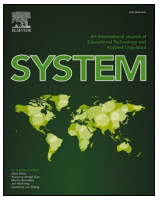
Multimedia enhanced vocabulary learning: the role of input condition and learner-related factors

System 122 (2024) 103275

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system



Multimedia enhanced vocabulary learning: The role of input condition and learner-related factors

Pengchong Zhang [a,*], Shi Zhang [b]

[a] Institute of Education, University of Reading, London Road Campus, 4 Redlands Road, Reading, RG1 5EX, UK
[b] College of Foreign Languages and Cultures, Chengdu University of Technology, 1 East Third Road, Erxianqiao, Chenghua District, Chengdu, 610059, PR China

ARTICLE INFO

Keywords:
Vocabulary
Multimodal input
Explicit instruction
Working memory
Individual difference

ABSTRACT

This study explored the effects of different types of input (verbal-only vs. verbal plus content-related nonverbal vs. verbal plus paralinguistic-related nonverbal) on vocabulary learning from multimedia. It also investigated how learning was moderated by three learner-related factors (prior vocabulary size, phonological short-term memory (PSTM) capacity, and comprehension of the input). Forty-three English learners of French first completed a French vocabulary size test, a vocabulary pre-test, and a PSTM test online. They then viewed three sets of multimodal teaching materials, each with a different type of input condition, followed by a vocabulary post-test and a comprehension test. Findings indicated that multimodal input including additional nonverbal information was more beneficial than traditional verbal-only input for productive vocabulary knowledge gains. Additionally, comprehension of the input was the most important moderator for the learning gains, especially when the input included paralinguistic-related nonverbal information. The findings provide novel insights into theories of multimedia learning and have pedagogical implications for the design of multimodal language learning materials.

1. Introduction

Learning foreign language vocabulary is challenging and is arguably much more so for young learners in Anglophone settings, where exposure to a foreign language outside of school is very limited due to the predominant nature of English as a global language. One reason for UK foreign language learners' lower levels of vocabulary may be that schools have limited curriculum time for foreign language instruction (Collen, 2022), a factor known to be important for foreign language progression (Graham et al., 2017). UK foreign language learners also make less use of multimedia or multimodal input, such as video clips and computer games, outside of school than is the case in Europe, where there is strong evidence of their beneficial impact on foreign language vocabulary development (De Wilde et al., 2021). It is worth noting, however, that existing empirical evidence examining the role of multimodal input in foreign language learning has mainly been drawn from studies investigating incidental vocabulary learning (e.g., Montero-Perez et al., 2014; Peters, 2019) whereby learners unconsciously "pick up" unknown vocabulary from the meaning-focused input (Hulstijn, 2001). To our knowledge, no study so far has explored the role of multimodal input in explicit vocabulary instruction, where vocabulary knowledge is intentionally delivered through pedagogical activities (Hulstijn, 2001). In addition, in those existing studies (Montero-Perez et al., 2014; Peters, 2019), multimodal input has largely taken a single format, i.e., captioned/subtitled video clips. Very limited research has

* Corresponding author. University of Reading, London Road Campus, 4 Redlands Road, Reading, RG1 5EX, UK.
E-mail addresses: anthony.zhang@reading.ac.uk (P. Zhang), shi.zhang@cdut.edu.cn (S. Zhang).

https://doi.org/10.1016/j.system.2024.103275
Received 13 September 2023; Received in revised form 21 February 2024; Accepted 11 March 2024

examined the effects of different combinations of multimodal input (incorporating different types of verbal and nonverbal elements) on learning and what additional scaffolding might be needed for lower proficiency learners to benefit from them.

How learner-related factors (e.g., working memory capacity, existing language proficiency) moderate learning from multimodal input is also an important area to investigate because perceptual learning styles (e.g., visual, auditory, kinesthetic, etc.) (James & Galbraith, 1985) and the level of sensitivity to nonverbal cues within the multimodal input (Gardner, 1999) are believed to vary significantly between individuals. In the case of vocabulary learning, studies have found that both learners' pre-existing levels of vocabulary knowledge and complex working memory capacity (Montero-Perez, 2020) or phonological short-term memory capacity (Teng, 2023a) significantly correlated with learning gains from multimodal input. The evidence, however, has been mainly gathered in relation to incidental learning. In addition, other important factors, such as how well learners are able to comprehend the multimodal input, have not been taken into account. The present study, therefore, investigates first, how different combinations of multimodal input affect vocabulary learning from explicit instruction; and second, how vocabulary gains are moderated by three learner-related factors, that is, prior vocabulary size, working memory capacity, and comprehension of the multimodal input.

2. Literature review

2.1. Foreign language vocabulary learning through multimodal input

With the wider application of educational technology, there has been an increasing focus on the benefits of using multimedia to assist foreign language teaching and learning. In addition to offering verbal information (spoken or written words), multimedia is multimodal in nature, also including nonverbal information which can take the form of static pictures, dynamic animations, or videos (Mayer, 2009). According to the cognitive theory of multimedia learning (Mayer, 2017), multimodal input is expected to deepen learners' understanding of the content on two levels. First, dual coding theory (Paivio, 1986) holds that in semantic memory, concepts can be processed by two different systems: a verbal-based system which processes linguistic information, and an imagery-based system which processes nonverbal information. Since individuals have limited capacity in each memory system at any given time (Baddeley, 1999), multimodal input involving both verbal and nonverbal information allows more content to be processed simultaneously in semantic memory and therefore enhances learning. Second, exposure to multimodal input requires learners to select, organise, and integrate information, which then tends to help them learn better (Mayer, 2009).

Studies investigating the effects of multimodal input on learning have been mainly in the first language context. With foreign language learners, limited studies to date have focused on exploring how much linguistic knowledge, in particular vocabulary knowledge, learners can acquire incidentally from multimodal input (Zhang & Zou, 2022). For example, Montero-Perez et al. (2014) examined among Flemish adult learners of English the effects of captioned videos on the acquisition of English words. They found that learners who watched the captioned videos outperformed those who watched videos without captions on both form and meaning recognition of the target words. Similar findings were confirmed in later studies by Peters (2019) and Teng (2022). The former found that secondary school Dutch learners of English made significantly larger gains in form recognition and meaning recall of the target words after watching audio-visual input with captions than upon viewing input without captions. In the latter, captioning was found to have benefited vocabulary gains in four knowledge dimensions (form recognition and recall; meaning recognition and recall) among adult Chinese English as a foreign language (EFL) learners.

It should be noted that in the above three studies (Montero-Perez et al., 2014; Peters, 2019; Teng, 2022), the evidence for the positive effects of multimodal input on vocabulary learning can be mainly attributed to the verbal element (i.e., captions) of the input. The fact that Peters (2019) also found that target words supplemented by imagery from the input had a significantly higher likelihood to be learnt than those without imagery raises an important question regarding the role of the nonverbal element within multimodal input in facilitating vocabulary learning. Two kinds of nonverbal input are potentially useful for vocabulary learning: content-related (e.g., a picture or video which is related to the meaning of a target word) (Pellicer-Sánchez et al., 2020) or paralinguistic-related (e.g., gestures and facial expressions from a teacher explaining a target word) (Batty, 2020; Sueyoshi & Hardison, 2005). Empirical studies so far have mainly examined the effects of content-related nonverbal input and have reached mixed conclusions.

