., IELTS, TOEFL). These studies have consistently shown that learners' vocabulary knowledge could be by far the most important correlate of successful L2 listening (r = .5-.7; see Zhang & Zhang, 2020 for a meta-analysis) and that developing robust L2 vocabulary thus needs to be prioritized in an L2 listening syllabus (e.g., Stæhr, 2009; Vafaee & Suzuki, 2020; Vandergrift & Baker, 2015; Wallace, 2022).

When we look at the existing literature, such vocabulary knowledge has been exclusively assessed via tests designed to tap into learners' form-meaning mappings, such as meaning recognition (Milton & Hopkins, 2006; McLean et al., 2015) and meaning recall (Cheng et al., 2022; González-Fernández & Schmitt, 2020). According to Nation's (2013) widely-cited framework, word knowledge essential for successful L2 listening can be considered to involve not only recognizing the sound and meaning of target words (form-meaning mapping) but also understanding how these words interact with other words in semantically, collocationally, and grammatically appropriate ways (use-in-context). The former pertains to recognition of single or two-to-three words, while the latter highlights the ability to comprehend words on broader levels (e.g., clauses, sentences). While much discussion concerns the measurements and training of form-meaning mappings (e.g., recognition vs. recall), surprisingly little is known about how to define, measure, and train the use-in-context aspects of word knowledge (Schmitt, 2019).

From a developmental perspective, the skill acquisition account of instructed L2 learning distinguishes three key types of knowledge: declarative, procedural, and automatized. These types reflect different stages of L2 phonological vocabulary learning, as outlined below:

- **Declarative knowledge** involves learners' explicit associations between word forms and meanings (knowing *what*). When a target word (or word strings) is presented in isolation, learners can recognize it in a list (McLean et al., 2015) and/or recall its definition (Cheng et al., 2022). However, this does not necessarily mean they can use the word appropriately in more global contexts (e.g., efficient retrieval of word meanings at the sentence level). The development of declarative knowledge can be achieved through explicit, form-focused, and controlled practice activities (e.g., multiple-choice quizzes, flashcards; Nakata et al., 2021).
- **Procedural knowledge** refers to the ability to use vocabulary in a context-appropriate manner without thinking about rules or meanings so explicitly (knowing *how*). It represents a gradual shift from more to less conscious and effortful processing in retrieving word meanings. Procedural knowledge develops through practice and involves using vocabulary in more extended, sentence-level contexts during listening (and speaking) activities (Marai, 2019).
- **Automatized knowledge** corresponds to the most advanced stage, where learners' use of vocabulary (i.e., meaning retrieval) becomes fast, stable, and effortless (i.e., automaticity). Achieving automatized knowledge is often considered the ultimate goal, particularly in communicatively authentic listening (and speaking) contexts. With repetitive practice, and through the proceduralisation of declarative knowledge via sustained exposure and engagement with target words in extended contexts, learners can eventually access this knowledge seamlessly in relation to surrounding words, forming chunks of linguistic constructions (Ellis, 2006).

By aligning Nation's framework of word knowledge with skill acquisition theory, form-meaning mapping is associated with the declarative dimension of phonological vocabulary, while use-in-context corresponds to the procedural and automatized dimensions of phonological vocabulary. Since our conceptualisation of vocabulary learning focuses on achieving advanced L2 listening proficiency, both the proceduralisation and automatization of declarative knowledge involve the processing of vocabulary at the sentence level in this study. To avoid redundancy, we use "automatization" to encompass both proceduralisation and automatization, indicating that the ultimate goal of instructed L2 learning is the automatization of proceduralised routines (Suzuki, 2023). In the present investigation, we adopted these paradigms to develop assessment and training materials, employing meaning recognition tasks for the form-meaning/declarative dimension and lexicosemantic judgment tasks for the use-in-context/automatized dimension.

It is important to emphasize that neither Nation's word knowledge framework nor the skill acquisition theory advocates discarding simple form-meaning mapping activities. Instead, these activities, which focus on memorizing patterns, are recognized as an essential initial step (for the suggested sequence of form-meaning and meaningful exposure, see Schmitt, 2008 and Nation, 2013). While form-meaning activities are integral and included in our current training paradigms (see the Method section), it is crucial to provide follow-up support that helps learners proceduralize and automatize the knowledge gained from these activities. This enables learners to access the knowledge more accurately, fluently, and subconsciously at more global sentence levels.

Measurement and Training of Automatized (vs. Declarative) L2 Knowledge

Within the skill acquisition framework, a range of single-modal tests have been adopted such as multiple-choice and fill-in-the-blanks, to assess participants' declarative dimensions of morphosyntactic knowledge. In such tests, participants can fully focus on their abilities to comprehend and use only target structures (e.g., past tense markers, articles). In contrast, to assess automatized knowledge, *dual-modal* tests have been devised to examine how L2 learners can access the target structures accurately, promptly, and subconsiciously. In such test formats, participants' specific linguistic knowledge is tested while at the same time they are using language for meaning conveyance. One most oft-used test is grammaticality judgement tests (GJTs).

In GJTs, learners read or listen to sentences featuring manipulated target morphosyntactic structures and must quickly and intuitively decide the correctness of these sentences. While their main test is to grasp the overall meaning of the sentences (i.e., reading or listening processing), they simultaneously need to check whether any grammatical features are used correctly or incorrectly in any parts of the sentences (i.e., morphosyntax processing; see Plonsky et al., 2020 for a review). Ample evidence shows that when L2 learners take metalinguistic tests and GJTs, their test scores cluster into two different factors. This suggests that metalinguistic tests and GJTs measure fundamentally different aspects of L2 morphosyntactic proficiency, with GJTs being assumed to capture automatized knowledge and metalinguistic tests reflecting declarative knowledge (Ellis, 2005; Gutiérrez, 2013).

While highly form-oriented instruction impacts only declarative knowledge, L2 learners can attain both declarative and automatized knowledge when they learn morphosyntactic accuracy from form-oriented lessons and also practice it repetitively via meaning-oriented instruction (Spada & Tomita, 2010). Additionally, performance on GJTs has been shown to correlate with key variables affecting high-level L2 proficiency, such as age of acquisition (Abrahamsson & Hyltenstam, 2009) and length of immersion (Faretta-Stutenberg & Morgan-Short, 2018).

According to Suzuki and Elgort's (2023) comprehensive review on automatized dimensions of various linguistic skills and training effects, much has been documented about L2 morphosyntax acquisition and about L2 phonetic acquisition to a lesser degree (cf. Saito & Plonsky, 2019). However, very few studies have examined the automatized dimensions of phonological vocabulary knowledge (the main focus of the current investigation).

Elgort (2011) investigated how learners can automatize their knowledge of pseudowords through intentional training involving meaning recognition and feedback. The study used lexical decision tests in which participants were shown strings of letters and asked to make an intuitive judgment about whether each string was a word or a non-word. The results showed that participants could recognize target words more quickly and accurately after being exposed to similar forms with related semantic meanings. These findings highlight the effectiveness of deliberate learning strategies in enhancing automatic lexical processing. For insights into the incidental approach to automaticity in lexical knowledge, see Elgort and Warren (2014).

Foster et al. (2014) employed a unique approach to examining L2 learners' highly advanced, automatized collocational knowledge. A total of 160 experienced Polish users of English (in the UK and Poland) were asked to read narratives and highlight non-nativelike word combinations embedded in the text (e.g., "a young man was strolling his way in the street"). The methodological rationale was that given that advanced L2 learners are assumed to store a range of multiword units as lexical chunks, the degree of nativelikeness in their collocational knowledge can be clearly observed when these word combinations *deviate* from what they are accustomed to. The results indicated that more advanced L2 participants, particularly those with an earlier age of arrival, demonstrated greater accuracy in identifying non-nativelike word selections.

Building on the methodological paradigm underlying GJTs, scholars have introduced the Lexicosemantic Judgement Test (LJT) to assess automatized phonological vocabulary knowledge (Saito et al., 2023; Uchihara et al., 2024). In the LJT, learners are aurally presented with grammatically correct sentences composed solely of high-frequency words. These sentences are categorized either "semantically appropriate" or "semantically inappropriate" based on the use of a target word (e.g., "estate"). Some sentences use the target word appropriately (e.g., "My grandfather bought an estate"), while others use it in a semantically incongruous manner (e.g., "My friend's estate was very kind"). Learners are asked to judge the semantic appropriateness of each sentence upon hearing it, with limited time to deliberate on the correct or incorrect use of the word.

In previous projects, over 400 Japanese L2 learners of English with varied experience and proficiency profiles took a general L2 listening proficiency test (TOEIC) as well as a set of tests designed to measure both declarative (meaning recognition, meaning recall) and automatized dimensions of phonological vocabulary knowledge (LJT). The results showed that their vocabulary test scores were factored into two latent variables (meaning recognition and recall being distinguishable from LJT). This led us to speculate (a) that LJT taps into a construct of lexical knowledge fundamentally different from declarative phonological knowledge (typically measured via meaning recognition and recall) and (b) that the construct could be the automatized dimension of phonological knowledge (Uchihara et al., 2024).

