+ MODEL

Available online at www.sciencedirect.com

ScienceDirect

System 41 (2013) 609–624

SYSTEM

www.elsevier.com/locate/system

Incidental vocabulary acquisition through L2 listening: A dimensions approach

Hilde van Zeeland [a,*], Norbert Schmitt [b]

[a] Room A93 Trent, University Park, Nottingham NG7 2RD, United Kingdom
[b] Room A61 Trent, University Park, Nottingham NG7 2RD, United Kingdom

Received 3 February 2012; revised 3 July 2013; accepted 18 July 2013

Abstract

This study investigated L2 learners' acquisition of three vocabulary knowledge dimensions through listening: form recognition, grammar recognition, and meaning recall. Whereas previous listening studies used only meaning-based vocabulary tests, which revealed very little vocabulary learning, the results of this study shows that learners start developing knowledge of a word (i.e. form and grammar recognition) long before they master the form-meaning link. Knowledge of the three dimensions immediately after listening was form > grammar > meaning, with the former two being more sensitive to attrition than the last. The effect of frequency of occurrence (3, 7, 11, or 15 exposures) on acquisition also differed between the three dimensions, but this effect was not strong overall. The acquisition of word meaning seemed particularly unaffected by frequency, a finding reminiscent of research on incidental learning from reading. For listening to be a valuable source for vocabulary learning, it appears that considerably more than 15 exposures are needed.

Keywords: Incidental vocabulary acquisition; L2 vocabulary knowledge; L2 listening; L2 reading

1. Introduction

Incidental learning occurs when learners acquire new aspects of their L2 without being focused on doing so. Researchers have since long been interested in the uptake of new vocabulary knowledge through incidental learning. This is an important question, for it gives an indication of how much input is required to reach certain degrees of vocabulary uptake.

The vast majority of such research has been carried out in the area of reading (e.g. Dupuy and Krashen, 1993; Hulstijn, 1992; Pitts et al., 1989). Although results vary, the learning gains found are generally small. The modest size of the gains may be due partly to the testing instruments used. Most studies used tests that require high levels of knowledge, mostly recognition and recall of meaning. However, vocabulary knowledge is a complex construct with multiple knowledge dimensions, such as form, grammatical characteristics, and collocations (Nation, 2001). In the case of those reading studies which only used meaning-based tests, it is likely that participants did gain knowledge of

* Corresponding author. Tel.: +44 7816757219.
E-mail addresses: aexhv@nottingham.ac.uk (H. van Zeeland), norbert.schmitt@nottingham.ac.uk (N. Schmitt).

0346-251X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.system.2013.07.012

some of these dimensions, but that the tests were insufficiently sensitive to reveal this. Researchers have therefore started to explore vocabulary learning from reading through the use of multiple vocabulary tests to provide insight into the acquisition of dimensions besides meaning. This vocabulary dimensions framework has revealed that more learning results from reading than was previously thought. For example, acquisition of the form-meaning link may require more exposures than that of spelling (Webb, 2007).

L2 listening has received relatively little research attention (Vandergrift, 2007), and this is also the case in the area of incidental vocabulary acquisition: considerably fewer studies have been carried out in the context of listening than reading. This is surprising, as spoken language is the primary medium of learning in many language classrooms, especially in the communicative language teaching context. Some early vocabulary studies have included listening, but they explored how auditory stimuli can reinforce acquisition from reading (e.g. Kelly, 1992), rather than acquisition from listening only. More recently, a few vocabulary studies have measured learning from listening directly. Their findings suggest that listening leads to even smaller gains than reading does (Brown et al., 2008; Vidal, 2011). Yet, similar to traditional reading research, these studies have used meaning-based tests only. This suggests that more learning may have occurred than has been revealed. The current study uses more sensitive vocabulary tests in order to find how much learning really occurs from listening, and how this compares to reading. It appears to be the first study to employ a vocabulary dimensions framework to investigate incidental vocabulary acquisition from L2 listening.

2. Literature review

2.1. Incidental vocabulary acquisition from reading

Several studies have used the vocabulary dimensions framework to explore the incidental acquisition of vocabulary from reading. Waring and Takaki (2003) explored learning from graded readers and found small gains overall: immediately after listening, learners recognised and recalled the meaning of 10.6 (42.4%) and 4.6 (18.4%) of the 26 target words, while they managed to recognise the form of 15.3 (61.2%). Three months later most of this knowledge was forgotten: learners generally remembered the meaning of only one word. Regarding the role of frequency, learners needed fewer encounters with a word to learn and retain form recognition (spelling) and prompted meaning recognition than unprompted meaning recognition. If learners had encountered a word 4–5 times, the chances of word form recognition were 24%, meaning recognition 16% and meaning recall 0%. With 15–18 occurrences, these chances of form and meaning recognition about doubled, while the chance of meaning recall increased to 6%. The authors suggest that more than 20 encounters may be required to gain full knowledge of a new word through reading.