On the one hand, Warren et al. (2018) found that for high-intermediate English as a second language (ESL) learners, meaning recognition of target words was significantly better when words were presented with content-related pictures than when they were given text-only definitions. A follow-up study by Teng (2023b) provided further evidence for the beneficial impact on vocabulary learning measured both receptively (form and meaning recognition) and productively (form recall) of content-related videos. In his study, Chinese EFL learners encountered target items in two conditions: 1) items appeared in a content-related video (a one-sentence film clip with images representing the meaning of the word) alongside other word meaning related verbal information (definition plus background information); 2) verbal only information about the items was given. Learning was significantly better in the first condition. On the other hand, no effects of content-related nonverbal input were detected in a study by Boers et al. (2017). They had three trials including two groups of EFL and one group of ESL learners. Each trial involved learners reading authentic English texts under two conditions: one with verbal glosses and the other with verbal glosses plus content-related pictures. Findings showed that for all three trials, there were no effects of condition for meaning recognition of the target words. Surprisingly, for one EFL and the ESL group, form recall of the target words was better when the condition was verbal-only glosses than when there were glosses plus content-related pictures, suggesting that the additional content-related pictures impeded rather than facilitated learning.

Compared to content-related nonverbal input, the effects of paralinguistic-related nonverbal input have been neglected in existing vocabulary research. To our knowledge, Arndt and Woore (2018) is the only study examining how paralinguistic-related nonverbal input (this term was not used explicitly by the authors) affects learning. Although they found that the amounts of foreign language

vocabulary acquired from viewing vlogs (involving the vloggers talking directly to the camera, hence providing paralinguistic-related nonverbal input) and reading written blogs did not differ significantly, learners who viewed L2 vlogs did score more highly on a test of word meaning recognition, with a small effect size. These findings suggest that participants' overall vocabulary learning in the vlogs group might have benefited from the additional paralinguistic-related nonverbal information contained in the videos. It should be noted, however, that the evidence is somewhat indirect as the two conditions were not highly controlled. In addition to the presence or absence of paralinguistic nonverbal information, the two conditions also differed in whether or not participants had access to the written form of the target vocabulary.

In sum, potential benefits of multimodal input for vocabulary learning have been confirmed empirically. Such benefits, however, are mainly attributed to the verbal element within the input. Few attempts have been made to investigate whether nonverbal input, paralinguistic-related nonverbal input in particular, may facilitate learning. In addition, no study to date has specifically compared the effects of different types of multimodal input (i.e., content-related nonverbal vs. paralinguistic-related nonverbal). Moreover, existing studies mainly examined incidental vocabulary learning. Little research attention has been given to how multimodal input can facilitate explicit vocabulary instruction. Considering the unique advantages that explicit vocabulary instruction has over incidental learning in prompting "noticing" (Schmidt, 1990) and hence facilitating learning (Laufer, 2006), the role of multimodal input in explicit vocabulary instruction is worthy of further exploration.

2.2. Learner-related factors in learning from multimodal input

When examining language learning from multimodal input, learner-related factors are believed to play a crucial role, and not just because of any supposed differences in individuals' sensitivity to nonverbal cues as per Multiple Intelligences theory (Gardner, 1999). In addition, learner-related factors are important because of the contradictory findings between the first language (L1) and foreign language context in terms of the redundancy principle (Mayer, 2017). According to this principle, when graphic information (such as a diagram) is accompanied by both written text and its correspondent narration, the written text is seen as redundant because it repeats the same information in the narration and takes up additional working memory. It therefore impedes rather than facilitates learning (Kalyuga & Sweller, 2014). This principle is empirically supported in the L1 context (Austin, 2009). For foreign language learners, providing information in two different forms concurrently has been found to support learning rather than impeding it. Hence the empirical evidence for the redundancy principle is contradictory across L1 and L2 contexts (Peters, 2019; Warren et al., 2018).

That contradictory empirical evidence for the redundancy principle is arguably not surprising. One way of interpreting the differing effect of redundancy in L1 and L2 might be as follows. For L1 learners, an image accompanying a word is more likely to be redundant, because of the greater probability that they know the word's meaning already, given it is in their L1. Hence the image has no relevance, can be distracting, and may create "a split attention situation" and therefore cognitive overload (Mayer, 2017, p. 413). For foreign language learners, by contrast, more words are likely to be unknown, and they would need to expend cognitive resources in working out their meaning (Mayer et al., 2014). In such instances, an image accompanying an unknown word, rather than being redundant or irrelevant, supports understanding of the word's meaning. Any split attention that arises is offset by a reduction in the overall cognitive burden of comprehending the word (Mayer et al., 2014). Hence the accompanying image helps rather than hinders. Yet while the foregoing explanation for contradictory findings across L1 and L2 contexts is potentially convincing, it lacks empirical support (Mayer et al., 2014) and does not take into account learner-related factors. Low proficiency learners may lack overall processing capacity when faced with multimodal input because of the cognitive burden posed by tasks such as speech segmentation and word recognition; it may therefore be difficult for them to make good use of the accompany images and cope with any negative impact from split attention. By contrast, higher proficiency learners, for whom such tasks are less effortful, may have greater processing space, which would allow them to better utilize the additional, supporting nonverbal input and hence benefit more from it.

Among the limited studies exploring how learner-related factors affect vocabulary learning through multimodal input, two factors have been considered: pre-existing vocabulary knowledge and working memory capacity (Montero-Perez, 2020). Both factors seem to have a positive effect on learning, meaning those with a larger vocabulary size or working memory capacity make larger vocabulary gains from multimodal input. For example, Montero-Perez et al. (2014) found that a significant positive correlation between adult Flemish learners' English vocabulary gains after watching captioned videos and their pre-existing vocabulary size, that is, learners with larger pre-existing vocabulary size made greater vocabulary gains. The effect sizes for these correlations were larger for meaning recognition and recall than for form recognition and clip association (the ability to associate a target word with the video clip it appeared). Similarly, Peters and Webb (2018) found that adult Flemish learners of French made significant meaning recognition and recall gains for the target words after viewing foreign language videos and such gains were larger for learners with higher levels of prior vocabulary knowledge.

Montero-Perez's (2020) study took one step further and explored how vocabulary gains after watching non-captioned foreign language videos were predicted by adult learners' pre-existing vocabulary knowledge and two forms of working memory capacity: phonological short-term memory or PSTM and complex working memory. The former represents the system for temporarily storing phonological information, and the latter refers to the complex system for storing and processing information (Baddeley, 2003). While the capacities of both PSTM and complex working memory are believed to be predictors of L2 vocabulary learning (for a review, see Service & Simard, 2022), previous studies mostly studied this using artificial languages in a lab setting (e.g., Martin & Ellis, 2012), while Montero-Perez's research was conducted in a more naturalistic learning setting. Montero-Perez's results suggested that learners' pre-existing vocabulary knowledge positively correlated with form and meaning recognition of the target words. In addition, only complex working memory significantly predicted form and meaning recognition gains. PSTM did not show any effect on learning gains. Quite different from Montero-Perez (2020), however, a later study (Teng, 2023a) examined among Hong Kong young learners

(mean age = 12.17) the role of the two forms of working memory in vocabulary learning through captioned videos and found that learning gains were moderated by PSTM but not by complex working memory capacity. Teng (2023a) discussed these different findings by attributing them to two factors. First, the overall proficiency level of the learners in Montero-Perez (2020) was much higher than the young learners in his study. With increased proficiency in the target language, learners tend to rely less on their PSTM to analyze the phonological structure of the language (Masoura & Gathercole, 2005). Second, additional verbal input in the form of captions was presented to the young learners in his study, whereas in Montero-Perez's (2020), non-captioned videos were used, which might have resulted in activating a different form of working memory.