Furthermore, the results also indicated that the learner factors predicting the attainment of L2 phonological vocabulary knowledge differed depending on the test formats. Those with more extensive L2 learning experience in classroom settings and greater working memory likely showed greater declarative phonological vocabulary knowledge (assessed via meaning recognition). In contrast, those who engaged in more extracurricular activities and exposure to communicatively authentic materials beyond classrooms (e.g., study abroad, viewing, conversations) likely showed greater automatized vocabulary knowledge (assessed via LJT; Saito & Uchihara, 2024). Finally, global L2 listening proficiency was more strongly predicted by LJT scores (r = .6-.7) than by scores for meaning recognition and recall (r = .4-.5; Saito et al., 2023).

The accumulated cross-sectional evidence has thus far suggested that there are two overlapping but essentially different dimensions of phonological vocabulary knowledge: form-meaning mappings and use-in-context. These dimensions can be measured via declarative vocabulary tests (meaning recognition) and automatized vocabulary tests (LJT). From the skill acquisition perspective, vocabulary learning progresses from explicit, rule-based knowledge (declarative) to fluent, context-appropriate use (procedural), and ultimately to automatic, effortless application (automatized) as learners practice and reinforce their skills (DeKeyser, 2017; Suzuki, 2023). Given the relative importance of vocabulary knowledge in global L2 listening proficiency (Wallace, 2022), the attainment of automatized vocabulary knowledge can help learners access words more spontaneously in real-life contexts, thereby enhancing their global L2 listening proficiency in the long run. By conducting an intervention study with a pre- and post-test design, the current study took the first step toward testing this hypothesis.

Enhancing Global Listening Proficiency via Vocabulary Training

To help learners attain L2 listening proficiency, numerous studies have focused on top-down approaches via metacognitive strategy training. In this training, students are encouraged to engage in strategic planning prior to listening through an awareness of their available strategies, self-monitoring during listening, and reflecting on post-listening outcomes (e.g., Kobayashi, 2018; Milliner & Dimoski, 2024; Yeldham, 2022). A meta-analysis by Dalman and Plonsky (2024) has confirmed that such studies can eventually impact L2 learners' global listening proficiency with medium effects ($d = 0.69$). Surprisingly, very few studies have explored how teaching phonological forms of a set of key vocabulary items can first impact vocabulary proficiency and then global L2 listening proficiency.

Zhang and Graham (2020) provided different types of phonological vocabulary training (e.g., explanations in L1 vs. L2) to a total of 137 young Chinese learners of English over the course of six weeks. The pre- and post-tests showed significant enhancement in their form-meaning mapping of the 60 target items introduced during the sessions (measured via meaning recall). Similarly, Uchihara et al. (2023) provided one hour of phonological vocabulary training on 40 target items to 80 Japanese college students of English. The pre- and post-tests (via picture naming test) demonstrated significant gains in form recall and prosodic accuracy after a brief training session. However, neither of these studies explored how such vocabulary gains ultimately impact L2 listening proficiency.

In Saito and Akiyama's (2018) longitudinal investigation, Japanese college students of English engaged in individual interaction with native interlocutors over the course of 12 weeks. Whenever any linguistic errors (including lexical ones) interfered with overall understanding, the native interlocutors provided corrective feedback to L2 participants' erroneous utterances and encouraged them to self-correct during conversational interactions with the aim of more easy-to-understand speech (i.e., negotiation for comprehensibility). The results showed that not only did the participants' L2 speech become more comprehensible and intelligible, but their global L2 listening proficiency (measured via TOEIC) also improved with large effects. Yet, the study did not isolate and link the participants' vocabulary development to their L2 listening proficiency development. The current study corresponded to these methodological concerns.

Current Study

To take a longitudinal look at the hierarchical relationship between declarative and automatized phonological vocabulary and global listening proficiency (form-meaning mappings of words → more appropriate, prompt, and stable access to words → global listening proficiency), the current study conducted an intervention study with a total of 133 Japanese learners of English with a pre- and post-test design. After taking the pre-tests (comprising vocabulary and global listening proficiency tests), they were divided into three groups—Automatization, Declarative, and Comparison, and engaged in a total of six sessions (30 minutes × 6 sessions = 3 hours). Those in the automatization group first worked on the form-meaning mapping of 80 target words via meaning recognition tasks (Sessions 1-3). Then, they shifted their attention to the use-in-context aspects of the target words via lexicosemantic judgment tasks (Sessions 4-6). As typical vocabulary training in many existing studies (e.g., Zhang & Graham, 2020) and many L2 classrooms all over the world, those in the declarative group learned the target words via meaning recognition tasks throughout the project (Sessions 1-6). Finally, those in the comparison group engaged in different types of EFL activities and took the pre- and post-tests. Their performance would serve as an indication of test-retest effects if any.

Although ample cross-sectional evidence suggests the critical roles of phonological vocabulary knowledge in L2 listening proficiency (e.g., Zhang & Zhang, 2022), the current study represents one of the first attempts to not only examine if training vocabulary can impact global L2 listening proficiency (Yeldham, 2022) but also determine what type of vocabulary knowledge—declarative vs. automatized—should be trained to lead to such gains in global L2 listening proficiency (Suzuki & Elgort, 2023). The two research questions and predictions were formulated as follows:

1. To what extent can different types of training (Automatization, Declarative, Comparison) differentially enhance the declarative and automatized dimensions of phonological vocabulary knowledge?

2. To what extent can the effects of different types of training be transferred to and lead to changes in higher-order, global L2 listening proficiency?

Regarding RQ1, both the automatization and declarative groups were expected to show comparable improvement in the declarative dimensions of the target words since both groups engaged in declarative training. However, with respect to the automatized dimensions of the target words, the automatization group was expected to outperform the declarative group. This is because the automatization group was encouraged to further reinforce their robust access to the target words and become more capable of using these words more accurately and promptly. For RQ2, we predicted that given the relative importance of declarative and automatized vocabulary knowledge in L2 listening proficiency (declarative < automatized; Saito et al., 2023; Uchihara et al., 2024), while both the declarative and automatization groups would enhance their global L2 listening proficiency, the gains would be larger for the automatization group than for the declarative group.

Method

All the research materials used in the current project have been deposited on the open science platform for researchers and teachers, L2 Speech Tools (Mora-Plaza et al., 2021).

Setup

Given the post-COVID context where health and safety were still prioritized for conducting large-scale training projects of this kind, a decision was made to implement the current investigation online using both Zoom (for general listening proficiency tests) and the online psychology experiment builder, Gorilla (Anwyl-Irvine et al., 2020; for training, pre-post tests, and questionnaires). We had ample similar research experience wherein approximately 10% attrition was achieved due to a range of precautionary measures (see below).

The project was advertised as a research-based listening training project at multiple universities across Tokyo, Japan. Once more than 200 interested Japanese college students contacted the research team, they were first invited to take both screening vocabulary and working memory tests in Gorilla. They were asked to use headphones for clearer audio without external noise and to ensure a stable internet connection on their computers. The vocabulary test was designed to select those who met the threshold for processing minimum levels of L2 comprehension. The test comprised recognition of the first 1,000 word families in the BNC-COCA corpus as per Vocab Profilers (Cobb, 2000). If participants' accuracy rate was below 80%, they were excluded (for a similar decision, see Dang, Webb, & Coxhead, 2020). The purpose of the working memory test was to check if they had the necessary computer skills to implement online research experiments without any technological issues. Participants who failed to type any numbers during the working memory tasks or/and who failed to recollect at least four numbers in the forward digit span tasks were eliminated.

Only after we confirmed that participants had successfully completed the preliminary tests, they were given a detailed instruction handout in Japanese to understand the rest of the experiment procedure (pre-test, training, and post-test). First, a group of 10-20 participants joined the pre-test sessions with a trained Japanese research assistant via Zoom. After the assistant once again explained the test procedure in Japanese followed by a Q&A session, they took a general listening proficiency test (TOEIC Version A) and then moved on to a battery of vocabulary tests (LJT and two different versions of aural MRT in this order), aptitude tests (tapping into auditory processing), and a bio questionnaire (surveying EFL experience and listening metacognition). The results of the auditory processing tests were not reported in this study (but will be reported elsewhere).

Subsequently, those who successfully completed the pre-test sessions were randomly assigned to one of the three group conditions (Automatized, Declarative, and Comparison) and proceeded to a total of six 30-minute training sessions via Gorilla. The length of the training (30 minutes of explicit vocabulary training on partially-known 80 items × 6 times) was determined based on similar previous studies aimed at the automatization of newly learned words (e.g., Elgort, 2011, for 5-6 sessions for the acquisition of 48 pseudowords). Once they received a URL link from a trained research assistant, they were given 24 hours to complete each session. Their progress was monitored by the assistant. If there was more than a one-day delay, these participants were eliminated from the project. To ensure participants receive the training in a regular pattern, the upcoming training sessions were locked until they were ready for the next scheduled training. One day after the completion of the final session, the participants were invited to join the post-test sessions. They took a different version of the general L2 listening proficiency test (TOEIC Version B). Then, they moved on to a set of vocabulary tests (LJT, MRT) and the exit questionnaires in Gorilla. A total of 133 participants completed the entire project and were included in the final analyses

Participants

The participants (N = 133) comprised 87 females and 46 males (M age = 23.5 years; Range = 19-28 years). Based on the results of the general L2 listening proficiency tests at pre-test (TOEIC Version A), their proficiency levels could be considered Upper-Basic (A2) to Independent (B1, B2) according to the Common European Framework of Reference (CEFR). The participants' previous EFL learning backgrounds varied widely in terms of age of learning onset (M = 10.4 years, SD = 4.5, Range = 3-14) and length of learning (M = 1619.0 hours, SD = 521.4, Range = 600-3150).