Pigada and Schmitt (2006) explored one participant's acquisition of word meaning, spelling and grammatical characteristics from reading graded readers. Of all 133 target words, knowledge of 97 (65.4%) was improved in one or more dimensions. As also found by Waring and Takaki (2003), the acquisition of spelling required very few exposures. The acquisition of meaning did not seem closely related to frequency, with the learning gains not being predictable from 2 to 3, 4–5, 6–10 and 10+ encounters. Only at the extremes the frequency effect was clearly visible: with 1 occurrence, only 3.4% of the words' meaning could be recalled, and with 20 or more occurrences this was 60%. Developing knowledge of grammatical information seemed more closely related to frequency, but the acquisition of articles required fewer encounters than that of prepositions. With 4–5 encounters, knowledge of the spelling of 39%, grammar of 27%, and meaning of 27% of the words improved. With 20 or more repetitions this was 67%, 80%, and 60%. Pigada and Schmitt suggest that words need to be encountered 20 or more times for all three knowledge dimensions to be learned.

Pellicer-Sánchez and Schmitt (2010) explored the acquisition of word form recognition (spelling), word class recall, and meaning recognition and recall through reading an authentic novel. Of all 34 target words, knowledge of 9.4 (28%) was enhanced somehow. Their results show a substantial increase in learning at 10 or more occurrences for all knowledge types. The authors collapse the data into two frequency bands: 1–8 and 10+ occurrences. Of the words met 1–8 times, spelling was learned in 28% of the cases, word class in 12%, meaning recognition in 36%, and meaning recall in 7%. Of words read 10 or more times, this was 76% for word form recognition, 63% for word class, 84% for meaning recognition, and 55% for meaning recall. This study thus confirms earlier reports that the acquisition of spelling requires fewer occurrences than that of other knowledge types, but contradicts Waring and Takaki's (2003) in finding that meaning recognition requires fewer exposures than form recognition does.

Results of these studies reveal the complexity of incidental vocabulary knowledge acquisition. Learners start to develop knowledge of a word from the very first exposure to it, long before they reach mastery of the word's form-meaning link.

2.2. Incidental vocabulary acquisition from listening: and how does it compare to reading?

Vidal explored incidental vocabulary acquisition from L2 listening (2003), and compared gains from listening and reading (2011). These studies analysed the effect of a large number of variables (e.g. frequency of occurrence, predictability from word form and parts) on learning. Knowledge gains of 36 target words were measured with a modified version of the Vocabulary Knowledge Scale, on which learners could effectively score 0 to 5. Out of the maximum score of 180, readers scored 40.85 (22.7%) on the immediate post-test and 19.14 (10.6%) on the one-month delayed test. Listeners scored 27.86 (15.5%) immediately after listening and 14.05 (7.8%) one month later.[1] The primary finding is that both reading and listening lead to vocabulary knowledge gains, with gains from reading being significantly higher than from listening. However, gains from listening were retained better than those from reading. An effect of frequency occurrence (1, 2, 3, 4, 5, or 6 occurrences) was found in both modes but this was considerably stronger in reading. More repetitions were needed in listening (5–6) than in reading (2–3) for it to have a positive effect on learning.

Brown et al. (2008) compared vocabulary learning from reading and listening (as well as from reading-while-listening). Their study involved two different vocabulary tests both measuring knowledge of meaning: a multiple choice test (meaning recognition) and a translation test (meaning recall). Immediately after reading, learners scored 44.8% on the MC and 14.6% on the recall test. After listening learners scored much lower at 29.3% and 2%. For all 28 target words, this equals recognition and recall of the meanings of 12.54 and 4.1 words after reading, and 8.2 and 0.56 words after listening. Thus, scores on the MC test were considerably higher than those on the translation test, indicating once more that recognition is acquired before recall (though correct MC test scores can be the result of guessing). The most notable difference between reading and listening was that the former lead to significantly more vocabulary learning. Another difference was that in the context of reading, but not listening, the translation scores tended to drop over time, reminiscent of Vidal's (2011) finding that knowledge gained from listening was retained better (although this may be partially due to the fact that the very low listening translation scores did not have the possibility of dropping very far). A third difference concerned the frequency effect: this was found to exist in both input modes, but was smaller in the listening mode. Even if a word was met 15–20 times, learners could provide a translation of it in only 2.7% (.19/7) of the cases immediately after listening.