The above findings point to the conclusion that more proficient foreign language learners seem to benefit more from multimodal input for vocabulary learning. One possible explanation is that receiving such input allows more proficient learners to pay closer attention to the nonverbal cues, thus reaching a higher comprehension level of the input (Pellicer-Sánchez et al., 2020). In the case of incidental vocabulary learning, how well learners are able to comprehend the input (either written or aural) has been shown empirically to have a clear impact on the learning outcome (Pellicer-Sánchez & Schmitt, 2010; Vidal, 2011). None of the studies so far, however, have examined how vocabulary learning from multimodal input is moderated by learners' comprehension of the input. Additionally, most of the studies involved adult participants. Considering that most existing studies have involved adult learners and the different findings regarding working memory in Montero-Perez (2020) and Teng (2023a), for high-intermediate adult learners in the former and low proficiency young learners in the latter, more studies are needed exploring the factors that influence how much young learners benefit linguistically from multimodal input.

3. Research questions

In light of the research gaps discussed above, the present study aims to answer two research questions.

1. How do different kinds of multimodal input (verbal only vs. verbal plus content-related nonverbal vs. verbal plus paralinguistic-related nonverbal) affect young foreign language learners' vocabulary learning from explicit instruction?
2. To what extent are the effects of vocabulary learning from different kinds of multimodal input moderated by learners' prior vocabulary size, phonological short-term memory capacity, and comprehension of the input?

4. Material and methods

4.1. Sampling

The initial sample included 55 Year 7 learners of French (aged 11–12) recruited through secondary schools in the southeast of England. The study was advertised to learners' parents through schools' newsletters. Interested parents then gave consent for their children's participation. Children's assent was obtained individually before they started the experiment. Learners participated online in two experiment sessions (see *Experiment design and procedures* section). Twelve learners did not attend the second session. Their data were, therefore, removed from the analysis, leaving a valid sample of 43 learners. Learners completed a language background questionnaire, adapted from Sabourin et al. (2016), gathering information about their first language and any additional languages they spoke outside of school. In cases where an additional language was indicated, learners were asked to rate their proficiency in that language from 0 to 10, zero meaning no proficiency, 10 meaning high proficiency. All learners spoke English as a first language and none of them spoke/used French outside of school. They had not previously studied any other foreign languages. They had just started learning French as a modern foreign language in secondary school and were of low proficiency overall. Upon completion of both experiment sessions, learners received a £20 e-voucher to reimburse their costs.

4.2. Experiment design and procedures

The data collection was conducted through an online experiment builder called *Labvanced* (https://www.labvanced.com/, Finger et al., 2017). The experiment adopted both between- and within-participant designs to maximise the statistical power. It involved two sessions, one pre-test session (30 min) and one language learning plus post-test session (60–70 min), with two weeks gap in between. This gap was determined to minimize the effects of the pre-test session on learning at the second session. In the pre-test session, learners completed a language background questionnaire, a French vocabulary size test, a vocabulary pre-test and a working memory test (see *Research instruments* section for detail). The language learning plus post-test session involved learners first watching three sets of multimodal vocabulary teaching materials. They then completed a vocabulary post-test, and a comprehension test (see *Research instruments* section). For all tests, the order of the test items was randomized between learners to prevent any order effect.

Each set of multimodal vocabulary teaching material included a short (2–3 min) French film clip with English-French bilingual subtitles and six PowerPoint slides, each explicitly teaching a target French word appearing in the film clip. The explicit vocabulary teaching was all delivered by one experienced French teacher in three different conditions: verbal-only vs. verbal plus content-related nonverbal vs. verbal plus paralinguistic-related nonverbal. Under the verbal-only condition, learners saw on the slides the original sentence from the film clip including the target word and the English translation for that sentence. They also saw the target word, and it's part of speech and meaning in English. Finally, an additional example French sentence for the target word was given alongside with its English translation (see Fig. 1). While watching the slides, learners were able to hear the French teacher reading out the above information on the slides. Under the verbal plus content-related nonverbal condition, in addition to the input that was given to the

verbal-only condition, at the right bottom corner of each slide, a picture representing the meaning of the target word was provided (see Fig. 2). Regarding the verbal plus paralinguistic-related nonverbal condition, at the right bottom corner of each slide, learners were given a video showing the French teacher reading out the slides (see Fig. 3). In these short video clips, learners were able to clearly see the gesture and facial expressions of the French teacher, hence having access to additional paralinguistic-related nonverbal information. Each learner experienced all three conditions for different items. The order of the three conditions and the three sets of multimodal input was counterbalanced between learners.

4.3. Research instruments

4.3.1. X-lex

A shorter version of X-Lex, adapted from the original X-Lex test (Meara, 1992), was used to measure learners' French vocabulary size. This adapted version is a Yes/No form recognition test assessing 120 words including 100 real French words (20 randomly selected from each of the first five 1000-word frequency bands) and 20 pseudowords. Learners need to choose Yes for the words that they know or can use and No for the words that they do not know or are invented. Fifty points were awarded for each real word ticked "Yes". For each pseudoword ticked Yes, however, a penalty of 250 points was given. The highest possible score was 5000. Overall, this test showed very good reliability, measured by Cronbach's alpha ($\alpha = 0.96$, 95% *CI* [0.94, 0.97]).

4.3.2. Phonological short-term memory capacity

Learners' phonological short-term memory (PSTM) capacity was measured using a backward digit-span task. Traditionally, a forward digit-span task is used to measure PSTM capacity, and a backward digit-span task is believed to assess complex working memory capacity (Baddeley, 2003). More recent studies (e.g., St Clair-Thompson & Allen, 2013; St Clair-Thompson et al., 2010), however, pointed out that, at least for young adults, both backward and forward digit-span tasks measure PSTM capacity. We used the backward digit span task rather than the forward digit span task because the latter is less demanding, and the learners might perform at ceiling and not show any difference in the forward digit span task. Furthermore, we did not use a nonword repetition task as some of the learners had access to family languages that were not English. Therefore, their representation of phonemes and phonological knowledge might be different, and this difference could affect their performance in the nonword repetition task. In our task, learners heard a sequence of digits (e.g., 4, 1, 7, 3) with a 1s interval between each digit, and were asked to recall the digits in reversed order (e.g., 3, 7, 1, 4) by logging them into a textbox. Before the formal test, the learners completed three 3-digit trials to familiarize themselves with this task. They were then tested on digit sequences of four to seven digits, with each sequence length consisting of three trials. The learners received one score for correctly responding to a sequence of digits (max = 12). The test showed good reliability as measured by Cronbach's alpha ($\alpha = 0.85$, 95% *CI* [0.74, 0.90]).

4.3.3. French film clips and target words

The three film clips were extracted from two films: *Harry Potter and the Philosopher's Stone*; and *Ratatouille*. They were selected from those available at *Online Language Learning for All* (https://pdcinmfl.com/online-language-learning-for-all-olla/), a site which houses materials that have been judged to be interesting by young learners and to be age and proficiency-level appropriate (Woore et al., 2020). All clips were between two and 3 min long and had both English and French subtitles. In total, eighteen target words were selected from the clips (six in each clip) by an experienced French teacher who was familiar with the level of the learners and the curriculum. All words were from the first two 1000-word frequency bands, judged by the MultilingProfiler (Finlayson et al., 2022).

4.3.4. Vocabulary pre-test and post-test

Knowledge of the target words was assessed through a vocabulary pre-test and a post-test, adapted from Montero-Perez et al. (2014). The pre-test measured three vocabulary knowledge aspects: form recognition (Yes/No test), meaning recall, and meaning recognition (four-option multiple choice questions: one correct answer and three distractors), in that order. Clip association (three-option multiple choice questions), i.e., the ability to associate a target word with the relevant film clip in which it appeared, was added as a fourth knowledge aspect in the vocabulary post-test. See Fig. 4 for an example test item for each knowledge aspect. The

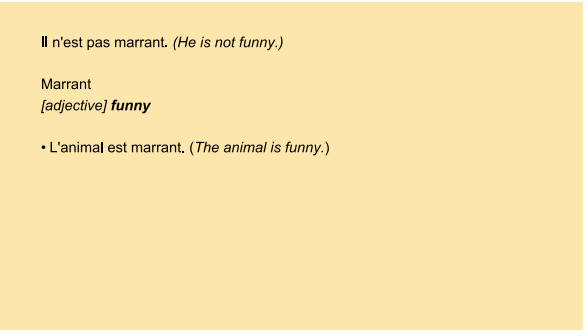


Fig. 1. Verbal-only condition.