Target Words

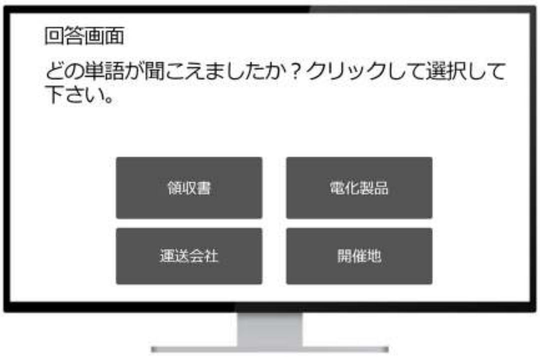
For the pre- and post-tests and training, a total of 80 target words were chosen to reflect real-life L2 vocabulary use. A speech corpus was developed using scripts from the TOEIC Listening test (Version B) for the post-tests, as the TOEIC test includes various forms of L2 discourse such as short sentences, conversations, and monologues. From this corpus, which contained 2,731 tokens, the top 80 words posing the most phonological challenges for Japanese EFL learners were carefully selected. These words were chosen based on their lower frequency profiles according to the BNC/COCA word family lists (Nation, 2012). Priority was given to words that Japanese learners might find difficult, including iambic words with multiple syllables, challenging segmentals like English [r] and [l], and consonant clusters (Saito, 2014). Additionally, loanwords were excluded as they could potentially facilitate L2 understanding (Uchihara et al., 2022).

Outcome Measures

In the current project, participants practiced on the 80 target words (derived from TOEIC Test Version B at post-tests). Whereas the impact of vocabulary training on participants' general L2 listening proficiency was evaluated based on their performance on TOEIC Test Version B, their pre-existing L2 listening proficiency was measured in a different version of TOEIC (Version A) at pre-tests. To examine the extent to which participants had enhanced the declarative and automatized dimensions of the target words, they took both meaning recognition tests (MRT) and lexicosemantic judgement tests (LJTs). Since the participants took the same format of the test for global L2 listening proficiency (TOEIC) and the same materials for the vocabulary tests (MRT, LJT) at pre- and post-tests, the performance of the comparison group served the presence of test-retest effects if any. The participants took the tests in the following order—(1) general L2 listening proficiency test (TOEIC), (2) lexicosemantic judgement test (LJT), and (3) meaning recognition test (MRT). All the test instructions were prepared in Japanese to facilitate participants' understanding of the procedures. To explain both methodological and conceptual details of the tests in a logically reasonable manner, the tests were described as follows in the following order: (3) MRT → (2) LJT → (1) TOEIC. The test materials are deposited in L2 Speech Tools (Mora-Plaza et al., 2022). The online versions of the tests are available as open materials in Gorilla and shared in **Supporting Information S1**.

Meaning Recognition Test

As operationalized in existing studies (e.g., McLean et al., 2015), the declarative dimension of participants' phonological vocabulary knowledge (i.e., form-meaning mapping of the target words) was measured via two different audio versions of MRT. One key aspect of form-meaning mapping in L2 speech development concerns generalization. Once learners become capable of recognizing words within training, they are believed to start generalizing their phonological representations from the contexts of trained talkers to other talker contexts (i.e., more real-life L2 listening contexts; Zhang, Cheng, & Zhang, 2021). To capture the robustness of generalization, participants took two versions of MRT, one featuring a trained voice (a male of General American; Talker A) and the other featuring an untrained voice (a female of General American; Talker B). A total of 160 trials were featured in MRT (n = 80 target words produced by Talkers A and B, respectively). In the Gorilla platform, the 160 stimuli were played in a randomized order. As shown in Figure 1, as soon as they listened to a target word, the participants were asked to choose the correct meaning from four options (one correct answer and three distractors). All four answer options were presented in Japanese and they served the same part of speech. The distractors were selected from a list of words frequently found in TOEIC test materials. They were also matched in terms of same parts of speech. Total scores were calculated out of a maximum of 160 points. The pre- and post-test MRT materials demonstrated high reliability, α = .93 and .94, respectively.



回答画面

どの単語が聞こえましたか？クリックして選択して下さい。

Figure 1 Screenshot of the Meaning Recognition Test (in Japanese). Participants listened to a target word without spelling (e.g., "appliance") and selected its meaning from four different

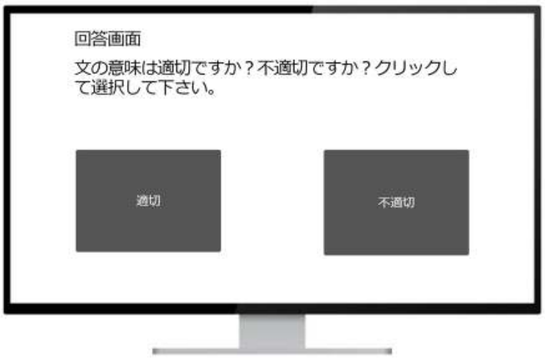
choices (e.g., “領収書” [receipt], “電化製品” [appliance], “運送会社” [carriers], “開催地” [venue]).

Lexicosemantic Judgement Test

Following the precursor projects on the development and validation of the test designed to assess the automatized dimension of phonological vocabulary knowledge (Saito et al., 2023; Uchihara et al., 2024), the same lexicosemantic judgement test was adopted in the current study. A total of 160 short sentences were prepared. Half of the sentences used the target words in a semantically appropriate manner, while the other half used them in a semantically inappropriate manner. For instance, for the word “estate,” participants heard “My grandfather bought an estate” (appropriate) and “My friend’s estate was very kind” (inappropriate). All other parts of the sentences were clear and easy to understand (comprising 1K word families)

The sentences were kept brief, ranging from 3 to 8 words, to minimize the strain on participants’ working memory during the test. To ensure that the test did not mainly evaluate syntactic knowledge or speech perception skills, we used simple syntax without complex subordination. Given that syntactic position can influence the salience of unfamiliar words, efforts were made to minimise such priming effects. First, we avoided placing target words at the beginning of sentences, as words appearing at the beginning or end of a sentence are typically more salient than those embedded in the middle (Moravcsik & Healy, 1998). Second, we avoided using subordinate clauses, as syntactic complexity affects the salience of words; words in main clauses are generally more salient than those in subordinate clauses (McKoon et al., 1993). Finally, all sentences were simple (without phrase boundaries) to prevent syntactic complexity from reducing word salience, as complex structures can lead to shorter eye fixations or cause words to be skipped altogether during sentence processing (Birch & Rayner, 2010). After drafting the sentences, three English experts reviewed them to ensure they conveyed the intended semantic appropriateness or inappropriateness. Participants listened to 160 short sentences spoken by the untrained talker (Talker B).

After hearing each sentence, they were asked to determine whether it was “semantically appropriate” or “semantically inappropriate” based on a single word (for a screenshot of the test, see Figure 2). To ensure participants listened to the entire sentence, the target word was placed somewhere other than at the beginning. The sentences were kept simple, ranging from 4 to 8 words in length, and were always grammatically correct. Most of the words (93%) came from the 1,000 most common words in English. As a measure of automatized vocabulary knowledge, participants were tested on their ability to process the conceptually appropriate (or inappropriate) use of target words produced by the new talker (Talker B) while at the same time processing other aspects of language on a sentence level (e.g., phonological, morphosyntactic, and pragmatic processing). The 160 sentences were then played to participants in a random order. Participants earned 1 point for each correctly judged sentence, with a maximum possible score of 160 points. The pre- and post-test LJT materials demonstrated high reliability, α = .92 and .90, respectively.



回答画面

文の意味は適切ですか？不適切ですか？クリックして選択して下さい。

Figure 2 Screenshot of the Lexicosemantic Judgement Test (in Japanese). Participants listened to a sentence with an embedded target word (without captions) and determined whether the sentence was semantically appropriate or inappropriate.

In the current investigation, consistent with the earlier validation studies demonstrating the strong predictive power of the LJT for L2 listening proficiency (Saito et al., 2023; Uchihara et al., 2024), we opted to use the untimed version of the LJT. Participants were informed that the test assessed both accurate and fluent processing of L2 English sentences, and they were encouraged to respond as quickly as possible, though no time limit was imposed for each trial. This approach was justified on both empirical and theoretical grounds.

Turning to the existing literature on L2 morphosyntax, researchers generally acknowledge that acceptability judgment tasks (GJTs) can measure the degree of automatization in L2 morphosyntactic knowledge. However, the role of time pressure remains a subject of debate. Some scholars advocate for timed tasks to better access automatized, potentially implicit, L2 knowledge by minimizing explicit, controlled, and monitored processing (e.g., Ellis, 2005; Gutiérrez, 2013). However, Plonsky et al.'s (2020) comprehensive review of GJT methodologies highlights that the effects of time pressure are particularly evident in *written* formats, where learners can revisit texts multiple times, which may lead to more deliberate, controlled processing. In contrast, such effects are not deemed necessary in *auditory* tasks, where participants hear a stimulus only once. The fleeting, continuous nature of auditory tasks inherently simulates real-life, relatively automatic language processing, reducing the need for external time constraints. As Plonsky et al. (2020) rightly pointed out, "by definition, aural language happens at a fixed pace for all participants, so is, in a sense 'timed', and so, generally with these comparisons between aural and written, 'modality' is inevitably conflated with timing" (p. 598).

