Measuring the visual in audio-visual input

The effects of imagery in vocabulary learning through TV viewing

Geòrgia Pujadas and Carmen Muñoz
University of Barcelona

This exploratory study investigates the effects of imagery on word learning through audio-visual input. A total of 82 adolescent EFL learners were exposed to 8 episodes of a TV series under four conditions, depending on the language of the on-screen text (L1 or L2) and whether they were pre-taught target words or not. The effects of co-occurrence of the word with its image, and the image time on screen (ITOS) were explored, alongside frequency, proficiency, and learning condition variables. Results showed that both image-related variables and frequency predicted word-form learning, while only ITOS predicted word-meaning recall, with a longer exposure to image associated to higher gains, suggesting that, at this age and proficiency level, the images associated with the words can be conducive to learning.

Keywords: imagery, TV viewing, frequency, vocabulary learning, adolescents

Watching TV series for language learning purposes has become a popular activity for foreign language teachers and students alike. TV series (as well as other audio-visual materials such as movies, documentaries, or short clips) have the potential to increase learners' exposure to rich, authentic input beyond the time constraints of the classroom setting and provide a necessary input flood for vocabulary learning. Unlike artificial material created specifically for language learning purposes, TV series provide foreign language (FL) learners with a source of naturalistic spoken input that resembles real life, as the images and the contextual clues present in the video allow them to "view" the message as well as to listen to it (Baltova, 1994; Danan, 2004).

Research on vocabulary learning has shown that learners can indeed pick up new words incidentally through TV viewing (Peters & Webb, 2018; Rodgers &

https://doi.org/10.1075/itl.22019.puj | Published online: 6 March 2023
ITL - International Journal of Applied Linguistics ISSN 0019-0829 | E-ISSN 1783-1490

Webb, 2020), and that learning can also be boosted by the addition of on-screen text (e.g. Montero Perez, Peters, Clarebout, & Desmet, 2014; Peters, 2019) as well as by pre-teaching the words that will appear in the input (Montero Perez, 2019; Pujadas & Muñoz, 2019). Research has also shown that learners' proficiency and vocabulary size play a major role in the learning outcomes, with more advanced students benefiting the most from this type of authentic input. Studies in this area have also investigated several word-related factors that may mediate word learning in this context, such as frequency of encounters in the input (Uchihara, Webb & Yanagisawa, 2019; Webb, 2007) – often regarded as one of the key factors for learning. But while it has long been hypothesised that the images present in the video could be beneficial for vocabulary learning, research has just recently started to explore the extent to which imagery supports aural information (Peters, 2019; Pujadas, 2019; Rodgers, 2018).

The present study aims at researching the effect of imagery on word learning through TV series by adolescent EFL learners, alongside the effects of frequency, the language of the on-screen text, the addition of pre-teaching of target words, and of learners' proficiency. The study also seeks to explore the predictive strength of two different image-related measures: the co-occurrence of the visual and aural representation of a target word, and the length of exposure to the target word's visual referent on-screen.

Background

TV series and vocabulary learning

TV series have several characteristics that make them a suitable tool for vocabulary learning. Firstly, they are consumed in large quantities and are more engaging than other traditional activities – such as reading – especially amongst young people (Lindgren & Muñoz, 2013). Secondly, because they are serial in nature, TV series allow learners to accumulate background knowledge and build up familiarity with the characters and story lines as they keep watching episodes from the same program, as people tend to watch multiple, consecutive episodes of the same TV series rather than isolated episodes (Rodgers, 2013). Thus, learners encounter novel words and expressions in contextualized, lifelike situations. Thirdly, TV series are also likely to contain repeated encounters with low frequency words, and word families from the 4,000 to the 14,000 levels are more likely to reoccur in episodes from the same series – or the same genre – than in a random sample (Rodgers & Webb, 2011). Therefore, the more episodes you watch, the more exposure you would get to those words, and the more likely you are to learn them.

Most studies on vocabulary learning through audio-visual materials have focused on incidental vocabulary learning, that is, learning vocabulary as a by-product of another activity (e.g., watching a documentary for its informational content), and results consistently show that incidental vocabulary acquisition can occur through viewing short videos, movies, and TV series (e.g., Peters & Webb, 2018; Rodgers, 2013). Studies have also shown that the addition of subtitles (native language [L1] text) or captions (L2 text) can facilitate vocabulary learning, especially for learners whose proficiency level is not high enough to cope with the fast speech rate and online nature of the videos. L2 captions can help with speech segmentation (Charles & Trenkic, 2015), written and aural form recognition (Markham, 1999; Sydorenko, 2010), and form-meaning mapping (Winke, Gass, & Sydorenko, 2010), while L1 subtitles provide on-line translations (Danan, 2004), are processed automatically, and allow viewers to understand the input regardless of their proficiency level.

The overall consensus is that L2 captions are more advantageous for language learning and vocabulary acquisition because they provide more exposure to the target language than L1 subtitles (e.g., Danan, 2004; Vanderplank, 2010; Winke et al., 2010), although a few studies have also found benefits derived from subtitling – especially for learners with a low proficiency level (e.g. Bianchi & Ciabattoni, 2008), and for word-meaning learning (e.g. Pujadas & Muñoz, 2019). While most research has focused on adult learners, studies with young learners have suggested that they can also benefit from exposure to audio-visual input enhanced with either captions or subtitles (e.g. Avello & Muñoz, forthcoming), albeit studies have reported mixed results depending on proficiency and age.

The benefits brought by audio-visual materials can be additionally boosted within the FL classroom by pre-directing learners' attention to vocabulary. Studies in the field of extensive reading suggest that deliberately focusing attention on lexical items (e.g. with input enhancement) can increase learning rates (Elley, 1989; Lee, 2007; Hulstijn, 2013; Nation, 2015). If attention is pre-directed to specific words, learners might spend more time on them because they are made more salient. If we assume that more attention leads to more learning (Boers, 2018; Robinson, Mackey, Gass, & Schmidt, 2012), the odds of learning those words increase. Guessing meaning from context is, however, challenging – even with the additional help of images – as this type of input has a fast-paced nature. Providing explicit access to the meaning of unknown words in the form of glossaries or short pre-viewing activities may help learners make an initial form-meaning connection (Chai & Erlam, 2008; Montero Perez, Peters, & Desmet, 2018; Pujadas & Muñoz, 2019; Sydorenko, 2010; Webb, 2010a, 2010b; Yang, 2014).

TV series and imagery

The most obvious and unique feature of TV series, compared to other language learning activities, is the presence of imagery, which is a powerful mode of meaning-making (*see* The Douglas Fir Group, 2016). The images allow viewers to construct meaning through an additional source of non-verbal information, which can be processed automatically and regardless of the L2 proficiency level. Due to their limited word segmentation skills, listeners with low proficiency level seem to rely more on top-down processing than bottom up-processing. Imagery – which can be seen as a "compensatory mechanism" (Vandergrift, 2007, p. 193) – provides contextual knowledge that allows beginner-level learners to focus on details of the story, which in turn can have a positive impact on comprehension (Rodgers, 2013, 2016).