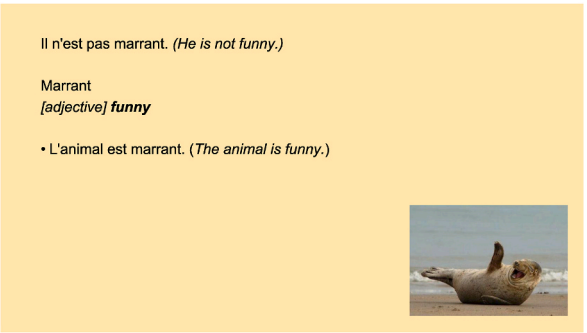


Fig. 2. Verbal plus content-related nonverbal.

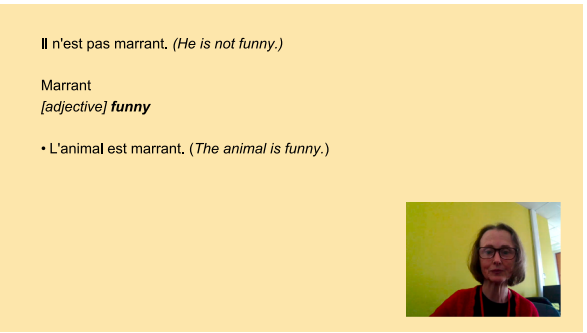


Fig. 3. Verbal plus paralinguistic-related nonverbal.

Form recognition
Have you seen "Bouger" before? (For the post-test, this was phrase as *Has "Bouger" been used in the clips?*)
☐Yes
☐No

Clip association
In which clip did "Bouger" appear?
☐Harry Potter and the snake
☐Ratatouille
☐Harry Potter at the station

Meaning recall
Translate the word into English
Bouger = _____

Meaning recognition
Choose the correct translation:
Bouger
☐To run
☐To move
☐To leave
☐To fly

Fig. 4. Example item for vocabulary pre-test and post-test.

reliability (Cronbach's alpha), given in Table 1, was judged to be good for all measurements, apart from meaning recognition which was acceptable to moderate. This was potentially due to an element of guessing involved as meaning recognition was assessed through multiple choice questions.

4.3.5. Comprehension test

Learners' comprehension of the film clips was measured through a 15-item multiple-choice comprehension test. Five questions were designed for each film clip. For each question, learners had to identify one correct answer from the three given options (see Fig. 5 for an example test item). The reliability for the test was moderate, $\alpha = 0.77$, 95% *CI* [0.57, 0.87].

5. Data analysis

The data were analyzed using Bayesian mixed effects models with the "brms" package (Bürkner, 2017, 2018, 2021) in R (R Core Team, 2022) both by-Participant and by-Item. Weakly informative priors (i.e., prior knowledge of the expected effects of the study) were adopted in order to reduce the effects of extreme outliers on the overall analysis. These weakly informative priors primarily followed the recommendations of Gelman et al. (2008) but also incorporated the follow-up suggestions by Gelman (2023), whereby all nonbinary fixed factors were first centered at the mean and scaled to have a *SD* of 0.50. A student's t distribution with four degrees of freedom was then set for all model terms. For the intercept term, the distribution was scaled to have a mean of 0 and *SD* of 10 whereas the distribution for all other predictor terms was centered at the mean and had a *SD* of 2.50.

The models for form recognition, meaning recognition, and meaning recall included five fixed factors: Time (Pretest vs. Posttest), Condition (Control vs. Content vs. Paralinguistic), PSTM (Phonological short-term memory), X-Lex (French vocabulary size), Comprehension (overall comprehension of the three film clips). Pretest was set as the baseline level for Time and Control was assigned as the baseline level for Condition. All the other continuous fixed factors were scaled to have a mean of 0 and *SD* of 0.50 as discussed above. Two-way Time × Condition interactions were included in all models to examine the first research question. In order to address the second research question, the moderation effects of learner-related factors (i.e., the three continuous fixed factors) on learning, three three-way interactions were added to the fixed effects structure including PSTM × Time × Condition, X-Lex × Time × Condition, and Comprehension × Time × Condition. Slightly different from the other models, the model for clip association only included four fixed factors, without Time, as it was not appropriate to measure learners' ability to associate a certain French target word with a film clip at the pre-test session before they had watched the clips. Three two-way interactions were included in the fixed effects structure to investigate the moderation effects of the three learner-related factors, namely PSTM × Condition, X-Lex × Condition, and Comprehension × Condition.

The random effects structure for all models included by-Participant and by-Item random intercepts. Additional by-Participant and by-Item random slopes for Time were added to the models for form recognition, meaning recognition, and meaning recall to further control the random effects of participants and items being tested twice. Initially, all models were fitted with a maximal model structure, that is, including all theoretically driven fixed factors and interactions. As the current study has a relatively limited number of observations (774 for the model for clip association and 1548 for all other models), backward model selection was performed through cross-validation by gradually removing interactions and fixed factors which did not show any effect. An effect was judged by whether the 95% credible intervals crossed 1.00, that is whether an odds ratio of 1.00, meaning no effect, fell into the range of the 95% credible intervals. For each model comparison, an "elpd_diff" (the expected log pointwise predictive density difference) value was obtained through the "loo" package (Vehtari et al., 2022). In cases where an "elpd_diff" value was smaller than 4, the simplified model was retained before continuing with further model selection (Sivula et al., 2020). When a meaningful effect was confirmed, the effect plot was created using the "plot_model" function within the "sjPlot" package (Lüdecke, 2022).

6. Findings

Descriptive statistics (see Table 2) were first calculated for the three baseline measurements and for each vocabulary measurement under each input condition at pre-test and post-test respectively.

6.1. Form recognition

The final model for form recognition included four fixed factors (Time, Condition, Comprehension, and X-Lex), two-way Time × X-Lex, Time × Condition, Time × Comprehension, and Condition × Comprehension interactions, and three-way Time × Condition × Comprehension interactions. PSTM was removed as a fixed factor during the model selection. Model fit, assessed by $R^2_{marginal} = 0.31$ and $R^2_{conditional} = 0.40$, indicated that fixed effects explained 31% of the variance of the predicted values and both fixed and random effects explained 40% of the variance of the predicted values (Gelman et al., 2019). The model results are given in Table 3 and the effects plots

Table 1
Reliability for vocabulary pre-test and post-test.

	Clip association	Form recognition	Meaning recall	Meaning recognition
Pre-test	–	$\alpha = 0.88$, 95% *CI* [0.78, 0.92]	$\alpha = 0.89$, 95% *CI* [0.80, 0.92]	$\alpha = 0.68$, 95% *CI* [0.56, 0.79]
Post-test	$\alpha = 0.82$, 95% *CI* [0.73, 0.87])	$\alpha = 0.87$, 95% *CI* [0.77, 0.91]	$\alpha = 0.90$, 95% *CI* [0.84, 0.93])	$\alpha = 0.76$, 95% *CI* [0.60, 0.84]

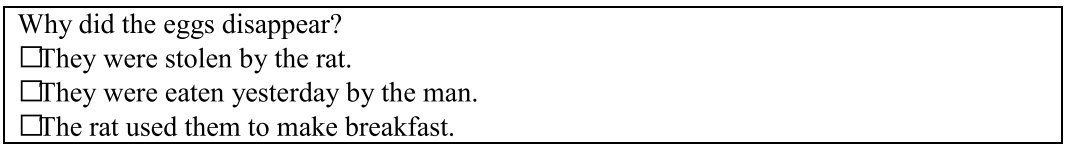


Fig. 5. Example item for comprehension test.