It is important to note that existing research suggests time pressure may not reliably distinguish between controlled and automatic processing across all learners, raising concerns about the use of time-limited tasks as a sole measure of automatization. For instance, Hulstijn et al. (2009) argue that faster reaction times under time pressure do not necessarily indicate automatization but could instead reflect more efficient retrieval without full automaticity. This makes it challenging to differentiate true automatization from mere speed enhancement even if both reaction time and coefficient of variation are considered. Additionally, Maie and Godfroid (2022) demonstrated that time pressure affects both controlled and automatic processes, particularly for L2 learners with slower lexical decoding and processing speeds, complicating its use as a reliable measure of automatization. Learners with processing difficulties may respond more slowly regardless of time pressure. Their eye-tracking study showed that time pressure might be more closely related to participants' lexical decoding skills and may not function as clearly as previously thought in distinguishing controlled from automatic processes, especially for learners with varying cognitive profiles.

To prepare for the current investigation and to strengthen the construct validity of LJT as a measure of automatized phonological vocabulary knowledge, we conducted an additional study to further examine the role of time pressure (Saito et al., forthcoming). While many GJT studies ambiguously set time limits based on native speaker norms, we carefully established an optimal response speed based on the LJT performance of advanced Japanese learners of English (CEFR = C1, C2). Although considerably slower than native speakers (600-700 ms), these advanced L2 learners were considered to have fully automatized phonological vocabulary knowledge, and their processing speed (2000-2400 ms) served as a benchmark for the university-level Japanese students in this study, who had relatively basic-to-intermediate L2 proficiency levels (CEFR = A2, B1, B2). Our findings indicated that participants' timed and untimed LJT scores similarly predicted their L2 listening proficiency ($r = .64$ and $.73$, respectively), with the difference in correlation coefficients failing to reach statistical significance ($z = 1.081$, $p = .140$). This lack of significant time pressure effects suggests that the construct of LJT as a measure of automatized phonological vocabulary knowledge is not substantially influenced by time constraints.

In essence, we find the timed approach problematic because enforcing a fixed time limit may disadvantage learners with weaker lexical decoding speeds (Maie & Godfroid, 2022) and may not accurately reflect the complex processes underlying the automatization of L2 knowledge, where the development of accuracy does not necessarily align with the development of fast and stable processing skills (Hulstijn et al., 2009). Therefore, we emphasized to participants the importance of both accuracy and fluency in the LJT, encouraging them to respond as quickly as possible without imposing specific time limits. To ensure participants adhered to these instructions and to identify those who might have spent excessive time responding to each stimulus, we defined slow outliers as individuals with an average reaction time of 6000 ms or more. This threshold was calculated as 1.5 standard deviations beyond the mean reaction time among the similar population—i.e., $n = 126$ Japanese EFL participants in Saito et al. (2023). None of the participants reached this threshold, and thus no data were excluded on this basis.

General L2 Listening Proficiency Test

Similar to previous studies (Cheng et al., 2022; Hamada & Yanagawa, 2023; Matthews et al., 2023; McLean et al., 2015), participants' general L2 listening proficiency was measured using the TOEIC test. Two different versions of the TOEIC test (Versions A and B) were adopted from the New Official Workbook provided by the Educational Testing Service in Volume 4. Version A was used to measure participants' pre-existing L2 listening proficiency. Version B was used to assess the impact of vocabulary training on L2 listening proficiency. As mentioned above (see the Target Word section), the 80 words identified as posing the greatest phonological challenges for Japanese EFL learners were carefully selected from the Version B TOEIC test and used as target words for training.

Each test comprised three different parts. Part 1 (30 items) asked participants to choose the best response from three options for single-sentence questions (5-10 words per sentence). Part 2 required participants to listen to a conversation between a male and a female speaker (consisting of 80-100 words) and then answer three comprehension questions by selecting the most appropriate response from four options. Part 3 involved listening to a business announcement delivered by a single speaker (80-100 words) and answering three comprehension questions by choosing the best response from four options. Participants' test scores were initially calculated as per Parts 1, 2, and 3, respectively (max scores = 30 points). High reliability was reported for Version A (α = .91) and Version B (α = .94).

Training

The participants were randomly assigned to one of the three training conditions—Automatization, Declarative, and Comparison. The training was implemented via Gorilla with a 24-hour timer. If participants did not complete each session within 24 hours, they would be eliminated from the project. As such, the timing they can access the training was strictly controlled by the researchers. The length/number of training sessions (30-minute × 6 sessions) was determined in accordance with previous similar studies (e.g., Elgort, 2011). To facilitate participants' understanding of the training procedure, a Japanese brochure was provided prior to the training, all the training instructions were displayed in Japanese, and research assistants were on standby to respond to any queries in Japanese. The online versions of the training materials are available as open materials in Gorilla and shared in **Supporting Information S2**. The content of training for the experimental groups (Automatization, Declarative) was summarized in Table 1 and detailed as follows.

Table 1 Summary of Automatization and Declarative Training

	Automatization Training	Declarative Training
Session 1	MRT + MRT (lenient timeout)	MRT + MRT (lenient timeout)
Session 2	MRT + MRT (lenient timeout) + MRT (mild timeout)	MRT + MRT (lenient timeout) + MRT (mild timeout)
Session 3	MRT + MRT (lenient timeout) + MRT (mild timeout)	MRT + MRT (lenient timeout) + MRT (mild timeout)
Session 4	LJT	MRT + MRT (lenient timeout) + MRT (mild timeout)
Session 5	LJT	MRT + MRT (lenient timeout) + MRT (mild timeout)
Session 6	LJT	MRT + MRT (lenient timeout) + MRT (mild timeout)

Notes. MRT for Meaning Recognition Task; LJT for Lexicosemantic Judgement Task; mild for 10.5 sec of lenient timeout; mild for 4.5 sec of timeout

Automatization Group

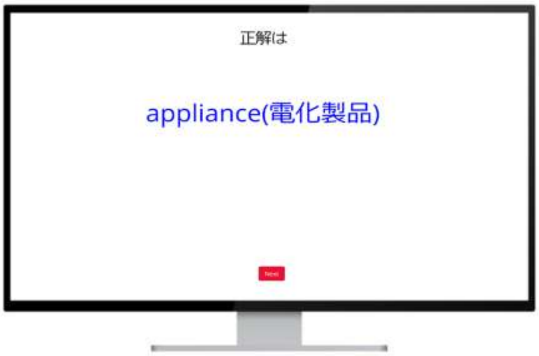
For the automatization group, the training program consisted of a total of six 30-minute sessions, with the first three sessions devoted to the development of declarative knowledge (Sessions 1-3) and the latter three sessions devoted to the development of automatized knowledge (Sessions 4-6).

Between Sessions 1 and 3, learners initially worked on form-meaning mappings of linguistic items. Following the notion of the skill acquisition theory for instructed L2 acquisition (DeKeyser, 2017; Suzuki, 2023), this process was promoted under a single task condition wherein they could fully focus on target areas of linguistic processing without much attention to other aspects of language. As operationalized in previous studies (Zhang & Graham, 2020), participants engaged in meaning recognition tasks to learn the 80 target words. Unlike the MRT in the pre- and post-tests, each trial in the training sessions was followed by feedback (see Figure 3).

Participants completed two rounds of MRT within each session, with certain time thresholds. In Session 1, participants engaged in the first round of MRT without any time limits, allowing them to take as much time as needed (i.e., no timeout). However, during the second round, a lenient timeout of 10.5 seconds was applied. In Sessions 2 and 3, participants again completed the first round of MRT without time constraints, followed by a second round with the lenient timeout (10.5 seconds) and a third round with a mild timeout (4.5 seconds).

To prepare for the current project, we conducted a series of pilot training sessions involving five experienced Japanese teachers of English, who engaged in meaning recognition and lexicosemantic judgment tasks under various timeout conditions. This helped us determine the optimal time windows for the participants in the current study (CEFR: A2, B1, B2). The purpose of implementing timeouts was to prevent L2 learners from spending excessive time on each trial, ensuring that each session could be completed within 30 minutes without undue delay or memory burden. Additionally, the timeouts aimed to facilitate a gradual shift from focusing solely on accuracy to increasing attention to fluency, thereby encouraging the proceduralization of any partially acquired vocabulary knowledge. Importantly, the goal of introducing timeouts was not to impose time pressure, as we consider the effects of time pressure problematic, particularly for cognitively diverse L2 learners. This rationale aligns with our adoption of an untimed approach, as previously justified.

To this end, we established 10.5 seconds as the lenient timeout and 4.5 seconds as the mild timeout. For the lenient timeout, our assumption was that none of the participants would fully use the 10.5 seconds; rather, the presence of the timeout would give them sufficient time to process each item, while also fostering an awareness that taking more than 10 seconds per item might not be ideal. Regarding the mild timeout, we assumed that 4.5 seconds would provide just enough time for participants to process each stimulus, instilling a slight sense of urgency to prioritize fluency alongside accuracy. Notably, this timeout approach differs from traditional time pressure methods. We ensured that participants (CEFR: A2, B1, B2) would not feel pressured by time constraints during the training but would be encouraged to balance accuracy and fluency in their phonological vocabulary knowledge and processing.