These studies both indicate smaller learning gains from listening than from reading. Why might this be the case? Possible reasons are related to the nature of spoken versus written language: spoken language requires fast processing, meaning that listeners simply have less time than readers to focus linguistic information. It is also continuous (i.e. no obvious boundaries between words), likely making it more difficult to notice new words in the input. In addition, many L2 listeners experience problems with speech segmentation, which often leads them to miss subsequent or forget previous parts of the input (e.g. Goh, 2000). Such problems can negatively affect comprehension and hinder acquisition.

Although both Vidal (2011) and Brown et al. (2008) provided a thorough comparative analysis of learning through spoken and written input, the primary concern we are left with is the actual uptake from listening. This is clearly notable in the fact that the two tests used by Brown et al. led to very different scores; this indicates that incidental vocabulary learning is more complex than could be revealed by these tests only. As the authors recognise, 'it suggests that a considerable amount of vocabulary knowledge was gained from the exposure, but was not assessed. Such knowledge might include the noticing of lexical phrases, collocational and colligational patterns, new nuances of meanings, improved lexical access speed, and so on. It is probably here that the true benefit of reading and listening extensively occurs' (p. 158).

This lack of sensitive vocabulary knowledge assessment in listening studies is surprising. As learning gains from listening have found to be small, even significantly smaller than those from reading, the dimensions approach should serve particularly well in revealing the smallest increments in learning.

[1] These are gain scores, i.e. pre-test scores subtracted from the post-test scores (Vidal, 2011). Gains after listening were similar to those found in the listening-only study (Vidal, 2003).

3. The study

3.1. Aims

This study aims to provide a more complete picture of vocabulary knowledge gains from listening through the use of multiple tests, both immediately and two weeks after listening. The following research questions are explored:

1. To what degree does L2 listening lead to the acquisition of three vocabulary knowledge dimensions: form recognition, grammar recognition and meaning recall?
2. To what degree is newly acquired knowledge of these three dimensions retained until two weeks after listening?
3. To what degree is the acquisition of these three vocabulary knowledge dimensions affected by words' frequency of occurrence in the input?

3.2. Methodology

3.2.1. Participants

A total of 30 participants took part in this study. All were postgraduate students at a British university. The university minimum level of acceptance is TOEFL iBT 100 or IELTS 6.0. As all participants must have met this minimum, it can be accepted with confidence that they were of high-intermediate or advanced level. Of the 30 participants, 19 were PhD students and 11 Master students, and 13 were male and 17 female. The pool included participants from 17 different L1s: Chinese (5), Dutch (3), Thai (3), Arabic (2), Hausa (2), Italian (2), Taiwanese (2), Vietnamese (2), Bengali (1), French (1), Hindi (1), Japanese (1), Kurdish (1), Romanian (1), Russian (1), Spanish (1), and Swedish (1). The average age was 27.6.

The participant pool was divided into two groups: 20 participants received the listening input followed by an immediate post-test, and 10 received the input followed by a two-week delayed post-test. This was done to avoid any risk of the so-called 'testing phenomenon', which refers to exposure in the immediate post-test positively affecting participants' scores on the delayed post-test (Glover, 1989). In order to minimise any possible differences between the two groups they had approximately the same ratio of Master and PhD students (13 and 7 in the immediate and 6 and 4 in the delayed post-test group), as well as a similar mean age (27.7 in the immediate and 27.5 in the delayed post-test group) and male-female ratio (9:11 in the immediate and 4:6 in the delayed post-test group). Both groups represented a wide variety of L1s.

3.2.2. Listening passages

Four listening passages were used (Table 1). The passages were of different genres, providing a more varied input context than other incidental learning studies such as those by Vidal (2003), who used only academic lectures, and Brown et al. (2008), who used only recordings of graded readers. All four passages shared a common theme (crime), so that they could contain the same target items. As they were taken from the internet, they represented real spoken English. However, they were slightly adapted for the purpose of the study: the target items needed to be inserted according to different frequency of occurrence bands, and the exposures needed to be spread out over the four passages.

Table 1
Overview of the four listening passages.