Table 2
Descriptive statistics for all measurements.

Measurement	Time	Input condition	*M*	*SD*	*Min*	*Max*
X-Lex	–	–	497.04	528.78	0.00	2450.00
PSTM	–	–	55.03%	28.14%	0.00%	100.00%
Comprehension	–	–	83.87%	16.16%	20.00%	100.00%
Clip association	–	Content	46.90%	50.00%	0.00%	100.00%
		Control	47.29%	50.02%	0.00%	100.00%
		Paralinguistic	46.51%	49.98%	0.00%	100.00%
Form recognition	Pretest	Content	25.64%	43.81%	0.00%	83.33%
		Control	28.21%	45.14%	0.00%	83.33%
		Paralinguistic	27.56%	44.83%	0.00%	100.00%
	Posttest	Content	70.93%	45.50%	0.00%	100.00%
		Control	72.48%	44.75%	0.00%	100.00%
		Paralinguistic	68.99%	46.34%	0.00%	100.00%
Meaning recognition	Pretest	Content	43.80%	49.71%	0.00%	100.00%
		Control	40.70%	49.22%	0.00%	83.33%
		Paralinguistic	42.25%	49.49%	0.00%	83.33%
	Posttest	Content	80.23%	39.90%	33.33%	100.00%
		Control	80.62%	39.60%	16.67%	100.00%
		Paralinguistic	82.17%	38.35%	16.67%	100.00%
Meaning recall	Pretest	Content	8.91%	28.55%	0.00%	66.67%
		Control	13.95%	34.72%	0.00%	83.33%
		Paralinguistic	9.30%	29.10%	0.00%	66.67%
	Posttest	Content	38.76%	48.81%	0.00%	100.00%
		Control	36.05%	48.11%	0.00%	100.00%
		Paralinguistic	48.06%	50.06%	0.00%	100.00%

Note. PSTM – Phonological short-term memory; Content – Content-related nonverbal; Paralinguistic – paralinguistic-related nonverbal; Control – Verbal-only.

for the interactions are presented in Figs. 6 and 7.

First, interpreting the three-way Time × Condition × Comprehension interactions, model results suggests that learners' comprehension of the video clips positively moderated how much they learnt under the three different conditions in terms of form recognition. With every unit increase in learners' comprehension of the video clips, they were 4.62 times more likely to make larger gains in the content condition than in the control (verbal only) condition. Similarly, when the input condition was Paralinguistic, with every unit increase of comprehension, learners were 4.71 times more likely to make larger gains than when the condition was control. In addition, the two-way Time × X-Lex interactions suggested that learners' pre-existing vocabulary size negatively moderated the gains in form

Table 3
Model results for form recognition.

Predictors	Form recognition	
	Odds Ratios	CI (95%)
Intercept	0.34	0.18–0.64
Time $_{\text{Post-pre}}$	10.42	6.26–17.82
Condition $_{\text{Content-Control}}$	0.82	0.45–1.41
Condition $_{\text{Paralinguistic-Control}}$	0.92	0.50–1.61
Comprehension	1.27	0.43–3.85
X-Lex	12.64	5.21–34.10
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Content-Control}}$	1.16	0.58–2.38
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Paralinguistic-Control}}$	0.91	0.46–1.90
Time $_{\text{Post-Pre}}$ × Comprehension	1.51	0.51–3.71
Condition $_{\text{Content-Control}}$ × Comprehension	0.29	0.09–0.80
Condition $_{\text{Paralinguistic-Control}}$ × Comprehension	0.29	0.09–0.86
Time $_{\text{Post-Pre}}$ × X-Lex	0.19	0.09–0.38
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Content-Control}}$ × Comprehension	4.62	1.30–17.96
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Paralinguistic-Control}}$ × Comprehension	4.71	1.35–19.03

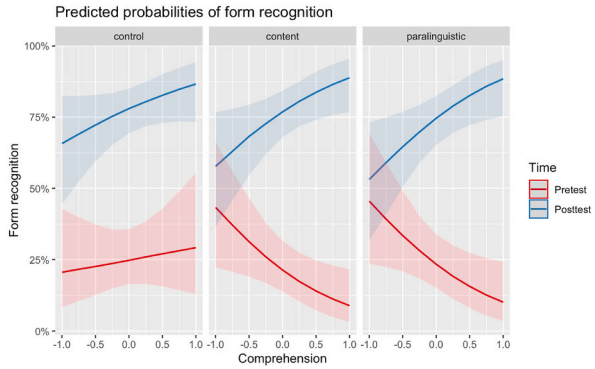


Fig. 6. Time × condition × comprehension interactions for form recognition.

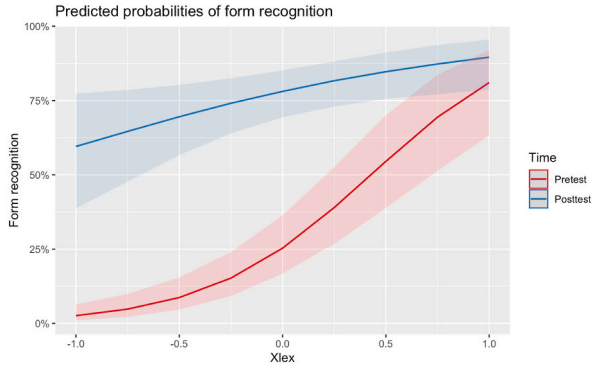


Fig. 7. Time × X-Lex interactions for form recognition.

recognition. When learners' X-Lex scores decreased by one unit, they were 5.26 (1/0.19) times more likely to make larger form recognition gains. Finally, there was no Time × Condition effect, suggesting that when learners' comprehension scores were centered at the mean, they benefited similarly from the three input conditions.

6.2. *Clip association*

Turning to the model for clip association, the final model only included two fixed factors (Comprehension and X-Lex). PSTM and Condition were removed through model selection, suggesting that they did not show any effect on clip association. Altogether 11% of the variance was explained by the fixed effects and 30% of the variance was explained by both the fixed and random effects. Model results indicated that there was an effect of X-Lex. With one unit of increase of learners' X-Lex scores, they were 3.56 times (95% *CI* [1.84, 6.82]) more likely to associate the target words with the correct film clips. Similarly, Comprehension also showed an effect. When learners' comprehension scores increased by one unit, they were 2.34 times (95% *CI* [1.22, 4.69]) more likely to correctly

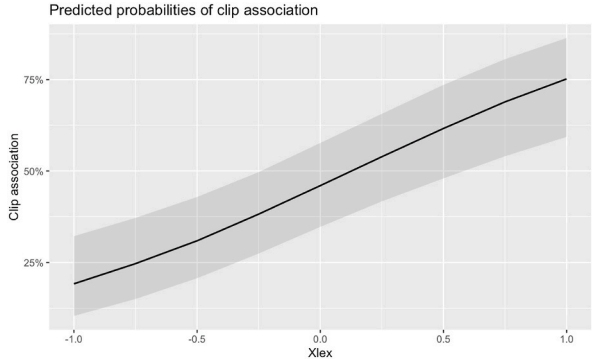


Fig. 8. The effect of X-Lex for clip association.

associate the target words with the clips in which they were embedded. The effect plots for X-Lex and Comprehension are given in Figs. 8 and 9 respectively.