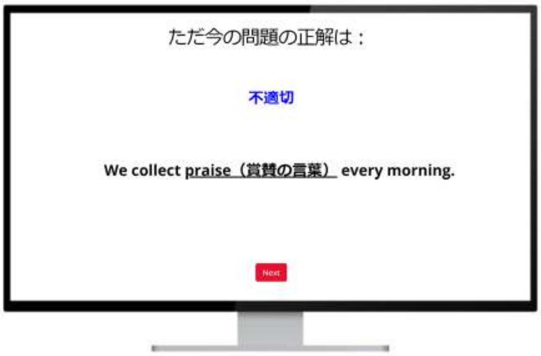


正解は

appliance(電化製品)

Figure 3 Screenshot of feedback in the Meaning Recognition training (in Japanese). The feedback displays the spelling of the target word (e.g., “appliance”) and its L1 translation (e.g., “電化製品”).

After participants established robust form-meaning mappings of target words, they proceeded to the automatization phase (Sessions 4-6). In each 30-minute session, the participants focused on the use-in-context aspects of the 80 learned vocabulary items via one round of lexicosemantic judgement tasks followed by feedback (160 trials). As stated in the skill acquisition theory (DeKeyser, 2017; Suzuki, 2023), automatization can be enhanced when learners engage in such dual task conditions. During the LJT, learners worked on accessing the target lexical items as a primary focus while using all other linguistic information (grammar, collocational, and pragmatic processing) at a sentence level. Each trial was followed by feedback (see Figure 4). Unlike MRT, participants did not have any timeouts.


ただ今の問題の正解は：

不適切

We collect praise（賞賛の言葉） every morning.

Figure 4 Screenshot of feedback in the Lexicosemantic Judgement training (in Japanese). The feedback indicates whether the sentence was semantically appropriate or inappropriate (e.g., "不適切" [inappropriate]) and provides the L1 translation for the target word in the sentence (e.g., "賞賛の言葉" [praise]).

Declarative Group

The participants worked only on the form-meaning mappings of the target words via MRT for all the sessions. In Session 1, they completed two rounds of MRT (without timeouts and with mild timeouts). In Sessions 2 to 6, they completed three rounds of MRT (with no, mild, and strict timeouts).

Comparison Group

The purpose of the comparison group was to determine if there was any test-retest effect resulting from taking the tests twice (MRT, LJT, TOEIC). As developed and piloted in another precursor project (Kachlicka et al., forthcoming), participants engaged in 30 minutes of adaptive vocabulary training. The training comprised picture matching tasks wherein participants read a target word prompt and chose a matching one out of four pictures. Word prompts and corresponding (and distractor) pictures were randomly selected from the word and picture banks. There were a total of 60 words randomly selected from each of the five vocabulary bands (*n* = 12 from 2000-, 3000-, 5000-, 10000-word-frequency families, and the academic vocabulary list, respectively) as set up in the BNC-COCA corpus in Vocab Profiler (Cobb, 2000). Within each session (30 minutes), participants worked on two rounds of picture matching of 60 words.

Results

The statistical analysis was performed using the R statistical environment (Version 4.3.2; R Core Team, 2024). To examine the extent to which the three groups of learners differentially improved different dimensions of L2 phonological vocabulary knowledge, the lme4 package was used (Bates et al., 2021) to construct a set of mixed effects models under different task conditions (MRT, LJT, and TOEIC, respectively). These models comprised participants' vocabulary performance as dependent variables relative to Group (Automatized, Declarative, Comparison) and Time (pre- vs. post-tests). Dimension was also featured as a predictor variable to check if participants' learning behaviors differed across different talkers in MRT (trained vs.

novel talkers) and different types of stimuli (semantically appropriate vs. inappropriate sentences). Where significant main or interaction effects were found, post-hoc multiple comparison analyses were conducted using the emmeans package by calculating and analyzing the estimated marginal means (Lenth et al., 2024). Cohen's d effect sizes were calculated and interpreted in accordance with Plonsky and Oswald's (2014) benchmarks for instructed L2 learning studies (d > 0.4 for small, > 0.7 for medium and > 1.0 for large effects).

Meaning Recognition Test

Descriptive statistics for the participants' MRT scores were summarized in Table 2 and visually plotted in Figure 5. Their MRT scores varied widely at the beginning of the project. After the training, the two experimental groups (Declarative, Automatized) demonstrated near-ceiling performance. To examine the extent to which the three groups (Automatized, Declarative, Comparison) differentially developed their MRT scores as per different talker dimensions, the following model was constructed: DV ~ group*time*dimension + (1|ID). Significant main effects of Time and Group (F = 30.450 and 962.412, p < .001) were found, but the effects of Dimension failed to reach statistical significance (F = 0.110, p = .540). The results indicated that participants successfully generalized their gains from the trained voice context (Talker 1) to the novel voice context (Talker 2). Interaction effects of Group and Time were significant (F = 130.374, p < .001). According to the results of multiple comparison analyses, the comparison group's performance showed some evidence of small test-retest effects (t = 2.46, p = .014, d = 0.69). Both the automatized and declarative groups significantly improved their performance over time (t = 31.25, 28.66, p < .001, d = 6.06, 5.45) and outperformed the comparison group at the time of post-tests with large effects (t = 18.88, 18.76, p < .001, d = 5.41, 5.38).

Table 2 Descriptive and Mixed Effects Analyses of Participants' Meaning Recognition Test Scores

A. Descriptive Results				
Group	Time/Dimension	*M*	*SE*	*95% CI*
Automatized	Pre/Trained	58.7	1.00	[56.7, 60.7]
	Pre/Untrained	58.7	1.00	[56.8, 60.7]
	Post/Trained	78.8	0.99	[76.8, 80.8]
	Post/Untrained	79.1	0.99	[77.1, 81.0]
Declarative	Pre/Trained	60.5	0.97	[58.5, 60.5]
	Pre/Untrained	60.8	0.97	[58.9, 62.7]
	Post/Trained	78.8	0.98	[76.9, 80.7]
	Post/Untrained	78.9	0.98	[76.9, 80.8]
Comparison	Pre/Trained	57.7	1.45	[54.8, 60.5]
	Pre/Untrained	57.8	1.45	[54.9, 60.6]
	Post/Trained	59.7	1.46	[56.8, 62.6]
	Post/Untrained	60.5	1.46	[57.6, 63.3]
B. Mixed Effects Modelling				
Fixed effects	*F*	*p*	$R^2_{conditional}$	$R^2_{mariginal}$

Group	30.450	< .001	.808	.205
Time	962.412	< .001		
Dimension	0.363	.540		
Group:Time	130.374	< .001		
Group:Dimension	0.033	.967		
Time:Dimension	0.061	.804		
Group:Time:Dimension	0.110	.895		

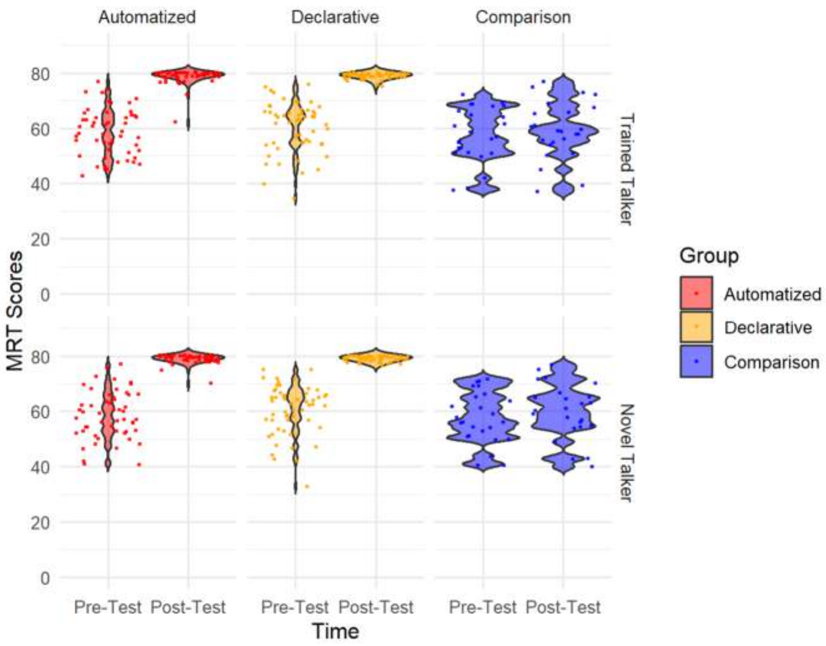


Figure 5 Participants' Meaning Recognition scores (out of 80 points) under two different talker conditions (trained vs. novel). While the comparison group's performance remained consistent between pre- and post-tests, the declarative and automatized groups showed significant improvement, reaching ceiling levels.