Text genre (and source)	Topic	Word count	Nonword count	Lexical coverage	Recording time
Informative, includes anecdote (TV talk show)	Explains a new justice scheme that enables victims of crime to meet their offenders.	1080	52	95.2%	5.15 min
Anecdote (TV interview)	A man explains how his house got broken into.	1156	57	95.1%	5.14 min
Informative (TV interview)	A teacher of crime scene investigation talks about his job.	883	43	95.1%	4.35 min
Life story (Informal lecture)	A former drug dealer tells his life story.	1322	64	95.2%	6.05 min

Based on the finding that 95% lexical coverage is sufficient for adequate listening comprehension of narrative text (van Zeeland and Schmitt, 2012) it was ensured that this coverage level was reached by all participants. This was ensured in two ways: 1) no more than 5% of the words in each passage were replaced by nonwords; and 2) by means of a frequency analysis with the Vocabulary Profile on the *Compleat Lexical Tutor* website (Cobb, n.d.) it was confirmed that almost all other words in the text were within the 2000 frequency band.[2] As the participants in this study were all proficient learners of English who undoubtedly mastered the 2000 level, it could be safely assumed that they reached 95% lexical coverage.

All passages were recorded by the same male native British English speaker. They were told as naturally as possible, in a speaking style which suited the text genre and was similar to that used in the original passages.

3.2.3. Target items

In total, 24 target items were used: six items in each frequency of occurrence band (Table 2). As one of the knowledge dimensions explored was grammar, the target items needed to be of different parts of speech. Each frequency of occurrence band contained two nouns, two verbs, and two adjectives. Adverbs were not used because their form (ending in *-ly*) would give away their grammatical function in a recognition test. Each frequency band furthermore contained three concrete and three abstract words.[3]

All items were finally replaced by nonwords (Table 2). These nonwords were created with the online *ARC Nonword Database* (Rastle et al., 2002). All conformed to the phonologic and orthographic constraints of English, were one syllable, and had four to six phonological neighbours.

3.2.4. Vocabulary tests

Participants' knowledge of the three vocabulary dimensions was measured with the tests described below. An example item from each of these three tests can be found in Appendix A.

1. Form recognition

Knowledge of word form was measured with a multiple choice recognition test. This receptive test format was considered appropriate because it measures the type of knowledge needed for listening: to understand a given word, learners need to recognise its spoken form and distinguish it from other word forms.

Incidental learning studies which included a multiple choice form recognition test have generally used distractors that are quite similar in form to the key (e.g. Chen and Truscott, 2010; Péllicer-Sanchez and Schmitt, 2010; Webb, 2005, 2007). Choosing the correct answer of course requires a higher level of knowledge when the distractors are similar to the key than when the distractors are dissimilar to the key (Bruton, 2007). The first plan was to include two form recognition tests: one with dissimilar distractors (revealing small gains in knowledge), and one with similar distractors (revealing greater gains). However, as learners scored similarly on both tests, only one form recognition test was included, i.e. the one with dissimilar distractors.

As participants had been exposed to the aural form of the words, the form recognition test also presented the items to the participants in spoken form. For each target item, participants heard four different nonword options on a recording, each preceded by A, B, C, or D. One of these options was a target item. Participants were given a piece of paper which only showed A, B, C and D, so they could not see the written form of the items. This ensured that only spoken word form recognition was measured. They were asked to tick either A, B, C or D, or to choose the *I don't remember any of these* option if none of the forms sounded familiar to them. The piloting showed that 4 seconds after each item was enough for participants to select one of the four options.

2. Grammar recognition

Participants' knowledge of the items' part of speech was also measured with a multiple choice recognition test. This format was considered appropriate because part of speech is a closed system (i.e. there are only a few

[2] For a small number of 3000 level words, no high-frequency word substitute could be found: *ambulance* (occurs once in Story 1), *offenders* (occurs once in Story 1), *doll* (occurs twice in Story 3), and *studio* (occurs once in Story 3). The lexical coverage figures in Table 2.7 do not count these lower-frequency items as unknown, but even if they are, the coverage of Stories 1 and 3 still reaches 95%.

[3] Concreteness/Abstractness was determined by a survey carried out with eight native speakers of English.

Table 2
Overview of the 24 target items.