Several studies support the idea that the imagery associated with videos can assist information processing. Research on listening has shown that imagery has a positive effect on comprehension (e.g. Jones & Plass, 2002) – especially for beginner learners (Maleki & Safaee Rad, 2011) – and that it helps reduce anxiety when encountering unfamiliar topics (Hasan, 2000). The positive effect of visual clues is also reported by Baltova (1994), who found that learners with access to audio and video almost doubled the comprehension scores of the group with access to audio only. Durbahn, Rodgers and Peters (2020) assessed comprehension of a documentary through questions that were imagery-based, audio-based, and imagery plus audio-based. Results showed that, when imagery was available, learners relied less on the spoken text, whereas for audio-based questions the factor that played the most significant role was vocabulary knowledge. The value of images for comprehension seems especially important, and it might be argued that images could also play a key role in the first stages of form-meaning mapping and word-meaning learning (Peters, 2019).

While it has long been acknowledged that images can be beneficial for language processing, research has just started to investigate the degree to which image supports aural information. Rodgers (2018) compared the extent to which the aural forms of 90 target words co-occurred simultaneously with the visual representations of those words in the first season of two television programmes from different genres (narrative TV and documentary). Following on the temporal contiguity principle of multimedia learning – which states that students learn better when words and pictures are presented simultaneously rather than successively (Mayer, 2014) – it can be assumed that, for a learner to be able to use the images to infer the meaning of the unknown word, the image associated with that word should appear in close proximity with its aural form. Building upon the theories of multimedia learning, Rodgers argued that the temporal proximity of the aural form of a word and its visual representation may support word learning, as it facilitates processing by allowing learners to hold separate representations of a word (visual and aural) and build a better mental connection between them. Results showed that the imagery in documentaries potentially supported vocabulary learning more than in narrative TV, with 65% of the images co-occurring simultaneously with the aural forms of the target words, and over 70% co-occurring within a 10-second timeframe. In contrast, only 29% of the target items were found to have a visual representation in the narrative programme. Findings suggest that the extent to which the word's visual image and the word's aural form co-occurs may support vocabulary learning through L2 television viewing, although further research is needed to investigate the degree to which co-occurrence contributes to word learning. Additionally, while learners may be assisted by co-occurrence in their form-meaning matching because of the temporal contiguity of word and image, this matching process may also be dependent on the total amount of time the image is on screen.

Empirical studies looking at the relationship between imagery and gains in vocabulary are, however, scarce. In a pilot study with university students, Pujadas and Muñoz (2018) found a positive association between the co-occurrence of a word's visual image and aural representation and the learning rates for that word, and that target items that presented co-occurrence were better recalled. Peters (2019), in a pioneer study on imagery and incidental vocabulary learning using a full-length documentary, also observed that the words that occurred in close proximity to their visual representation were three times more likely to be learnt than the ones without image support, and that this was true for both form recognition and meaning recall. While results from her research also indicate that learning was mediated by other word-related factors (i.e. cognateness, frequency, and word familiarity), taken together these studies suggest that images provide some kind of automatic visual semantic support – which appears to vary across genres – and that the presence (or absence) of such support has a direct impact on vocabulary learning. Although not directly measuring imagery, a study by Suárez, Gilabert and Moskvina (2021) found higher vocabulary gains from watching an animal documentary relative to other genres. They also found a significant influence of vocabulary size in the other genres (sitcom, police procedural, edutainment) but not in the documentary. These findings led the researchers to suggest an explanation based on the higher imagery and context support in documentary. Further research on image support will shed light on the predictive power of this variable against other word-related variables.

Frequency of occurrence in viewing

Vocabulary studies suggest that one of the most prominent word-related characteristics that mediates word learning is frequency of occurrence. Research on incidental vocabulary learning through reading (e.g., Horst, Cobb & Meara, 1998; Pellicer-Sanchez & Schmitt, 2010), listening (e.g., Vidal, 2011; Van Zeeland & Schmitt, 2013) and more recently viewing (e.g., Muñoz, Pujadas, & Pattemore, 2023; Peters, Heynen & Puimège, 2016; Peters & Webb, 2018; Rodgers, 2013) has provided strong evidence that repeated encounters with unknown words in the input can facilitate learning, though the number of occurrences needed for substantial learning remains unclear (Uchihara et al., 2019), with a wide variation conditional on input mode (i.e. reading, listening, viewing) and word learning conceptualization (i.e. form recognition, meaning recall) (van Zeeland & Schimitt, 2013). Research has suggested that the importance of frequency might be less salient in spoken input than in written input (e.g., Brown, Waring & Donkaewbua, 2008; Van Zeeland & Schmitt, 2013; Vidal, 2011), and that listening requires more encounters given the transient nature of the mode, but that fewer encounters may be necessary when gestures are present (Gullberg, De Bot, & Volterra, 2008).

Studies on the effects of repetition in audio-visual input argue that frequency may play a different role in this media because of the presence of the images (Peters & Webb, 2018; Rodgers, 2013), and propose that 5 occurrences might be enough (e.g. Webb & Rodgers, 2009a) as the presence of visual support might compensate for a smaller number of repetitions. Due to the nature of TV materials – in which the number of encounters with a target word cannot be modified –, studies on this type of input have generally considered frequency of occurrence as an explanatory factor, instead of attempting to establish the exact number of occurrences needed for learning. Nevertheless, studies looking at the effect of frequency in video materials have generally reported a positive effect of repetition on incidental word learning. Rodgers (2013) found a small but significant correlation between frequency of occurrence and word learning gains in a demanding test of meaning recognition, although the correlation disappeared with an easier test. Peters, et al. (2016) also found a positive correlation between frequency and word learning, both for word form recognition (+10%) and meaning recall (+11%), but they found that the effect of frequency was mediated by the interaction of this variable with the learners' vocabulary size, with higher odds of learning a word when both parameters increased. Peters and Webb (2018) also found that frequency was positively related to word learning. Results from their first experiment showed that the odds of recalling a word's meaning were 25% higher per each additional occurrence of that word in the input (i.e. a full-length documentary). In their second experiment, assessing meaning recognition, the odds of a correct response were 20% higher when frequency of occurrence increased. That is, per every five occurrences of the target item, the chances of recognising it doubled. Data from both experiments suggest that the effect of frequency was slightly stronger for meaning recall than meaning recognition. Finally, Muñoz et al. (2023) looked at the effect of frequency for vocabulary learning together with the language of the on-screen text (L1 or L2), and found positive correlations between language gains and frequency in the input. Results also showed that the number of encounters was significantly associated with word-meaning gains for the subtitles group and with word-form gains for the captions group.

While widely regarded as a key variable in vocabulary learning, frequency has, in some cases, not emerged as a significant predictor for word learning (e.g. Webb & Chang, 2015). A recent study by Feng and Webb (2020) comparing vocabulary learning through reading, listening, and viewing, found that frequency of occurrence was not significantly related with incidental vocabulary learning in any of the three modes. The authors suggest that, in the case of viewing, the overlap between the images and the words could also have increased the potential for word-meaning learning.

Aim and research questions

The aim of the present study is to investigate the effects of imagery on word-form and word-meaning learning through the viewing of successive episodes of a TV series by adolescent EFL learners – an under-researched age group. The study will explore the potential effects of two image-related measures on word learning: the co-occurrence of the image and target word, and the amount of time the image associated with a target item appears on-screen, an aspect that, to the best of the authors' knowledge, has yet to be investigated. A second aim is to investigate whether vocabulary gains are also mediated by the frequency of encounters in the input, the language of the on-screen text, the pre-teaching of the target items prior to viewing, and the learners' proficiency level. More specifically, the study seeks to answer the following research questions:

1. What is the effect of imagery on word learning through audio-visual input with adolescent learners?
   1.1 To what extent do the two image-related measures predict learning?
2. What is the effect of frequency on word learning through audio-visual input with adolescent learners?