Lexicosemantic Judgement Test

As summarized in Table 3 and visually plotted in Figure 6, participants' LJT scores were substantially higher when tested for their ability to accept semantically appropriate stimuli compared to their ability to reject semantically inappropriate sentences. The mixed effect model found significant interaction effects of Group, Time, and Dimension (F = 5.012, p = .007). The results of post-hoc multiple comparison analyses revealed several patterns. First, unlike MRT, the test-retest effects were unclear in LJT as the comparison group's performance over time was not statistically significant for both semantically appropriate and inappropriate stimuli (p > .05). Second, the automatized group significantly improved their LJT performance with large effects, especially in the context of semantically inappropriate stimuli (t = 6.388, p < .001, d = 2.82) relative to semantically appropriate stimuli (t = 2.817, p = .005, d = 0.76). The automatized group outperformed the comparison group in both stimulus contexts (t = 2.128, 5.731, p = .046, < .001, d = 0.84, 2.04) and the declarative group in semantically inappropriate stimulus contexts (t = 4.469, p < .001, d = 1.27). Whereas the declarative group's performance over time was significant in both stimulus contexts (t = 4.730, 6.338, p < .001, d = 1.25, 1.68), they outperformed the comparison group only in semantically appropriate stimuli (t = 4.08, p < .001, d = 1.44) but not in semantically inappropriate stimuli (t = 2.225, p = .068, d = 0.78).

Table 3 Descriptive and Mixed Effects Analyses of Participants' Lexicosemantic Judgement Test Scores

A. Descriptive Results				
Group	Time/Dimension	*M*	*SE*	*95% CI*
Automatized	Pre/Appropriate	55.9	1.28	[53.4, 58.5]
	Pre/Inappropriate	44.5	1.28	[42.0, 47.0]
	Post/Appropriate	60.8	1.29	[58.3, 63.3]
	Post/Inappropriate	62.4	1.29	[59.9, 65.0]
Declarative	Pre/Appropriate	56.7	1.26	[54.2, 59.1]
	Pre/Inappropriate	43.6	1.26	[41.1, 46.1]
	Post/Appropriate	64.7	1.27	[62.2, 67.2]
	Post/Inappropriate	54.3	1.27	[51.8, 56.8]
Comparison	Pre/Appropriate	54.9	1.87	[51.3, 58.6]
	Pre/Inappropriate	44.5	1.87	[40.9, 48.2]
	Post/Appropriate	55.3	1.91	[51.6, 59.1]
	Post/Inappropriate	49.2	1.91	[45.5, 53.0]
B. Mixed Effects Modelling				
Fixed Effects	*F*	*p*	$R^2_{conditional}$	$R^2_{mariginal}$
Group	7.176	.001*	.432	.365
Time	88.9518	< .001*		
Dimension	101.218	< .001*		
Group:Time	8.490	< .001*		
Group:Dimension	7.869	< .001*		
Time:Dimension	16.622	< .001*		
Group:Time:Dimension	5.012	.007*		

Note * for p < .05

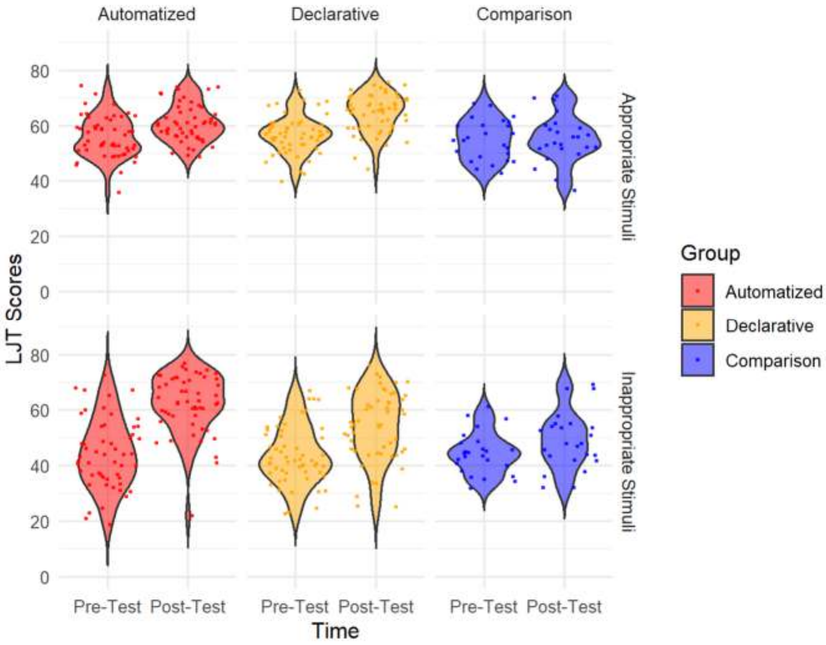


Figure 6. Participants' Lexicosemantic Judgement scores (out of 80 points) under two different stimulus conditions (semantically appropriate vs. inappropriate). The automatized group outperformed the declarative group in rejecting semantically inappropriate stimuli, demonstrating large effect sizes.

General Listening Proficiency Test

Participants' general L2 listening proficiency, measured via TOEIC scores, is summarized in Table 4 and visually plotted in Figure 7. The test comprised three different parts (Parts 1-3), and their separate scores in Parts 1-3 (max = 30 points) were used for the analyses. The descriptive statistics showed that most of the participants' listening scores improved over time regardless of the group conditions. Different from the models for MRT and LJT, the section parts (Parts 1-3; labelled "Dimension") were used as random effects in the mixed effects analyses—DV ~ group*time + (1|ID) + (1|Dimension). The results yielded only significant main effects of Time (p < .001). The analysis of effect size noted large effects for the Automatized (d = 1.31) and Declarative (d = 1.53) groups, and medium effects for the Comparison group (d = 0.93). Given that the interaction effects of group and time failed to reach statistical significance (p = .632), there were two key observations. First, the three groups' performance prior to the project was comparable. Second, the participants appeared to enhance their general listening proficiency test scores regardless of group conditions. Thus, the effects of vocabulary-focused training (MRT with and without LJT) on global L2 listening proficiency remained unclear. In light of the effect sizes, there is an emerging pattern suggesting that while the test-retest effect (indicated by the comparison group) could be medium (d < 1.00), vocabulary training (indicated by the experimental groups) could result in a large effect (d > 1.00).

Table 4 Descriptive and Mixed Effects Analyses of Participants' General L2 Listening Proficiency Test Scores.

A. Descriptive Results				
Group	Time	*M*	*SE*	*95% CI*
Automatized	Pre	16.4 (54.6%)	0.477	[15.5, 17.4]
	Post	19.5 (65.0%)	0.477	[18.6, 20.5]
Declarative	Pre	16.7 (55.6%)	0.469	[15.7, 17.6]
	Post	19.8 (66.0%)	0.469	[18.8, 20.7]
Comparison	Pre	16.8 (56.0%)	1.87	[15.4, 18.2]
	Post	19.3 (64.3%)	1.87	[17.9, 20.7]
B. Mixed Effects Modelling				
Fixed Effects	*F*	*p*	$R^2_{conditional}$	$R^2_{mariginal}$
Group	0.068	.068	.504	.105
Time	138.217	< .001*		
Group:Time	0.452	.632		

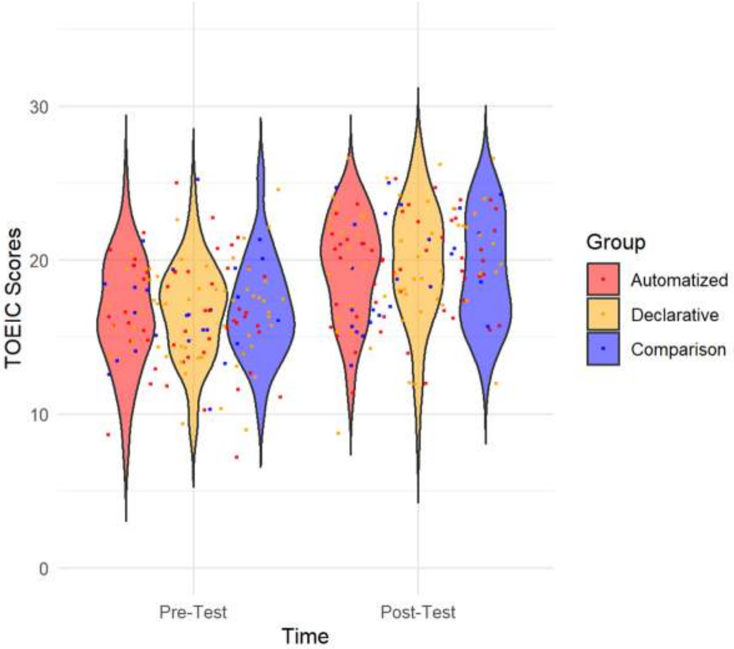


Figure 7. Participants' general listening proficiency scores showing similar gain patterns over time (out of 30 points).

A follow-up analysis was conducted to further examine to what extent participants' general L2 listening proficiency could be linked not only to what types of training they received (Automatized, Declarative, Comparison) but also to how much they *developed* their automatized and declarative phonological vocabulary knowledge throughout the project. To index changes in participants' automatized and declarative phonological vocabulary knowledge, relative gain scores were calculated by dividing their raw gain scores (post-test minus pre-test scores) by pre-test scores. Relative gain scores for LJT were labelled as R_LJT, and relative scores for MRT were labelled as R_MRT. The relative (rather than raw) gain scores are assumed reflect the extent to which gains took place within the period of the project while their initial performance was statistically controlled for. The following model was constructed: DV ~ group*time*R_LJT + group*time*R_MRT + (1|ID) + (1|Dimension). As shown in the results (see Table 5), the main effects of R_LJT (but not those of R_MRT) were found to be statistically significant (F = 4.168, p = .043). However, any interaction effects of R_LJT (nor R_MRT) failed to reach statistical significance (p < .05). The results indicated that participants who developed their automatized phonological vocabulary knowledge were likely to attain more advanced L2 listening proficiency regardless of group conditions.