6.3. Meaning recognition

The final model for meaning recognition included three fixed factors: Time, Comprehension, and X-Lex. There were also two-way Time × Comprehension interactions. PSTM and Condition were removed during model selection, indicating they were unlikely to have any effect on meaning recognition. These fixed effects explained in total 26% of the variance and 36% of the variance was explained by both the fixed and random effects. Model results first showed that learners' comprehension positively moderated how much they gained in meaning recognition (see Fig. 10 for the effect plot). When their comprehension scores improved by one unit, learners were 4.28 times (95% *CI* [2.40, 7.86]) more likely to make gains in meaning recognition of the target words. In addition, X-Lex showed an overall positive effect (see Fig. 11 for the effect plot). With every unit increase of X-Lex scores, learners were 2.95 (95% CI for ORs [1.82, 4.83]) more likely to successfully recognize the meaning of the target words. The fact that X-Lex did not interact with the fixed factor of Time, however, suggested that learning gains were not moderated by X-Lex.

6.4. Meaning recall

Lastly, regarding meaning recall, the final model (see Table 4) included four fixed factors (Time, Condition, Comprehension, and X-Lex). There were also two-way Time × Condition, Time × Comprehension, and Condition × Comprehension and three-way Time × Condition × Comprehension interactions. Model fit indices showed that in total 32% of the variance was explained by the fixed effects and 56% of the variance was explained by both the fixed and random effects. First, interpreting the three-way Time × Condition × Comprehension interactions (see Fig. 12), model results indicated that the difference in meaning recall gains between the Control and Paralinguistic conditions was positively moderated by learners' comprehension. With every unit increase in comprehension, learners were 20.35 times more likely to make gains when they experienced the Paralinguistic condition than when they were given the Control condition. Learners' comprehension scores, however, did not show any moderation effect on the learning differences between the Content and Control conditions ($OR = 0.91$).

In addition, the two-way Time × Condition interactions suggested that when learners' comprehension scores were set at the mean, the likelihood of making meaning recall gains was 3.31 times higher when the Condition was Content than when it was Control and was 4.35 times higher when the Condition was Paralinguistic than when it was Control. Learners' X-Lex scores showed an overall effect on meaning recall regardless of the test time point and input condition (see Fig. 13). With every unit increase in X-Lex scores, learners were 11.03 times more likely to successfully recall the meaning of the target words. This effect, however, did not interact with the fixed effect of Time, meaning that X-Lex did not show any moderation effect on learning gains.

7. Discussion

7.1. How do different kinds of multimodal input affect foreign language learners' vocabulary learning from explicit instruction?

Our findings demonstrate that the effects of Condition (Control vs. Content vs. Paralinguistic) were only observed for meaning recall. Learners made larger meaning recall gains for the target words when they were given either additional paralinguistic-related or content-related nonverbal input than when they experienced the control, i.e., verbal-only condition. Within the two nonverbal conditions, the effects were slightly larger for the paralinguistic-related nonverbal than for the content-related nonverbal. These findings overall support the cognitive theory of multimedia learning (Mayer, 2017). Learners were able to process the additional nonverbal information simultaneously with the verbal input in their semantic memory (Paivio, 1986). Doing so has thus deepened learners' overall understanding of input and facilitated their learning from the input.

The present study, however, found that for all other knowledge aspects, i.e., form recognition, clip association, and meaning

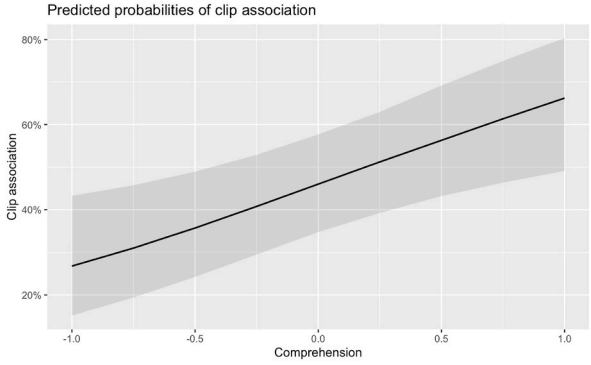


Fig. 9. The effect of comprehension for clip association.

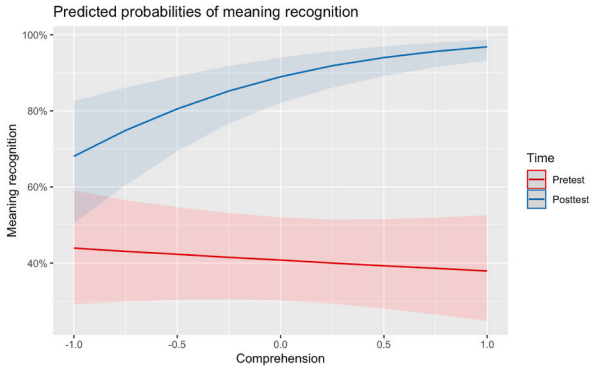


Fig. 10. Time × comprehension interactions for meaning recognition.

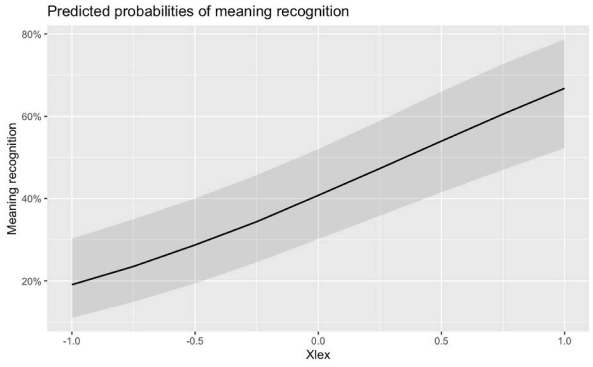


Fig. 11. The effect of X-Lex for meaning recognition.

Table 4
Model results for meaning recall.

Predictors	Meaning recall	
	Odds Ratios	CI (95%)
Intercept	0.01	0.00–0.05
Time $_{\text{Post-pre}}$	23.14	6.69–109.48
Condition $_{\text{Content-Control}}$	0.40	0.16–0.93
Condition $_{\text{Paralinguistic-Control}}$	0.54	0.22–1.24
Comprehension	10.76	1.52–128.69
X-Lex	11.03	4.16–32.53
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Content-Control}}$	3.31	1.24–8.70
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Paralinguistic-Control}}$	4.35	1.64–11.87
Time $_{\text{Post-Pre}}$ × Comprehension	0.49	0.04–4.16
Condition $_{\text{Content-Control}}$ × Comprehension	1.07	0.12–9.19
Condition $_{\text{Paralinguistic-Control}}$ × Comprehension	0.19	0.02–1.46
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Content-Control}}$ × Comprehension	0.91	0.09–9.96
Time $_{\text{Post-Pre}}$ × Condition $_{\text{Paralinguistic-Control}}$ × Comprehension	20.35	1.84–325.49

recognition, learners benefited similarly across the three conditions. This finding, albeit in line with Boers et al. (2017) where meaning recognition gains for the target words did not significantly differ regardless of the presence or absence of the content-related nonverbal input, is different from most other previous studies investigating the role of nonverbal input in incidental vocabulary learning. In such studies, having content-related nonverbal input was found to be beneficial for form recognition (Teng, 2023b) and meaning recognition (Teng, 2023b; Warren et al., 2018), and receiving paralinguistic-related nonverbal input helped meaning recognition (Arndt & Woore, 2018). This is not surprising considering that the explicit vocabulary instruction employed in the current study likely promoted learners' "noticing" (Schmidt, 1990) of the target words. Such noticing may be particularly useful for gaining types of vocabulary knowledge that are less demanding, such as form and meaning recognition (receptive knowledge of the form and meaning). Therefore, for these types of knowledge, any effects that the additional nonverbal input had on learning gains may have been less observable. For those studies examining incidental vocabulary learning (Arndt & Woore, 2018; Teng, 2023b; Warren et al., 2018) where the element of