3. What are the effects of learner factors (i.e. proficiency) and learning conditions (i.e. language of the on-screen text, and pre-teaching) on word learning, and to what extent do they interact with imagery and frequency measures?

Methodology

Participants

A total of 106 secondary school learners (65 females, 41 males) in Grade 8 (13–14 years old) from a state school in the area of Barcelona were initially selected for the study. They were Catalan-Spanish balanced bilinguals, their proficiency level in English ranged from beginner to low-intermediate (from Pre-A to B1 according to the Common European Framework of Reference (CEFR), as measured by the Oxford Placement Test), and they had a mean vocabulary size of 1,959 words (as measured by the X_Lex test (Meara & Milton, 2003)). Prior to the intervention, around 55% of participants reported watching movies or TV series in English on a weekly basis (with or without on-screen text).

Four intact classes participated in the study, which took place over the course of an academic term (i.e. 3 months). Participants had been randomly allocated to classes by the school, and each class was assigned to a different viewing condition according to the language of the on-screen text (L1 or L2) and whether they were taught target vocabulary or not. Although all students took part in the study, only those with 85% attendance or more and who had completed all the tests were included in the analysis, leaving a total of 82 participants (51 females, 31 males), which were distributed as follows: captions + instruction (CI) (*n* = 23), captions + no instruction (CNI) (*n* = 21), subtitles + instruction (SI) (*n* = 21), and subtitles + no instruction (SNI) (*n* = 17).

Audio-visual materials

The first 8 consecutive episodes from the TV series *Fresh off the Boat* (Khan et al., 2015) were selected for this study. The series was chosen for its appropriate format (a sitcom), length (20-minute episodes), and age-appropriate content, as well as for the fact that it had not been aired in Spain at the time the study took place, which minimized the possibility that participants had watched any of the episodes before.

The 8 episodes were analysed using the RANGE software (Nation & Heatley, 2002). The analysis of the lexical profile showed that, on average, the episodes reached 94.37% coverage at the 2,000 word-level plus proper nouns and marginal words. Since participants in the present study had a mean vocabulary size of around 2,000 words[1] and they had the additional support of the on-screen text (in the L1 or L2), it was considered that the input was challenging enough to promote learning but not overwhelming (Krashen, 2003). According to Webb and Rodgers' (2009b) corpus study, the mean lexical coverage of comedy is 93.99% at the 2,000-word level, which makes this series a typical example of the genre.

Instruments

Initial general proficiency was assessed through the Oxford Placement Test (OPT) which was administered at the beginning of the term. The OPT scores (hereafter *proficiency*) were used because they provide a general measure of proficiency, including a section on listening, which was deemed relevant in this learning environment. Knowledge of the target items was assessed through a pre- and post-test, which consisted of two parts: (1) an aural form recognition and written form transcription test, and (2) a meaning recall test. Participants had to listen to each target item twice, write down the English word, and then provide a translation or a short definition. This type of test was chosen to be congruent with the input-modality (Jelani & Boers, 2018).

A total of 40 items (5 per episode) were originally selected, according to frequency of occurrence within the episode (they had to appear at least twice), and the low likelihood of being known known by participants at this level of proficiency, after consultation with their teachers. Due to the nature of the two image-related variables used in the study, the present analysis focused on nouns ($n = 20$).[2] The distribution of target nouns (TNs) across episodes was not regular.

Imagery measures

The extent to which imagery in the TV series supported aural information was explored through two measures: image co-occurrence and image time on screen.

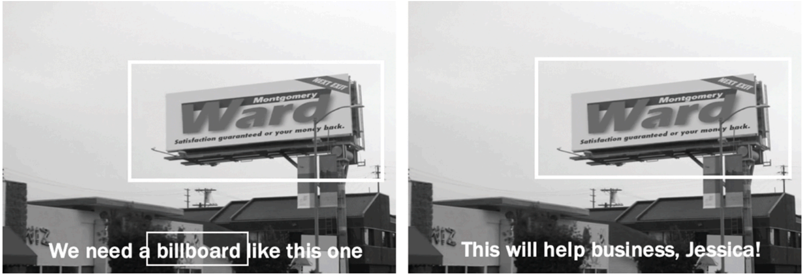
[1] Note that the RANGE software (measuring the episodes' lexical coverage) and the X_Lex test (measuring learners' vocabulary size) are based on different word-lists. A validation study by Miralpeix (2012), however, has shown that the results of the Levels Test and the X_Lex are comparable. The X_Lex test provides a total score out of 5,000 words by adding up the knowledge in each of the 5k word-families but does not provide information at each level. However, it seems logical to assume that – out of a score of 2,000 words – most of the words would be from the first or second thousand word-bands, even if some of the known words come from higher bands (Miralpeix, personal communication, December 18, 2018).

[2] The study was embedded in a larger study, that included an analysis of learning gains for all 40 target items (see Pujadas & Muñoz, 2019).

Similar to Rodgers' (2018) and Peters' (2019) studies, co-occurrence (i.e. CoO) was operationalized as the simultaneous occurrence of the visual representation of a target item and its aural (sound) and written (on-screen text) forms in a timeframe of five seconds before or after the occurrence of the item. The rationale behind this timeframe is that the on-screen text only appears on-screen for a maximum of six seconds, and using a +5/−5 seconds limit ensures that the word would have occurred "within an established processing amount" (Rodgers, 2018, p. 201). CoO was coded as a binary variable (words were either image-supported or not image-supported).

For the present study, a new measure was developed: the image time on screen (i.e. ITOS). ITOS refers to the amount of time the image of a TN appears on screen. As research suggests that the co-occurrence of a word with its image can be conducive to learning (Peters, 2019), it might be the case that a longer exposure to the image of the word also facilitates learning, and especially word-meaning learning. A longer ITOS – and therefore longer access to the visual, semantic representation of an unknown word – could make the word more salient, and allow L2 learners to have more time to process the information.

ITOS was operationalized as the total amount of time (in seconds) in which the visual representation of a TN was present on screen, independently of when or if the TN was uttered simultaneously. Figure 1 exemplifies the difference between the two image-related variables. The image on the left represents co-occurrence of the visual, aural, and written form of the TN "billboard;" the image on the right illustrates the measure ITOS, where the image of a billboard is present on the screen (for X number of seconds) without the word billboard being simultaneously uttered.


NEXT EXIT
Montgomery
Ward
Satisfaction guaranteed or your money back.
We need a billboard like this one
NEXT EXIT
Montgomery
Ward
Satisfaction guaranteed or your money back.
This will help business, Jessica!