			Concreteness (1–10)[a]	Original word	Substitute word
24 target items	6 items with 3 occurrences (total of 18 occurrences)	2 nouns	9.00	Book	*vinse*
			3.25	Community	*grike*
		2 verbs	8.38	To touch	*to lulk*
			3.88	To understand	*to cluss*
		2 adjectives	7.25	Sunny	*droil*
			4.00	Difficult	*quirt*
	6 items with 7 occurrences (total of 42 occurrences)	2 nouns	8.00	Burglar	*brouth*
			1.88	Idea	*nunce*
		2 verbs	8.25	To sleep	*to belve*
			3.50	To try	*to glabe*
		2 adjectives	7.13	Loud	*krung*
			2.88	Important	*tulse*
	6 items with 11 occurrences (total of 66 occurrences)	2 nouns	8.38	Blood	*sulp*
			4.00	Life	*troice*
		2 verbs	8.50	To run	*to treb*
			4.00	To help	*to voadge*
		2 adjectives	7.50	Quick	*drepe*
			2.63	Normal	*yince*
	6 items with 15 occurrences (total of 90 occurrences)	2 nouns	9.13	House	*grath*
			3.00	Time	*zouch*
		2 verbs	7.75	To say	*to nersh*
			2.88	To need	*to strose*
		2 adjectives	7.75	Big	*clauve*
			3.13	Good	*mulse*

[a] Concreteness was rated by eight native speakers of English on a scale of 1 = very abstract and 10 = very concrete. The scores reported are means.

possibilities to choose from, making multiple choice a suitable format), and also because it measures the knowledge necessary for listening: in order to understand a word in context, learners need an understanding of what part of speech it is. A format was used similar to that of Webb (2005, 2007). Participants were presented with each target item on the recording and on paper, and were asked to tick the correct box if they knew what part of speech the item was: noun, verb or adjective. They were also given an *I don't know* option. Short sentences were provided on paper to illustrate how the noun, verb or adjective form of the item could be used in context. These short sentences were only meant as support and reading them was optional. Six seconds of silence followed each item on the recording, which gave participants enough time to read the example sentences and tick one of the four boxes. If participants needed more time, they gave a sign to the researcher and the recording was paused.

3. Meaning recall

Knowledge of meaning was measured with a recall test. Although this format is different from the form and grammar tests (which measure recognition), it was considered most appropriate because it measures the type of knowledge needed for listening. That is, for a learner to understand a word's meaning in listening, s/he must be able to recall (rather than recognise) that meaning once the form is recognised.

Participants were presented with all 24 target items on paper and asked to write down anything they knew about their meaning. This could be a translation into English, a synonym, an explanation, or anything else that demonstrated their knowledge. Although this written test format was incongruent with the spoken input, participants had been presented with both the spoken and written form of the target items in the grammar test taken before this meaning test. By now, they were expected to be able to link the spoken form to the written form they saw.

As the first few pilots showed that presenting the target items in isolation was too difficult (as also found by Donkaewbua, 2007), it was decided to present the target items in sentence contexts. The sentences were the same as those used in the grammar test. To make sure the sentences did not give away the items' meaning, native speakers

were asked to read the sentences and to guess the meaning of the nonwords. Where correct guessing was revealed, changes were made until the items were no longer guessable in their contexts (these changes were also made in the grammar test).

On all three tests answers were scored as either correct (one point) or incorrect (0 points).

3.2.5. Procedure

Participants listened to the passages and completed the tests in a one-to-one session with the first author. Before listening, they were told that the passages contained some unknown words, but that they should try to comprehend them as well as they could. After listening to each passage participants were asked two questions: a general question about the passage content, and a more specific question which focused on more detailed information. Correct answers to these questions were taken as an indication that they had paid attention to the input.

Participants completed the three vocabulary tests immediately ($N = 20$) or two weeks ($N = 10$) after listening. The order of the tests was: 1) form recognition, 2) grammatical knowledge, and 3) meaning recall. This order was chosen to ensure that participants were not given any clues to the correct answers in the following tests. Instructions were given before each test. In the form and grammar tests, participants were asked to carefully consider the options before selecting one. In order to minimise guessing, participants were asked to choose the 'I don't know/remember' option when they did not know the answer. The first author guided the subjects through the completion of each test. After the first few test items, subjects were asked how confident they felt selecting one of the four options. When a subject expressed his/her difficulty with recognising any of the options as correct, but still appeared to be providing random responses (i.e. guesses), the author encouraged him/her to choose the 'I don't know/remember' option instead. This further ensured that minimal blind guessing occurred. The whole experiment lasted approximately 50 min.