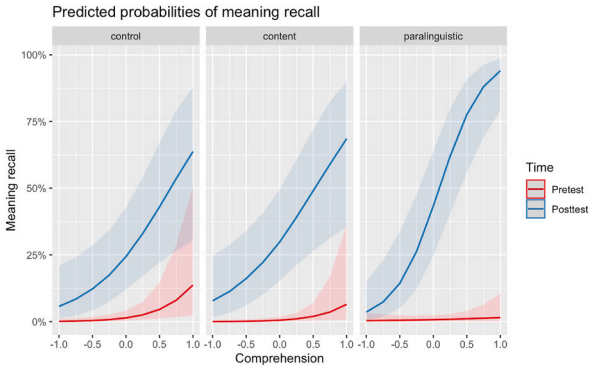


Fig. 12. Time × condition × comprehension interactions for meaning recall.

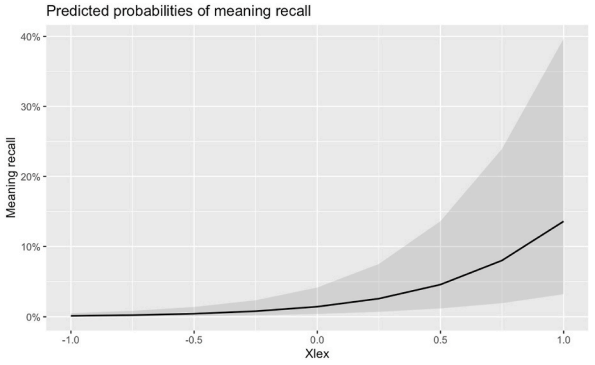


Fig. 13. The effect of X-Lex for meaning recall.

noticing was likely much reduced however, the additional benefits that nonverbal input brings were more observable.

7.2. To what extent are the effects of vocabulary learning from different kinds of multimodal input moderated by learner-related factors?

7.2.1. Comprehension of the input

The current study, for the first time, investigated how learners' comprehension of the multimodal input (i.e., the three film clips) moderated vocabulary learning. Findings suggest that comprehension overall positively moderated vocabulary learning gains. Regardless of the input conditions, learners with better comprehension of the film clips made larger gains in meaning recognition and clip association in comparison with their counterparts with poorer comprehension. Better comprehension may have allowed those learners to draw richer comprehensible input out of the film clips, which then allowed them to associate more words with relevant clips and recognize more words' meanings.

For form recognition and meaning recall, comprehension further moderated the learning gains between different input conditions, evidenced by the three-way Time × Condition × Comprehension interactions. Learners benefited more from the two nonverbal conditions than from the control (verbal-only) condition in terms of form recognition when their levels of comprehension increased. They were also more advantaged by the paralinguistic condition than by the other two conditions (content-related nonverbal and verbal-only) for meaning recall within every unit increase of their comprehension levels. These complex findings suggest that although content-related nonverbal input seemed to have helped learners with better comprehension to recognize more target word forms, paralinguistic-related nonverbal input went one step further by helping these learners recall more word meanings, and therefore facilitated more in-depth learning. Possible explanations for these findings could be that learners with better comprehension of the film clips were those who were more sensitive to the nonverbal cues (Gardner, 1999) and hence paid more attention to both forms of nonverbal input. As verbal and nonverbal information are processed in different parts of semantic memory (Mayer, 2017), those learners then gained more information overall from the explicit vocabulary instruction than their peers who were less sensitive to the nonverbal cues. In addition, paralinguistic-related nonverbal input took the form of videos whereas content-related nonverbal input appeared as static pictures. This meant that the former provided richer nonverbal input than the latter, hence triggering more in-depth learning (meaning recall).

7.2.2. Prior vocabulary size

Regarding the moderation effects of learners' prior vocabulary size measured by X-Lex, a Yes/No form recognition test, results

revealed that it positively moderated clip association. Learners with a larger prior vocabulary size tended to correctly associate more words with their corresponding film clips than their peers with a smaller prior vocabulary size. This supports what was found by Montero-Perez et al. (2014). A larger French vocabulary size may have helped relieve learners of some cognitive burden while processing the three film clips. This then allowed them to pay more attention to the target words appearing in the clips and form a stronger connection between those words and the theme of the clips.

Unlike previous studies which also found that learners' pre-existing vocabulary knowledge positively moderated gains in form and meaning recognition (Montero-Perez et al., 2014; Teng, 2023a) and meaning recall (Peters & Webb, 2018), in the current study, however, X-Lex scores did not show any moderation effects on meaning recognition or recall. Surprisingly, we found X-Lex scores negatively moderated the form recognition gains, meaning the gains were smaller for learners with greater vocabulary knowledge than for those with less knowledge. This contradicts previous studies and also our own findings for clip association. There are potentially two reasons behind this. First, the nature of the explicit vocabulary instruction that the current study employed is very different from the incidental vocabulary learning (Hulstijn, 2001) investigated in the previous studies. The explicit instruction may have been particularly useful for learners with less vocabulary knowledge, drawing their attention to the target words and their forms and meanings. For those with higher levels of pre-existing vocabulary knowledge, on the other hand, although they also benefited from such instruction, the gains were not as large as their peers who knew fewer words than them before the experiment. Second, clip association was only measured at the post-test but not at the pre-test as learners had not yet been exposed to the clips at the pre-test. Clip association is also less closely related to the explicit vocabulary instruction. All other knowledge aspects, however, were measured at the pre-test to control any existing knowledge learners had and were more likely to be affected by the explicit vocabulary instruction. A different learning pattern was therefore observed for these knowledge aspects to that observed for clip association.

7.2.3. capacity

For all measurements, the present study did not find any moderation effects of PSTM capacity. During the process of model selection, PSTM capacity was removed as a fixed factor as it did not explain further any variance within the data. This is different from Montero-Perez (2020), who found an effect of complex working memory capacity (measured by backward digit span and operation span) but failed to find an effect of PSTM capacity (measured by forward digit span), and Teng (2023a) who revealed an effect of PSTM (measured by nonword repetition) but not an effect of complex working memory capacity (measured by operation span). One potential explanation could be that comprehension of the input which was found to have moderated gains on all knowledge aspects was not examined in the previous studies. In the current study, the dominant moderation effects found for comprehension may have cancelled out other factors such as working memory, making the effects of those factors less observable in the analysis. Moreover, compared to incidental vocabulary learning, where learners had to pick up the target words within a restricted period of time, the explicit vocabulary instruction in this study allowed the learners to spend more time on memorizing the target words, which might ease the load on PSTM capacity. Therefore, differences in this factor may not contribute to explaining the vocabulary gains.

8. Limitations

The current study would certainly benefit from having a bigger sample size. Although the sample was recruited from different schools in England, representing a multi-site sample (Vitta et al., 2021), and had decent statistical power judged by the analysis, the overall sample size is still relatively small. A larger sample would allow future studies to include more complex model structures, further exploring the interplay between the learner-related factors on vocabulary learning through multimodal input. In addition, considering the contradictory findings between the present study and two previous studies (Montero-Perez, 2020; Teng, 2023a) regarding the effects of working memory, follow-up studies may consider adopting a more comprehensive combination of different working memory tests for low proficiency young foreign language learners to capture the moderation effects of working memory more accurately. Finally, the study was not able to administer a delayed post-test due to time constraints. A longitudinal study consisting of multiple intervention sessions with a delayed post-test can provide a clearer picture of the long-term effects of learning from such type of explicit vocabulary instruction using multimodal input.