Example of co-occurrence
of the aural/written word and the image

Example of image time on-screen
(independently of the aural/written word)

Figure 1. Comparison of CoO and ITOS measures
Note: Still images from the series *Fresh off the Boat* (Khan, et al., 2015)

The ELAN software was used to calculate both imagery measures. This program, developed by The Language Archive (URL: https://tla.mpi.nl/tools/tla-tools/elan), allows to create frame by frame annotations on video and audio files. To calculate the ITOS, an annotation was made for every time a TN's image appeared on screen, setting the beginning of the annotation the second the image appeared and closing it the moment it disappeared. Then, the total number of seconds per annotation was calculated and each instance was then added up to obtain the total amount of image time on-screen per each TN. A second rater assessed the TNs' image-support, reaching a 96% agreement. Unclear cases were discussed until an agreement was reached. Note that the location of the TN in the scene (e.g., in the background, in a close-up) was not taken into account for either of the two measures, and that ITOS refers to the total amount of time a word's referent was shown on-screen across the 8 episodes (see *Spacing* in Table A in the Appendix). While most TNs only appeared in one episode, 5 appeared in more than one.

Procedure

The classroom intervention took place during 11 weeks and was embedded as a part of regular English lessons. Participants were pre-tested at the beginning and at the end of the term to assess their knowledge of the corresponding 40 TIs included in the 8 episodes for that term, from which half of them were TNs ($n=20$). The proficiency test was administered first (week 1) and the pre-test was administered a week before the first session (week 2), to reduce pre-test effects. Then, participants had 8 viewing sessions, one per week. The post-test was administered a week after the last viewing session (week 11).

The two groups with instruction (one with captions, one with subtitles) started the sessions with a short, 5-minute pre-viewing task aimed at teaching the aural and written form as well as the meaning of the five TIs (plus three distractors) appearing in the episode. These activities – which included matching exercises, word searches, fill-in-the blanks tasks and crosswords – were completed autonomously by the learners and were corrected orally by the teacher. Then, participants watched the episode (with either captions or subtitles, depending on the group) and completed two immediate post-viewing tasks, namely a vocabulary task and a content comprehension task,[3] which were not corrected in class. The

[3] The vocabulary post-tasks were aural form transcription and meaning recognition tasks: participants heard a word twice, wrote it down, and selected the correct translation from 5 options provided. The task included 5 target items plus 3 distractors. The comprehension

two groups without instruction did not complete the pre-viewing task, but the rest of the session (i.e. episode viewing and post-tasks) was identical.

Scoring of vocabulary tests

Pre- and post-tests were scored dichotomously. For word-form, 1 point was given when the word was correctly spelled. Due to the large variability of transcriptions in the sample, a strict spelling-based criterion was adopted, and this test format could show the potential advantage given by captions, which expose learners to the written form of the words. For word-meaning, translations and short definitions were scored by two raters, with an interrater reliability of 94.5% (conflicting cases were discussed until an agreement was reached). A list of the accepted translations was elaborated to ensure that the same correction criteria was followed from pre- to post-test.

Preliminary analysis

Detailed information on all 20 target nouns (TNs) can be found in Table A in the Appendix. In the present sample, 13 TNs presented co-occurrence of the word utterance and its visual representation, whereas 7 TNs did not. The ITOS ranged from 4 to 128 seconds, with a mean length of approximately 40 seconds on screen. Since it was not linearly distributed, this variable was re-categorised in four levels: none (no visual representation), low (4–11 seconds), mid (17–36 seconds), and high (90–128 seconds). Because of the strong association between both image-related variables, analyses were run separately to compare their predictive power. Concreteness ratings (Brysbaert, Warriner & Kuperman, 2014) were used to check whether TNs with and without onscreen imagery were comparable in terms of concreteness (or imageability) of meaning, as this factor is known to facilitate word learning and meaning recall (e.g. Deconinck, Boers & Eyckmans, 2017). A Welch's ANOVA showed that there were no significant differences in terms of concreteness between image-supported and non-image-supported TNs ($F(1,7.764) = 4.226$, $p = .075$), although image-supported TNs had a slightly higher concreteness rating (*M* 4.66; *SD* 0.42) than the ones without image support (*M* 3.98; *SD* 0.82).

Frequency of occurrence varied from 2 to 10 encounters, with a mean frequency of 5.25. While most of the encounters with the TNs were generally in the episode in which they were taught in the groups with instruction, TNs could also

tasks included 10 items (5 multiple-choice, 5 true-false) assessing content comprehension (see Pujadas & Muñoz, 2020).

be encountered again in successive (or prior) episodes. Since learning gains were assessed at the end of the term, all encounters across the 8 episodes were taken into account for this measure. Similar to ITOS, frequency of occurrence was also re-categorized in low (2–3 encounters), mid (4–7 encounters), and high (8–10 encounters).

There were no significant differences between experimental groups in terms of proficiency ($F(3,78) = .829$, $p = .482$) nor vocabulary size ($F(3,72) = .845$; $p = .474$). Exploration of the data showed that this variable was not linearly distributed, so it was re-categorised into three levels, distributing participants in the following CEFR groups: Pre-A ($n = 28$), A1 ($n = 38$) and A2/B1 ($n = 16$).

Results

Table 1 shows the number of correct and incorrect responses for the 20 TNs, divided by experimental condition. Items known in the pre-test were excluded from the analysis.[4] As can be observed, overall, the two groups who had been pre-taught the words had 14.1% more correct responses than the other two groups in form transcription, and 8.4% in meaning recall.

Table 1. Number (and percentage) of correct and incorrect responses to TNs

	Form		**Meaning**	
	Correct responses	**Incorrect responses**	**Correct responses**	**Incorrect responses**
CI	101 (23.2%)	334 (76.8%)	56 (12.3%)	399 (87.7%)
SI	75 (18.5%)	330 (81.5%)	38 (9.1%)	379 (90.9%)
CNI	26 (6.3%)	390 (93.8%)	12 (2.8%)	422 (97.2%)
SNI	28 (7.6%)	339 (92.4%)	7 (1.9%)	365 (98.1%)
Total	**230 (14.2%)**	**1392 (85.8%)**	**113 (6.7%)**	**1565 (93.3%)**

Note: CI = Captions (L2-English) + Instruction; SI = Subtitles (L1-Spanish) + Instruction; CNI = Captions + No Instruction; SNI = Subtitles + No Instruction

[4] The percentage of known words in the pre-test represented a small percentage over the total amount of answers. More concretely, the percentages were as follows: 6.5% (CI), 4.1% (SI), 5.4% (CNI), and 2.8% (SNI).

Word-form learning

A first Generalized Linear Mixed Model (GLMM) with repeated measures was run to explore the effect of co-occurrence along with frequency, learning condition-related variables and learner-related variables. A first model was run with word-form gains (0 or 1) as the dependent variable, and co-occurrence with image (yes, no), frequency of occurrence (low, mid, high), language of the on-screen text (L1, L2), instruction (yes, no) and proficiency (Pre-A, A1, A2/B1) as fixed factors, as well as all two-way interactions between co-occurrence and the rest of the potentially mediating variables, and frequency and the rest of the variables. The model was based on 1493 observations. All non-significant interactions and main effects ($p < .10$) were then removed one by one.

Table 2 presents the final fitted model, and Table 3 provides information on the significant main effects. The model revealed that all fixed factors – except for language of the on-screen text – significantly contributed to the model ($p < .05$), and that there were no interactions.