Table 5 Results of Follow-Up Mixed Effects Analyses of Participants' General L2 Listening Proficiency Test Scores vs. Relative Vocabulary Gains

Fixed Effects	*F*	*p*	$R^2_{conditional}$	$R^2_{mariginal}$
Group	4.142	.018*	.616	.210
Time	22.605	< .001*		
R_LJT	4.168	.043*		
R_MRT	0.147	.701		
Group:Time	0.236	.789		
Group:R_LJT	2.635	.075		
Time:R_LJT	0.059	.808		
Group:R_MRT	1.967	.144		
Time:R_MRT	1.775	.183		
Group:Time:R_LJT	0.693	.500		
Group:Time:R_MRT	0.791	.453		

Note. R_LJT for relative gain scores in LJT; R_MRT for relative gain scores in MRT

Discussion

Although much research has supported the significant roles of vocabulary knowledge in L2 listening proficiency (e.g., Zhang & Zhang, 2020), some scholars have argued that real-life L2 comprehension concerns not only how much learners know about the form and meaning of words but also how accurately and promptly they can access these words in relation to surrounding words at the sentence level (Nation, 2013; Schmitt, 2019). This distinction relates to the declarative and automatized dimensions of L2 knowledge stated in the skill acquisition theory for instructed L2 learning (DeKeyser, 2017; Suzuki, 2023). Given the lack of research in this area, recent studies have proposed an assessment framework of instruments tapping into the two different aspects of L2 phonological vocabulary knowledge—meaning recognition for declarative knowledge and lexicosemantic judgement for automatized knowledge (Saito et al., 2023; Uchihara et al., 2024). The main objective of the current investigation was to examine the roles of declarative and automatized dimensions of phonological vocabulary knowledge in global

L2 listening proficiency from a longitudinal perspective. With133 Japanese EFL students, we conducted a training study wherein participants worked on the development of declarative knowledge of target words via meaning recognition tasks, with and without the automatization of such learned knowledge via lexicosemantic judgement tasks.

Effects of Training in Phonological Vocabulary

As for the first research question, which asked about the comparative effects of different vocabulary training on different vocabulary knowledge, both the automatized and declarative groups demonstrated substantial learning gains in the declarative dimensions of phonological vocabulary knowledge. However, regarding automatized vocabulary knowledge, the declarative group's improvement was somewhat limited. After a series of form-meaning mapping practice activities, they began to show some form of automatization, but only when their abilities were tested to evaluate whether the target words were used in a semantically appropriate manner. In contrast, the automatized group significantly enhanced their ability to not only accept the semantically appropriate use of the target words but also reject the semantically inappropriate use of the target words. In essence, the gains in automatized vocabulary knowledge were significantly greater in the automatized group compared to the declarative group.

The findings here—training declarative and automatized dimensions of phonological vocabulary leading to different learning behaviors—supported our overall assumptions inspired by the skill acquisition theory that declarative and automatized phonological vocabulary are partially overlapping but essentially different constructs (DeKeyser, 2017; Suzuki, 2023). Turning to the L2 morphosyntax literature, automatized knowledge has been measured not only in terms of how learners can accept when target linguistic items are used correctly but also how they can reject when the items are used incorrectly (i.e., grammaticality judgement task). Our results, showing that the effects of automatized training were most clearly observed in the context of semantically inappropriate stimuli, concur with previous findings that it is the correct rejection behaviors that may best reflect the degree of integration and automatization of L2 knowledge (Ellis, 2005).

To date, existing L2 vocabulary research has exclusively focused on the form-meaning mapping of words. L2 learners' phonological vocabulary knowledge has typically been assessed via meaning recognition and recall tests. From a pedagogical perspective, many training techniques have focused on the development of declarative knowledge via recognition and recall activities (e.g., Webb, Yanagisawa, & Uchihara, 2020). However, few empirical studies have explored how to help L2 learners enhance accurate, prompt, and stable access to these words in global contexts so that they can actually use these words in communicatively authentic contexts (Schmitt, 2019).

We echo certain scholars who assert that it is important to start vocabulary training with a range of form-meaning mapping activities to help learners notice, understand, and incorporate what target words sound like and mean. To this end, L2 learners' declarative knowledge can be strengthened through both recognition and recall activities (Cheng et al., 2022). With a view to attaining robust automatized vocabulary knowledge, which ultimately matters for successful L2 listening comprehension in real-life settings, we argue that it is important to devise pedagogical activities that further help L2 learners use these target words in a semantically, collocationally, and grammatically appropriate manner in relation to surrounding words (Nation, 2013). Lexicosemantic judgement tasks could be used to promote the automatization of already-learned phonological vocabulary knowledge (DeKeyser, 2017; Suzuki, 2023).

Effects of Training on Global Listening

As for the second research question, which asked about the ultimate impacts of different vocabulary training on global L2 listening proficiency, we initially predicted that the effects of vocabulary training could positively impact participants' general L2 listening proficiency at the time of the post-test, given that the learning of the 80 target words was assumed to equip the participants with the knowledge of all the lexical items used in the TOEIC test materials. Contrary to our predictions, however, the findings did not clearly support the link between the robust improvement (thanks to automatization training) of the target words and *substantial* changes in participants' global listening test scores (measured via TOEIC). The results of the follow-up analysis rather suggest that enhanced L2 listening proficiency may be predicted by the extent to which participants improved in the automatized, rather than declarative, dimensions of target words, *regardless* of the training conditions.

Broadly speaking, these findings align with cross-sectional evidence supporting the hierarchical roles of declarative and automatized vocabulary knowledge in global L2 listening proficiency, with automatized knowledge having a stronger impact than declarative knowledge (Saito et al., 2023; Uchihara et al., 2024). However, the absence of a direct link between training type and improved listening proficiency does not clearly support the superiority of either automatized or declarative training methods. In the following section, we present suggestions on refining the content of vocabulary training in this study, considering the nature of the participants in this particular study (CEFR A2-B2) and communicatively authentic L2 listening comprehension (high-level variability). Such refinements may enhance the efficacy of vocabulary training and its ultimate impact on global L2 listening proficiency.

Future Directions

First and foremost, given the relatively low English proficiency range of participants in this study (CEFR A2-B2), L2 listening comprehension could be a particularly demanding task. It is important to consider the possibility that the duration of automatization training (3 hours) may not have been sufficient to induce substantial automatization of the target words, allowing participants to readily use them during global L2 listening comprehension (cf. Elgort, 2011). As previous studies have shown, the automatization of newly acquired knowledge often requires a relatively extensive training period (e.g., DeKeyser, 1997, with eight weeks of practice). Therefore, future studies should investigate the effects of a more extended training period on the development of global listening proficiency among low-to-intermediate L2 learners (e.g., Vandergrift & Tafaghodtari, 2010 for over an academic semester).

Next, future studies need to revisit the quality of materials, especially during the phase of the lexicosemantic judgment training, wherein learners were trained and pushed to establish robust phonological vocabulary knowledge that could be accessed despite the demanding nature of real-life listening contexts. In the current study, participants listened to target words produced by only a single talker. Such materials may not reflect the reality of L2 listening, where learners are exposed to and required to understand L2 utterances produced by many different talkers. This limitation may have hindered participants from attaining robust phonological representations of words suitable for such challenging listening tasks. Not only can the spectral representation of these words differ due to speakers' anatomical differences (e.g., longer vocal tracts producing lower formant values; Adank, Smits, & van Hout, 2004), but these words can also be produced at different speeds due to individuals' idiosyncratic speech behaviors (e.g., some making more repetitions thus speaking faster; De Jong, Groenhout, Schoonen, & Hulstijn, 2015). Repeated exposure to multiple talkers plays a key role in inducing both L1 and L2 listeners to develop abstract phonemic and lexical categories and enhance their accessing speed to these representations (Zhang et al., 2021; but see Brekelmans et al., 2022).

Furthermore, one more key characteristic of developing robust L2 listening skills is the presence of noise. Unlike the training treatment in the current study wherein participants were trained on target words in quiet without any sound distraction, real-life listening takes place in contexts filled with different types of noise, such as environmental sounds and multi-talker voices. Prior research has tested how L1 and L2 listeners perceive segmental sounds when speech is masked with white and pink noise (masking energy in speech) or with background talkers (masking information in speech). The results have shown that such noise factors could substantially affect L2 listeners' performance (but not necessarily L1 listeners; Bradlow & Bent, 2002; Lecumberri, Cooke, & Cutler, 2010; Peng & Wang, 2016). Multi-talker variation could be more challenging to L2 listeners when such speech is delivered under noise conditions (Bent, Kewley-Port & Ferguson, 2010).

In terms of training, although limited, there is emerging evidence that providing training under noise conditions can be more effective than delivering the same materials in silence. Such training helps develop robust L2 phonological representations (Mi, Tao, Wang et al., 2021). The acquisitional impact of noise can be significant, particularly when training induces learners to focus on lexical (rather than sub-lexical/phonemic) units (Mora, Ortega, Mora-Plaza et al., 2022) and when noise levels are adjusted to participants' individual auditory processing profiles (Leong, Price et al., 2018). These studies have demonstrated that participants who received noise-based training showed more accurate and fluent access to new L2 sounds and words during post-tests delivered in silence, compared to those who received training without noise.