9. Conclusion

The study is the first to investigate, among low-proficiency young learners, the effects of different kinds of input (verbal-only vs. verbal plus content-related nonverbal vs. verbal plus paralinguistic-related nonverbal) on foreign language learners' vocabulary learning from explicit instruction and how the learning was moderated by three learner-related factors (prior vocabulary size, phonological short-term memory, and comprehension of the input). Findings highlighted that multimodal input including additional content-related and paralinguistic-related information was more beneficial than traditional verbal-only input in helping foreign language learners learn more demanding productive vocabulary knowledge. All types of input were equally helpful for learning less demanding receptive vocabulary knowledge. At a pedagogical level, these findings suggest that when selecting materials for vocabulary instruction, attention needs to be given to balancing the use of verbal and nonverbal input. Adding additional content-related pictures representing the meaning of the target words and providing video-based narrations with paralinguistic nonverbal input for any vocabulary teaching materials can be particularly useful for retaining productive knowledge of new vocabulary.

Regarding the three learner-related factors, findings revealed that comprehension of the multimodal input seemed to be the most important moderation variable for the learning gains, followed by learners' pre-existing vocabulary size. This was particularly true when the input included paralinguistic-related nonverbal information. These findings provide novel insights into how the redundancy

principle (Mayer, 2017) can be interpreted in the foreign language learning context. That is, although multimodal input seems to have positive effects on foreign language learning overall, there is a linguistic threshold for that supportive aspect to kick in. Less proficient foreign language learners may not benefit from multimodal input as much as their counterparts with higher proficiency levels. From a pedagogical perspective, these findings suggest that how comprehensible the multimodal input is should be a crucial factor to take into account when designing vocabulary learning activity, perhaps particularly for school-aged learners. Making sure that learners can reach certain levels of comprehension of the input independently is essential for them to further acquire knowledge from the input, which is an important goal in language instruction.

CRediT authorship contribution statement

Pengchong Zhang: Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Shi Zhang:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation.

Acknowledgements

This study was supported by University of Reading Research Fellowship (Improving Foreign Language in the UK: Transforming Classroom Language Teaching through Multimedia; A368700).

References

Arndt, H. L., & Woore, R. (2018). Vocabulary learning from watching YouTube videos and reading blog posts. *Language, Learning and Technology, 22*(3), 124–142, 10125/44660.

Austin, K. A. (2009). Multimedia learning: Cognitive individual differences and display design techniques predict transfer learning with multimedia learning modules. *Computers & Education, 53*(4), 1339–1354. https://doi.org/10.1016/j.compedu.2009.06.017

Baddeley, A. D. (1999). *Human memory*. Allyn & Bacon.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829–839. https://doi.org/10.1038/nrn1201

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing, 38*(4), 511–535. https://doi.org/10.1177/2F0265532220951504

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System, 66*, 113–129. https://doi.org/10.1016/j.system.2017.03.017

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software, 100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Collen, I. (2022). *Language Trends 2022: Language teaching in primary and secondary schools in England*. British Council. https://www.britishcouncil.org/about/press/british-council-language-learning-schools-2022.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Young learners' L2 English after the onset of instruction: Longitudinal development of L2 proficiency and the role of individual differences. *Bilingualism: Language and Cognition, 24*(3), 439–453. https://doi.org/10.1017/S1366728920000747

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *In International conference on computational social science*. Cologne.

Finlayson, N., Marsden, E., & Anthony, L. (2022). *MultilingProfiler (version 3)*. University of York [Computer software] https://www.multilingprofiler.net/.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. Basic Books.

Gelman, A. (2023). Prior choice recommendations. GitHub https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician, 73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning, 67*(4), 922–958. https://doi.org/10.1111/lang.12251

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press.

James, W. B., & Galbraith, M. W. (1985). Perceptual learning styles: Implications and techniques for the practitioner. *Lifelong Learning, 8*(4), 20–23.

Kalyuga, S., & Sweller, J. (2014). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbooks in psychology. The Cambridge handbook of multimedia learning* (pp. 247–262). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.013.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *The Canadian Modern Language Review, 63*, 149–166. https://doi.org/10.3138/cmlr.63.1.149

Lüdecke, D. (2022). *sjPlot: Data visualization for statistics in social science*. R package version 2.8.11 https://CRAN.R-project.org/package=sjPlot.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*(3), 379–413. https://doi.org/10.1017/S0272263112000125

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory, 13*, 422–429. https://doi.org/10.1080/09658210344000323

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning, 33*(5), 403–423. https://doi.org/10.1111/jcal.12197

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28*(5), 653–660. https://doi.org/10.1002/acp.3050

Meara, P. (1992). *EFL vocabulary tests*. CALS University of Wales Swansea.

Montero-Perez, M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition, 42*(4), 749–773. https://doi.org/10.1017/S0272263119000706

Montero-Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language, Learning and Technology, 18*(1), 118–141, 10125/44357.

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

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language, 22*(1), 31–55.

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, Á. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition, 42*(3), 577–598. https://doi.org/10.1017/S0272263120000091

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *Tesol Quarterly, 53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

R Development Core Team. (2022). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing [software] https://www.R-project.org. July 2023.

Sabourin, L., Leclerc, J.-C., Lapierre, M., Burkholder, M., & Brien, C. (2016). The language background questionnaires in L2 research: Teasing apart the variables. In L. Hracs (Ed.), *Proceedings of the 2016 annual conference of the Canadian linguistics association*. Calgary, AB.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Service, E., & Simard, D. (2022). How measures of working memory relate to L2 vocabulary. In J. W. Schwieter, & Z. E. Wen (Eds.), *The cambridge handbook of working memory and language* (pp. 529–549). Cambridge University Press. https://doi.org/10.1017/9781108955638.030.

Sivula, T., Magnusson, M., & Vehtari, A. (2020). *Uncertainty in Bayesian leave-one-out cross-validation based model comparison. arXiv:2008.10296* https://arxiv.org/abs/2008.10296.

St Clair-Thompson, H., & Allen, R. J. (2013). Are forward and backward recall the same? A dual-task study of digit recall. *Memory & Cognition, 41*(4), 519–532. https://doi.org/10.3758/s13421-012-0277-2

St Clair-Thompson, H., Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. *Educational Psychology, 30*(2), 203–219. https://doi.org/10.1080/01443410903509259

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Teng, M. F. (2022). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System, 105*. https://doi.org/10.1016/j.system.2022.102736

Teng, M. F. (2023a). Effectiveness of captioned videos for incidental vocabulary learning and retention: The role of working memory. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2023.2173613

Teng, M. F. (2023b). The effectiveness of multimedia input on vocabulary learning and retention. *Innovation in Language Learning and Teaching, 17*(3), 738–754. https://doi.org/10.1080/17501229.2022.2131791

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2022). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.5.1 https://mc-stan.org/loo/.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Vitta, J. P., Nicklin, C., & McLean, S. (2021). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition, 44*(5), 1424–1448. https://doi.org/10.1017/S0272263121000541

Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading: Evidence from eye-tracking. *Studies in Second Language Acquisition, 40*(4), 883–906. https://doi.org/10.1017/S0272263118000177

Woore, R., Graham, S., & Arndt, H. L. (2020). Online language learning for all (OLLA). PDC in MFL https://pdcinmfl.com/online-language-learning-for-all-olla/.

Zhang, R., & Zou, D. (2022). A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign language learning. *Computer Assisted Language Learning, 35*(9), 2790–2816. https://doi.org/10.1080/09588221.2021.1896555