Table 2. GLMM results for word-form learning, with CoO

						95% CI for Exp Coeff[a]	
Terms	**Coeff**	**SD**	**t**	**Sig**	**Exp Coeff**	**Lower**	**Upper**
Intercept	−1.397	.2746	−5.085	<.001	.247	.144	.424
CoO (no)	−.453	.1761	−2.575	.010	.636	.450	.898
CoO (yes)	0[b]	.	.	.	.	.	.
Frequency (low)	−.525	.1794	−2.927	.003	.592	.416	.841
Frequency (mid)	−.498	.1985	−2.508	.012	.608	.412	.897
Frequency (high)	0[b]	.	.	.	.	.	.
Instruction (yes)	1,338	.2378	5.624	<.001	3.810	2.390	6.075
Instruction (no)	0[b]	.	.	.	.	.	.
Proficiency (Pre-A)	−1.309	.2810	−4.659	<.001	.270	.165	.469
Proficiency (A1)	−.937	.2405	−3.896	<.001	.392	.244	.628
Proficiency (A2/B1)	0[b]	.	.	.	.	.	.

a. Confidence interval for Exponential Coefficient b. Coefficient is set to zero because is redundant

CoO appeared as a significant predictor ($F(1,1486) = 7.304$, $p = .007$), with words with visual support receiving a significantly higher percentage of correct responses than words without image representation (+4.6%). Frequency also emerged as a significant predictor of word-form learning ($F(2,1486) = 4.401$,

Table 3. Results from the GLMM 1: Fixed main effects for word-form learning

	Mean (SE)	M Diff (SE)	df	F	Sig.
CoO (yes)	14.00 (1.6)	4.60 (1.7)	1, 1486	7.304	.007
CoO (no)	9.40 (1.5)				
Frequency (low)^{a}	9.70 (1.5)	$^{a–b}$ 0.20 (1.8)	2, 1486	4.401	.012
Frequency (mid)^{b}	10.00 (1.6)	$^{b–c}$ 5.40 (2.2)			
Frequency (high)^{c}	15.40 (2.1)	$^{a–c}$ 5.70 (2.0)			
Instruction (yes)	20.20 (2.1)	14.00 (2.3)	1, 1486	36.049	<.001
Instruction (no)	6.20 (1.2)				
Pre-A^{a}	6.90 (1.4)	$^{a–b}$ 2.80 (1.9)	2, 1486	8.798	<.001
A1^{b}	9.70 (1.5)	$^{b–c}$ 11.80 (3.4)			
A2/B1^{c}	21.50 (3.3)	$^{a–c}$ 14.60 (3.5)			

$p = .012$). Pairwise contrasts revealed, however, that there were no significant differences in gains between TNs with low and mid frequency ($p = .981$), but only between low and high frequency ($p = .014$) and mid and high frequency ($p = .026$). The lack of interaction between the image-related and word-related variables and the other mediating variables indicates that the effect of imagery and frequency was present independently of the language of the on-screen text, instruction, and learners' proficiency.

As for the learning conditions and learner-related variables, overall, the two groups with instruction significantly outperformed the other two ($F(1,1486) = 36.049$, $p < .001$), with 14% more gains in word-form learning than their counterparts. Proficiency also emerged as a significant predictor ($F(2,1486) = 8.049$, $p < .001$), with the more advanced group (A2/B1) clearly, and significantly, outperforming the other two. While language of the on-screen text did not emerge as a predictor in the model, a closer examination of the data reveals that the captions group has a higher percentage of correct responses when combined with instruction, but the subtitles group performs better when instruction is not provided.

A second GLMM was run again – following the same procedure – with the measure ITOS as the image-related variable. The model revealed similar results, as can be observed in Table 4. The exploratory measure ITOS emerged as significant predictor of word-form learning ($F(3,1484) = 15.279$, $p < .001$), although only the words with the highest ITOS (i.e. +90 seconds) were significantly better learnt, compared to the other three levels. Frequency appeared again as a significant predictor ($F(2,1484) = 13.716$, $p < .001$), and, again, only TNs with the highest frequency were significantly better learnt than words with low and mid frequency (both $p<.001$), with no significant difference between low and mid frequency ($p=.191$).

Table 4. Results from the GLMM 2: Fixed main effects for word-form learning

	Mean (SE)	M Diff (SE)	df	F	Sig.
Instruction (yes)	24.00 (2.4)	16.70 (2.7)	1, 1484	37.575	<.001
Instruction (no)	7.30 (1.4)				
Pre-A^{a}	7.80 (1.6)	$^{a-b}$ 3.60 (2.2)	2, 1484	9.552	<.001
A1^{b}	11.50 (1.8)	$^{b-c}$ 14.50 (4.2)			
A2/B1^{c}	26.00 (4.0)	$^{a-c}$ 18.20 (4.2)			
Frequency (low)^{a}	8.50 (1.4)	$^{a-b}$ 2.60 (2.0)	2, 1484	13.716	<.001
Frequency (mid)^{b}	11.10 (1.9)	$^{b-c}$ 13.90 (3.2)			
Frequency (high)^{c}	25.00 (3.2)	$^{a-c}$ 16.50 (3.2)			
ITOS (none) ^{a}	8.50 (1.5)	$^{a-d}$ 28.50 (4.4)	3, 1484	15.279	<.001
ITOS (low) ^{b}	12.40 (2.3)	$^{b-d}$ 24.70 (4.5)			
ITOS (mid) ^{c}	7.30 (1.3)	$^{c-d}$ 29.80 (4.4)			
ITOS (high) ^{d}	37.10 (4.3)				

Word-meaning learning

The two models (one with CoO, one with ITOS) were run again to assess the effect of imagery on word-meaning learning. The models were based on 1544 observations. Results from the first model (see Table 5) showed that, in contrast with word-form learning, neither co-occurrence ($F(1,1535)=.810, p=.368$) nor frequency of encounters ($F(2,1535)=2.457, p=.086$) predicted word-meaning learning. Looking at the overall results, it can be observed that TNs with image support still tended to be better learnt, though the difference was not significant, while no clear pattern was observed for frequency.

Again, a second model was run with ITOS as a mediating variable (see Table 6). In contrast with the results for co-occurrence, the model revealed that there was a positive relationship between ITOS and word-meaning learning ($F(3,1535)=5.778$, $p=.001$). As with word-form learning, ITOS only predicted learning when the image was present the longest (+90 seconds) compared to any other ITOS length ($p<$ 001), while there were no significant differences in learning between none, low and mid ITOS. Similar to the prior model, frequency of occurrence did not appear to predict meaning learning ($F(2,1535)=2.542, p=.079$).