There are a number of potential benefits from noise-based training. While learners may use a range of cues to perceive target sounds in clear speech, noise is believed to push L2 learners to focus on the most salient, reliable, and primary acoustic correlates of target sounds (Cooke, García Lecumberri, & Barker, 2008). For instance, Japanese listeners can rely on various acoustic information (e.g., second and third formants, duration, and intensity) to differentiate between English [r] and [l]. However, these cues are not equally reliable, as some are more sensitive to surrounding phonetic contexts (e.g., the second formant can be influenced by preceding and following vowels; Saito, 2013). It is likely that exposure to noise-based input could encourage Japanese listeners to identify and rely on the most reliable cue—such as third formant variation, which native speakers use as a primary acoustic correlate of the English [r]-[l] contrast. Moreover, noise-based training offers a range of cognitive advantages. Training under noisy conditions has been shown to improve listeners' ability to direct attention to important phonetic details, suggesting that this form of training may promote better overall auditory attention and selective listening skills (Rogers, Lister, & Febo, 2006)

In essence, noise-based training can help L2 learners develop optimal sound and word processing strategies, and refine or strengthen their representations more effectively than when they are exposed to clean input without noise. It would be intriguing if future studies explored how to enhance the effectiveness of LJT training under simulated real-life listening conditions, wherein target words are produced by multiple talkers under various noise conditions. Such studies may not only improve the automatization of phonological vocabulary knowledge but also boost its ultimate impact on global L2 listening proficiency.

Another issue worthy of discussion concerns individual differences *within* learners. It is important to remember that the follow-up analysis suggested that participants' relative gains in LJT (but not MRT) have some impact on global L2 listening proficiency throughout the project. This provides indirect evidence that a greater degree of automatization of phonological vocabulary knowledge may lead to the attainment of more advanced L2 listening proficiency regardless of the type of vocabulary training L2 learners undergo. This raises the possibility that some individuals may achieve robust automatization even through repeated engagement with brief form-meaning mapping training (3 hours), without the need for specific automatization training, which may be sufficient to positively impact their global L2 listening proficiency. It would be valuable to investigate which types of learners are able to automatize vocabulary knowledge and translate it into enhanced L2 listening proficiency more efficiently. In light of cross-sectional evidence, such learners may possess stronger aptitude profiles, such as enhanced working memory and auditory processing abilities (Wallace, 2023). Given the exploratory nature of this analysis, we call for future studies to further explore the perceptual-cognitive individual differences that may determine the optimal combinations of training to maximize vocabulary-listening links.

Finally, the findings presented in this paper are specific to the population of this study—college-level Japanese learners of English. The target words for training and assessment were selected based on factors unique to these learners (e.g., the TOEIC, which is widely used among this population, was chosen to develop the word corpus; particular phonological contrasts such as English [r] and [l] and word stress were emphasized in target word selection). Future studies should aim to replicate these findings with learners from different L1-L2 pairings (e.g., linguistically close vs. distant languages), under varying learning conditions (e.g., naturalistic vs. classroom), and at different phases of acquisition (e.g., early vs. ultimate attainment).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language learning*, *59*(2), 249-306.

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the acoustical society of America*, *116*(5), 3099-3107.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. Behavior research methods, 52, 388-407.

Bent, T., Kewley-Port, D., & Ferguson, S. H. (2010). Across-talker effects on non-native listeners' vowel perception in noise. *The Journal of the Acoustical Society of America*, *128*(5), 3142-3151.

Birch, S., & Rayner, K. (2010). Effects of syntactic prominence on eye movements during reading. *Memory & Cognition*, *38*, 740-752.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, *112*(1), 272-284.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074-2085.

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352.

Cheng, J., Matthews, J., Lange, K., & McLean, S. (2022). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3137

Cobb, T. (2007). Computing the vocabulary demands of L2 reading.

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*, 414-427.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141-201.

Dalman, M., & Plonsky, L. (2022). The effectiveness of second-language listening strategy instruction: A meta-analysis. *Language Teaching Research*, 13621688211072981.

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*, 617-641. https://doi.org/10.1177/1362168820911

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223-243.

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in second language acquisition*, *19*(2), 195-221.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In *The Routledge handbook of instructed second language acquisition* (pp. 15-32). Routledge.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367-413.

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365-414.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*, *27*(2), 141-172.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics*, *27*(1), 1-24. https://doi.org/10.1093/applin/ami038

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second Language Research*, *34*(1), 67-101.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, *36*(1), 101-132. https://doi.org/10.1017/S0272263113000624

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*(2), 269-297.

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.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in second language acquisition*, *35*(3), 423-449.

Kachlicka, M., Symons, A., Ruan, Y., Saito, K., Dick, F., & Tierney, A. (forthcoming). Effects of targeted perceptual training on L2 suprasegmental weighting strategies.

Kamiya, N. (2022). The limited effects of visual and audio modalities on second language listening comprehension. *Language Teaching Research*, 13621688221096213.

Kobayashi, A. (2018). Investigating the effects of metacognitive instruction in listening for EFL learners. *Journal of Asia TEFL*, *15*(2), 310.

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech communication*, *52*(11-12), 864-886.

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PloS one*, *13*(10), e0204888.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., ... & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language learning*, *63*(3), 530-566.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760. https://doi.org/10.1177/1362168814567889

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. *Journal of Memory and Language*, *32*, 56-75.

Mi, L., Tao, S., Wang, W., Dong, Q., Dong, B., Li, M., & Liu, C. (2021). Training non-native vowel perception: In quiet or noise. *The Journal of the Acoustical Society of America*, *149*(6), 4607-4619.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

Milliner, B., & Dimoski, B. (2024). The effects of a metacognitive intervention on lower-proficiency EFL learners' listening comprehension and listening self-efficacy. *Language Teaching Research*, *28*(2), 679-713.

Mora, J. C., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (2022). Training the pronunciation of L2 vowels under different conditions: the use of non-lexical materials and masking noise. *Phonetica*, *79*(1), 1-43.

Mora-Plaza, I., Saito, K., Suzukida, Y., Dewaele, J-M., & Tierney, A. (2022). Tools for second language speech research and teaching. http://sla-speech-tools.com. http://doi.org/10.17616/R31NJNAX

Moravcsik, J. E., & Healy, A. F. (1998). Effect of syntactic role and syntactic prominence on letter detection. *Psychonomic Bulletin & Review*, *5*(1), 96-100.

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. *Journal of Memory and Language*, *32*(1), 56-75.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520-531.

Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2021). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*, *55*, 248-270.

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* https://www.wgtn.ac.nz/lals/about/staff/paul-nation

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

Peng, Z. E., & Wang, L. M. (2016). Effects of noise, reverberation and foreign accent on native and non-native listeners' performance of English speech comprehension. *The Journal of the Acoustical Society of America*, *139*(5), 2772-2783.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language learning*, *64*(4), 878-912.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, *27*, 465-485.

Saito, K. (2013). Age effects on late bilingualism: The production development of/ɹ/by high-proficiency Japanese learners of English. *Journal of Memory and Language*, *69*, 546-562.

Saito, K., & Akiyama, Y. (2018). Effects of video-based interaction on the development of second language listening comprehension ability: A longitudinal study. *Tesol Quarterly*, *52*(1), 163-176.

Saito, K., Hosaka, I., Macmillan, K., Takizawa, K., Suzukida, Y., & Uchihara, T. (forthcoming). Timed vs. untimed lexicosemantic judgment task for measuring automatized phonological vocabulary knowledge.

Saito, K., & Uchihara, T. (2024). Experiential, perceptual, and cognitive individual differences in the development of automatized and declarative phonological vocabulary knowledge. *Bilingualism: Language and Cognition*. https://doi.org/10.1017/S1366728924000609

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2023). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S027226312300044X

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

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274. https://doi.org/10.1017/S0261444819000053

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language learning*, *60*(2), 263-308.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in second language acquisition*, *31*(4), 577-607.

Suzuki, Y. (Ed.) (2023). *Practice and automatization in second language research: Perspectives from Skill Acquisition Theory and Cognitive Psychology*. New York: Routledge.

Suzuki, Y., & Elgort, I. (2023). Measuring automaticity in second language comprehension. In Y. Suzuki (Ed.), *Practice and Automatization in Second Language Research* (pp. 206-234). Routledge.

Taguchi, N. (2011). Teaching pragmatics: Trends and issues. *Annual review of applied linguistics*, *31*, 289-310.

Uchihara, T., Saito, K., Kurokawa, S., Takizawa, K., & Suzukida, Y. (2024). Declarative and automatized phonological vocabulary knowledge: Recognition, recall, lexicosemantic judgement, and listening-focused employability of L2 words. *Language Learning*. https://doi.org/10.1111/lang.12668

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*, *44*(2), 357-380.

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383-410.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191-210.

Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language Learning*, *60*, 470-497.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390-416.

Vandergrift, L., & Goh, C. (2012). Teaching and learning second language listening: Metacognition in action. *New York*.

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.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715-738.

Yeldham, M. (2022). Examining the Interaction between two Process-based L2 Listening Instruction Methods and Listener Proficiency Level: Which form of Instruction Most Benefits Which Learners? *TESOL Quarterly, 56*(2), 688-712.

Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802-4825.

Zhang, P., & Graham, S. (2020). Learning vocabulary through listening: The role of vocabulary knowledge and listening proficiency. *Language Learning*, *70*(4), 1017-1053.

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696-725.