Table 5. Results from the GLMM 1: Fixed main effects for word-meaning learning

	Mean (SE)	M Diff (SE)	df	F	Sig.
Instruction (yes)	9.00 (1.5)	6.90 (1.5)	1, 1535	20.659	<.001
Instruction (no)	2.00 (0.7)				
Pre-A^{a}	2.20 (0.7)	$^{a–b}$ 1.90 (1.1)	2, 1535	5.300	.005
A1^{b}	4.10 (1.0)	$^{b–c}$ 4.80 (2.1)			
A2/B1^{c}	8.90 (2.2)	$^{a–c}$ 6.70 (2.1)			
Frequency (low)^{a}	6.30 (1.4)	$^{a–b}$ 2.60 (1.3)	2, 1535	2.457	.086
Frequency (mid)^{b}	3.70 (1.0)	$^{b–c}$ 0.30 (1.1)			
Frequency (high)^{c}	3.40 (1.0)	$^{a–c}$ 2.90 (1.4)			
CoO (yes)	4.80 (1.0)	0.90 (1.0)	1, 1535	.810	.368
CoO (no)	3.90 (1.0)				

Table 6. Results from the GLMM 2: Fixed main effects for word-meaning learning

	Mean (SE)	M Diff (SE)	df	F	Sig.
Instruction (yes)	10.20 (1.6)	7.80 (1.7)	1, 1535	20.761	<.001
Instruction (no)	2.40 (0.8)				
Pre-A^{a}	2.40 (0.8)	$^{a–b}$ 2.60 (1.3)	2, 1535	5.911	.003
A1^{b}	5.00 (1.1)	$^{b–c}$ 5.60 (2.5)			
A2/B1^{c}	10.50 (2.5)	$^{a–c}$ 8.10 (2.5)			
Frequency (low)^{a}	5.90 (1.2)	$^{a–b}$ 2.30 (1.2)	2, 1535	2.542	.079
Frequency (mid)^{b}	3.60 (1.0)	$^{b–c}$ 2.40 (1.5)			
Frequency (high)^{c}	6.10 (1.5)	$^{a–c}$ 0.20 (1.6)			
ITOS (none)	4.20 (1.0)	$^{a–d}$ 10.10 (2.8)	3, 1535	5.778	.001
ITOS (low)	4.00 (1.1)	$^{b–d}$ 10.30 (2.7)			
ITOS (mid)	2.60 (0.8)	$^{c–d}$ 11.70 (2.9)			
ITOS (high)	14.30 (2.9)				

Discussion

Imagery

Our first research question aimed at exploring the extent to which imagery supports vocabulary learning through audio-visual input, and investigating the respective value of two different image-related measures. The first measure explored was co-occurrence of the word's visual representation and its aural/written form. Results from the GLMMs showed that co-occurrence was positively related to word-form learning, with TNs that occurred simultaneously with their image having higher percentage of correct responses than TNs without CoO. More concretely, words that had the support of imagery were 1.57 times more likely to be learned than words without imagery. However, there was no significant effect of co-occurrence for meaning learning. Results fall partly in line with the findings reported by Peters (2019), who found positive benefits of co-occurrence for both form recognition and meaning recall, and that words that had imagery associated to them were three times more likely to be learned than words without imagery (Peters, 2019). This discrepancy in results in word-meaning learning might be due to the fact that participants in the present study were younger and less proficient, and that episodes were viewed only once, which in turn may explain the overall low vocabulary gains (cf. studies with repeated viewings; e.g. Naghizade & Darabi, 2015; Peters et al., 2016). While co-occurrence might have drawn their attention to form, it might not have been enough to help them make the connection with its meaning.

The support provided by images was also investigated through a new measure: image time on screen (ITOS). This variable, which takes into account how long the image of a target word is present on the screen, emerged as a significant predictor for both word-form and word-meaning learning. Results revealed, however, that only the words with the highest ITOS (+ 90 seconds) were significantly better learnt. This may suggest that, at this age and proficiency level, a minimum time on-screen may be necessary to benefit from the presence of the image, especially for word-meaning learning. For word-form learning, seeing the image of the word may make the word more salient and encourage deeper processing – and this could be achieved just by presenting the image and the words simultaneously. For meaning recall, however, it may seem that co-occurrence alone is not enough, but a longer appearance on screen could facilitate form-meaning connection, as learners do not need to hold the image in their mind but can access it on the screen while processing. Results from this analysis indicate that ITOS could be a better predictor than CoO when assessing the impact of imagery on learning. Because the viewing materials were authentic, full-length episodes of a TV series, however, there was little control over the items' image support and their distribution across the sessions, with a time on-screen ranging from as little as 4 seconds to up to 128, and so any generalization should be made with care.

The effects of these variables have been found in both instructed and non-instructed conditions, and across all three proficiency levels. This suggests that learners may make use of the image independently of whether words were pre-taught or not, and independently of their L2 skills. Considering that the group who received instruction had higher percentage of gains in both form and meaning, it is possible that for them the images worked as some kind of reinforcement tool, boosting the benefits of the instruction received. For the other two groups, on the other hand, the images might have served as a compensatory mechanism for the lack of instruction. Further research looking into how learners make use of the image in this context (e.g. immediate protocol recalls) could provide insight on this matter.

Frequency of encounters

Our second research question looked at the effect of frequency on word learning. Frequency of encounters was found to be a significant predictor of word-form learning, with higher frequency leading to higher word-form gains, but it did not predict word-meaning learning. This falls partly in line with results from previous studies on incidental vocabulary learning through audio-visual input, which also found a positive effect of increased frequency on learning (e.g., Peters, et al., 2016; Peters & Webb, 2018; Rodgers, 2013). For word-form learning, however, frequency only emerged as a significant predictor when the words were encountered 8 times or more, a number of repetitions slightly higher than other viewing studies (e.g. Webb & Rodgers, 2009a; Uchihara et al. 2019). This is not surprising, as research indicates that recalling words requires more encounters, especially for meaning recall (e.g., Pellicer-Sanchez & Schmitt, 2010; Webb, 2007). While the positive effect of repetition was found independently of L2 proficiency level, it is possible that, again, participants in the present sample – younger and less proficient than the typical study populations – may need a higher number of repetitions in order to benefit from them. This might also explain why no significant effects were found for word-meaning learning. As suggested by Feng and Webb (2020), it is also possible that, in the context of viewing, there might be other factors that play a more prominent role than frequency.

An unexpected finding was the lack of interaction between instruction and frequency of encounters, as it would seem that having been pre-taught the words would reduce the need for repetition, as the learners would already be aware of the upcoming unknown words and identify them more easily when encountered again. Further research on the effects of repetition combined with instruction could shed light on that regard.

Learning conditions and learner variables

Finally, our last research question looked into the effects of the language of the on-screen text, the addition of pre-teaching and the participants' proficiency level – three relevant variables within the EFL classroom context – and whether these variables mediated the effects of imagery and frequency in word-form and word-meaning learning. As shown above, results from the GLMMs revealed that there were no significant interactions between instruction, language and proficiency and the image- and word-related variables, indicating that the effect of imagery and frequency is independent of them.

Results showed that both pre-teaching and proficiency significantly predicted word learning (both form and meaning), independently of whether participants were watching the series with English captions or Spanish subtitles. The positive effect of pre-teaching is not surprising, as it is well known that intentional learning is significantly more efficient than incidental learning (Hulstijn, 2003), and results fall in line with findings in prior studies that show that a minimum amount of instruction can already yield significant positive effects on learning (Pujadas & Muñoz, 2019). Proficiency also emerged as a significant predictor of word learning, with more advanced learners obtaining higher gains, also in concordance with prior research findings in the field (e.g. Chen, Liu & Todd, 2018). Language of the on-screen text did not appear to predict learning, but an exploration of the data showed an interesting tendency: when words were pre-taught, the group with captions outperformed the group with subtitles in both form transcription and meaning recall, while it was the subtitles group that performed slightly better in meaning recall when there was no prior instruction. This may suggest that the instruction condition allowed learners to make a first connection between form and meaning in the pre-viewing activities, and having the oral and written forms in the same language (captions) allowed them to reinforce that connection (Webb & Nation, 2017), which in turn would help meaning recall. In contrast, for learners that were not pre-taught the words, having access to subtitles might have compensated for the lack of instruction, as they could use the L1 translations to connect the meaning to the L2 oral form, while this shortcut could not be used with captions. It is possible that the SNI group – with access to L1 translations – could follow the story more easily and thus devote more attentional resources to the unknown words' meanings.

Conclusions

This exploratory study has focused on investigating the effects of imagery in vocabulary learning through TV series by beginner, adolescent EFL learners, while taking into account the effects of frequency, the language of the on-screen text, the addition of pre-teaching, and learners' general proficiency, and it provides valuable evidence of the importance of image support for vocabulary learning through viewing, with data from multiple, successive full-length episodes of a TV series.

Results from this study show that visual support was accessed independently of the learning conditions and learner-related variables in this study (i.e. language of the on-screen text, instruction, and proficiency), and that the image associated with target words supported learning in narrative TV, confirming that the simultaneous presentation of a word and its visual representation can facilitate learning (Rodgers, 2018; Peters, 2019). An additional contribution has been the development of an exploratory new measure of image support (i.e. ITOS), which revealed that a longer image time on screen better supported word-form and – especially – word-meaning learning. This suggests that, while image co-occurrence might not be sufficient for recalling meaning at this age and proficiency level, a much longer exposure to the image associated with the word might allow these younger, less proficient students create a semantic match between the image and its aural / written form (Peters, 2019; Sydorenko, 2010).

Compared to other studies on incidental learning through viewing, the vocabulary gains in the non-instructed groups were relatively small. Differences may be due to learners' age (i.e. adolescents), proficiency level (i.e. beginners), and background (i.e. Spain, a traditionally dubbing country with few opportunities for exposure to the L2). Additionally, participants only viewed the episode once, while other studies with young learners – which reported higher gains – included repeated viewings (e.g. Naghizade & Darabi, 2015; Peters, et al., 2016). It is possible that, with a longer intervention and/or as learners become older and more proficient, these gains might increase. The objective of integrating extensive viewing in the classroom is that students improve over time, and to raise learners' awareness of the additive value of autonomous L2 television viewing in the long run.

While the main focus of research has been incidental learning settings, a unique feature of this study is that it provides evidence from two learning conditions concurrently – having explicit instruction and not having it (an incidental-like situation). Findings contribute to the emerging evidence that imagery in topic-related episodes can support learning, while suggesting that the benefits may not be limited to the incidental learning context.

The study presents several limitations that should be acknowledged. Firstly, it should be noted that the statistical power of the study is small; only a small number of items – with a variety of characteristics – were analysed, and thus our results are contingent on the TNs selected. Findings from this exploratory analysis, however, provided initial evidence that the image associated with videos supports word learning, and can provide a starting point for future research. Another drawback of the study might be the type of test used to evaluate learning, since a recall test (e.g., a translation test) is more difficult than a recognition test (e.g., multiple-choice test) (Jones, 2004), and it might have failed to measure partial knowledge of the word meanings – thus explaining the low gains obtained across conditions, especially in meaning recall. If a student could not identify orally a target item first, they could not provide a translation, but this does not signify that the learners could not recognise the word form if they encountered it, or that they did not know the meaning of the word. The strict spelling-based criterion may have also contributed to misrepresent what participants had learnt. Another aspect to consider is that, while the classroom-based setting makes the study more ecologically valid, the longitudinal nature of the study does not totally exclude outside learning.

Finally, the study did not consider other word-related factors, such as spacing, recency, saliency, relevance, or frequency of occurrence in corpus, which have been shown to play a role in word learning. The reduced number of target items and the variability within the sample (besides the four learning conditions and the range of proficiency levels), however, would have hindered the reliability of the results. More research controlling for imageability of meaning is also needed, as images are often ambiguous and it is hard to disentangle the effects of on-screen imagery and concreteness of meaning.

Future research should look into the effects of imagery while controlling for these other word-related variables to better understand the contribution of visual support to word learning, and further explore the effect of ITOS at other ages and proficiency levels. Finally, considering that outside classroom settings learners generally watch TV series without any sort of instruction – a variable that may have overpowered the effect of imagery – further research on the effects of imagery specifically on incidental word learning alone could provide results that are more ecologically valid. Since the objective is to promote extensive viewing at home, exploring the benefits of imagery in this context would be a valuable contribution to the field.

Funding

Research funded by Ministerio de Ciencia e Innovación (PID2'19-110594GB-100) to Carmen Muñoz and by Ministerio de Universidades (Margarita Salas Grant) to Geògia Pujadas.

References

Avello, D. & Muñoz, C. (forthcoming). The use of captioned-video viewing to support the development of L2 reading skills. *Proceedings of the fifth ELTRIA Conference*, Barcelona.

Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, *50*(3), 507–31.

Bianchi, F., & Ciabattoni, T. (2008). Captions and subtitles in EFL learning: an investigative study in a comprehensive computer environment. In Baldry, A., Pavesi, M., & Taylor Torsello, C. (Eds.) *From didactas to ecolingua: An ongoing research project on translation and corpus linguistics* (pp. 69–90). Trieste: Edizioni Università di Trieste. http://hdl.handle.net/10077/2848

Boers, F. (2018). Intentional versus incidental learning. *The TESOL Encyclopedia of English Language Teaching*, 1–6.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a foreign language*, *20*(2), 136–163.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904–911.

Chai, J., & Erlam, R. (2008). The effect and the influence of the use of video and captions on second language learning. *New Zealand Studies in Applied Linguistics*, *14*(2), 25.

Charles, T., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bimodal input. In Gambier, Y., Caimi, A., & Mariotti, C. (Eds.) *Subtitles and language learning: Principles, strategies and practical experiences* (pp. 173–198). Bern: Peter Lang.

Chen, Y. R., Liu, Y. T., & Todd, A. G. (2018). Transient but effective? Captioning and adolescent EFL learners' spoken vocabulary acquisition. *English Teaching & Learning*, *42*, 25–56.

Danan, M. (2004). Captioning and subtitling: undervalued language learning strategies. *Meta: Journal Des Traducteurs*, *49*(1), 67–77.

Deconinck, J., Boers, F., & Eyckmans, J. (2017). 'Does the form of this word fit its meaning?' The effect of learner-generated mapping elaborations on L2 word recall. *Language teaching research*, *21*(1), 31–53.

Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*, *100*(S1), 19–47.

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

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly* *24*(2), 174–187. https://www.jstor.org/stable/747863.

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, *42*(3), 499–523.

Gullberg, M., De Bot, K., & Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gestures 8*(2), 149–179.

Hasan, A. (2000). Learners' perceptions of listening comprehension problems. *Language, Culture and Curriculum*, *13*(2), 137–53.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language 11*(2), 207–23.

Hulstijn, J. H. (2003). Incidental and Intentional Learning. In Doughty, C. & Long, M. H. (Eds.), *The handbook of second language acquisition* (Vol. *19*, pp. 349–381). MA: Blackwell.

Hulstijn, J. H. (2013). Incidental learning in second language acquisition. In Chapelle, C. A. (Ed.) *The encyclopedia of applied linguistics* (Vol. 5, pp. 2632–40). Chichester: Wiley-Blackwell.

Jelani, N. A., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL-International Journal of Applied Linguistics*, *169*(1), 169–190.

Jones, L. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language learning & technology 8*(3), 122–143. ISSN 1094-3501

Jones, L., & Plass, J. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *The Modern Language Journal*, *86*(4), 546–61.

Khan, N., Kasdar, J., Melvin, M., Blomquist, R., Huang, E., & McEwen, J. (2015). Fresh off the Boat [TV series] Los Angeles, CA: ABC.

Krashen, S. D. (2003). *Explorations in language acquisition and use*. Portsmouth, NH: Heinemann. ISBN: 0-325-00554-0

Lee, S. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning*, *57*(1), 87–118.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129.

Maleki, A., & Safaee Rad, M. (2011). The effect of visual and textual accompaniments to verbal stimuli on the listening comprehension test performance of Iranian high and low proficient EFL learners. *Theory and Practice in Language Studies*, *1*(1), 28–36.

Markham, P. (1999). Captioned videotapes and second-language listening word recognition. *Foreign Language Annals*, *32*(3), 321–28.

Mayer, R. (2014). *The Cambridge handbook of multimedia learning*. Cambridge: Cambridge University Press.

Meara, P., & Milton, J. (2003). *X_Lex: the Swansea levels test*. Berkshire: Express Publishing.

Miralpeix, I. (2012, March). *X_Lex and Y_Lex: A validation study*. Paper presented at the 22nd Lexical Studies Conference, Newtown, UK.

Montero Perez, M. (2019). Pre-learning vocabulary before viewing captioned video: an eye-tracking study. *The Language Learning Journal*, *47*(4), 460–478.

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: the effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*(1–2), 1–26.

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), 118–41.

Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research 39*(*1*), 13–38.

Naghizadeh, M., & Darabi, T. (2015). The impact of bimodal, Persian and no-subtitle movies on Iranian EFL learners' L2 vocabulary learning. *Journal of Applied Linguistics and Language Research*, *2*(2), 66–79.

Nation, P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language 27*(1): 136–145.

Nation, P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from: https://www.victoria.ac.nz/lals/about/staff/paul-nation

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.

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

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551–77.

Peters, E., Heynen, E., & Puimège, E. (2016). Learning Vocabulary through Audiovisual Input: The Differential Effect of L1 Subtitles and Captions. *System 63*, 134–48.

Pujadas, G. (2019). Language learning through extensive TV viewing: A study with adolescent EFL learners (Unpublished doctoral dissertation). University of Barcelona.

Pujadas, G., & Muñoz, C. (2018). *What words do we learn better through TV series? The effects of age, proficiency and type of instruction*. Paper presented at the Second Language Research Forum, Montreal, Canada.

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

Pujadas, G., & Muñoz, C. (2020). Examining Adolescent EFL leaners' TV viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, *42*(3), 551–575.

Robinson, P., Mackey, A., Gass, S. M., & Schmidt, R. (2012). Attention and awareness in second language acquisition. In Gass, S. M., & Mackey, A. (Eds.) *The Routledge Handbook of Second Language Acquisition* (pp. 247–267).

Rodgers, M. P. H. (2013). English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions (Unpublished doctoral dissertation). Victoria University of Wellington.

Rodgers, M. P. H. (2016). Extensive listening and viewing: The benefits of audiobooks and television. *The European Journal of Applied Linguistics and TEFL*, *5*(2), 43–57.

Rodgers, M. P. H. (2018). The images in television programs and the potential for learning unknown words: The relationship between on-screen imagery and vocabulary. *ITL – International Journal of Applied Linguistics 169*(1), 191–211.

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

Rodgers, M. P. H., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *ITL – International Journal of Applied Linguistics*, *171*(2), 191–220.

Suárez, M. D. M., Gilabert, R., & Moskvina, N. (2021). The mediating role of vocabulary size, working memory, attention and inhibition in early vocabulary learning under different TV genres: An exploratory study. *TESOL Journal*, *12*(4), e637.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning and Technology*, *14*(2), 50–73.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The Effects of Repetition on Incidental Vocabulary Learning: A Meta-Analysis of Correlational Studies. *Language Learning*, *69*(3), 559–99.

van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*(3), 609–624.

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

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching 43*(1), 1–37.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

Webb, S. (2010a). Pre-learning low-frequency vocabulary in second language television programmes. *Language Teaching Research*, *14*(4), 501–15.

Webb, S. (2010b). Using Glossaries to Increase the Lexical Coverage of Television Programs. *Reading in a Foreign Language*, *22*(1), 201–221.

Webb, S., & Chang, A. C. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Rodgers, M. (2009a). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–66.

Webb, S., & Rodgers, M. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–27.

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning and Technology*, *14*(1), 65–86.

Yang, H. Y. (2014). The effects of advance organizers and subtitles on EFL learners' listening comprehension skills. *Calico Journal*, *31*(3), 345–373.

Appendix

Table A. Target nouns descriptives

Target item	**Session**	**Recency**	**Spacing**	**Corpus frequency**	**Concret.**	**Frequency**	**CoO**	**ITOS (seconds)**
janitor	1	2	2	5.73	4.68	3	1	4
jukebox	2	2	1	2.27	4.93	3	1	26
napkin	2	2	1	3.61	4.93	6	1	19
crouton	2	2	1	0.25	4.9	8	1	26
nightmare	3	4	2	22.39	2.96	3	0	–
rib	3	3	1	5.9	4.9	6	1	17
mall	5	8	2	18.9	4.83	10	0	–
real estate	5	5	1	0.02	4.25	4	0	–
AC	5	5	1	2.16	4.21	10	1	7
buckle	6	6	1	5.04	4.92	2	1	8
carpool	6	6	1	0.71	3.9	6	1	128
knockoff	6	6	1	0.45	2.85	3	0	–
billboard	6	6	1	1.35	4.83	10	1	29
franchise	6	6	1	2.37	3.72	6	0	–
hedgehog	7	7	1	0.29	4.93	2	1	11
ride	7	7	2	135.37	3.75	8	1	36
principal	7	7	1	13.75	4.79	2	1	91
realtor	8	8	2	1.8	4.61	7	0	–
shield	8	8	1	8.2	4.66	4	0	–
hairdryer	8	8	1	0.22	4.97	2	1	108
Mean	–	–	–	**11.54**	**4.43**	**5.25**	–	**39.2**

Note: Session = target episode where TIs were pre-taught; Recency = last episode where TIs were encountered; Spacing = massed encounters (1) vs. spaced encounters (2); Corpus frequency = frequency per million according to the SUBTLEX-US corpus; Concret. = Concreteness ratings by Brysbaert, Warriner and Kuperman (2014); Internal frequency = Mean frequency of encounters with TIs within the term; CoO = Co-occurrence of TIs and its image; ITOS = TIs image time on screen (in seconds).

Address for correspondence

Geòrgia Pujadas
University of Barcelona
Gran Via de les Corts Catalanes, 585
08003 Barcelona
Spain

georgia.pujadas@ub.edu
https://orcid.org/0000-0002-0290-1158

Co-author information

Carmen Muñoz
University of Barcelona

munoz@ub.edu
https://orcid.org/0000-0002-7001-4155

Publication history

Date received: 30 September 2022
Date accepted: 25 January 2023
Published online: 6 March 2023