4. Results

4.1. Absolute learning: how much incidental learning from listening?

Table 3 provides an overview of the number of participants who answered each target item correctly on the immediate and delayed form, grammar and meaning test. Out of the total of all target items and all three knowledge dimensions, on the immediate post-test learning was found in 29.2% of the cases, which is an average of 7.05 out of the 24 target items. On the delayed post-test learning was found in 19% of the cases, or in 4.56 of the target items. This shows that considerable learning has taken place overall. Yet it should be kept in mind that these percentages do not reflect full knowledge of the words, but rather reflect that one of more aspects of knowledge had started to develop.

In fact, there are notable differences between the learning gains of the three different knowledge dimensions (see also Fig. 1). Immediately after listening, participants demonstrated receptive knowledge of form in 45.8% of the cases (11 items), and of grammar in 33.7% of the cases (8.1 items). Participants managed to recall the meaning of only 8.5%, or of 2.05 words. Results were notably lower two weeks later: 25% (6) for form, 24.6% (5.9) for grammar, and as little as 7.5% (1.8) for meaning recall. Durable learning from listening seems to be modest, especially in the case of meaning recall. Even if we look at the very best case of learning, which was the item *sulp* (*blood*), scores are low: two weeks after listening, only four of the ten participants (40%) remembered its meaning (Table 3).

4.2. Acquisition of the three vocabulary knowledge dimensions

Participants scored highest on the form recognition test, followed by the grammar recognition test, and considerably lower on the meaning recall test (Table 3, Fig. 1). Two Friedman tests were carried out to compare scores on the form, grammar and meaning tests for both the immediate and post-test results. This showed the difference in scores between the three test was significant on both the immediate (χ^2 (2, $n = 20$) 31.564, $p < .001$) and delayed test (χ^2 (2, $n = 10$, 16.270, $p < .001$). Wilcoxon signed rank tests were used to compare the relative performance on the three dimensions. This revealed that participants' immediate post-test scores were significantly higher on the form test than on both the grammar test ($z = -2.864$, $p < .01$) and the meaning test ($z = -3.925$, $p < .001$). Scores on the meaning test were significantly lower than those on the grammar test (z −3.834,

Table 3
The number of times each target item was answered correctly on the three tests, categorised according to the items' concreteness-abstractness, part of speech, and frequency of occurrence.

Target word	Nonword	C/A	POS	FOO	Form		Grammar		Meaning	
					Imm Max20	Del Max10	Imm Max20	Del Max10	Imm Max20	Del Max10
Book	*vinse*	C	N	3	11	1	8	2	2	2
Community	*grike*	A	N	3	3		2	1		
To touch	*to lulk*	C	V	3	1		2	2		
To understand	*to cluss*	A	V	3	3		1	1	1	2
Sunny	*droil*	C	A	3	10	1	5	4	2	1
Difficult	*quirt*	A	A	3	5	3	2	1		
Burglar	*brouth*	C	N	7	13	4	11	4	2	3
Idea	*nunce*	A	N	7	13	3	10	4	1	
To sleep	*to belve*	C	V	7	8	2	10	3	2	
To try	*to glabe*	A	V	7	3	1	8		1	
Loud	*krung*	C	A	7	10	1	2			
Important	*tulse*	A	A	7	10	1	1	1		
Blood	*sulp*	C	N	11	19	7	17	6	14	4
Life	*troice*	A	N	11	12	2	11	2	1	
To run	*to treb*	C	V	11	8	5	6	6		
To help	*to voadge*	A	V	11	16	7	11	3	4	
Quick	*drepe*	C	A	11	9	7	3	1	1	1
Normal	*yince*	A	A	11	7	1	6	2		
House	*grath*	C	N	15	17	5	15	7	4	3
Time	*zouch*	A	N	15	8	1	5	2	1	2
To say	*to nersh*	C	V	15	14	3	9	2	2	
To need	*to strose*	A	V	15	8	2	8	2		
Big	*clauve*	C	A	15	6	1	8	1	1	
Good	*mulse*	A	A	15	6	2	1	2	2	
Total of correct answers					220	60	162	59	41	18
(Max 480 on immediate, 240 on delayed)					45.8%	25.0%	33.7%	24.6%	8.5%	7.5%
Mean number of correct items (Max 24)					11.00	6.00	8.10	5.90	2.05	1.80

Imm = immediate post-test.
Del = delayed post-test.
C/A = concrete or abstract item.
POS = part of speech (Noun, Verb, Adjective).
FOO = frequency of occurrence.

$p < .001$). On the delayed post-test, scores on the form test were still significantly higher than those on the meaning recall test ($z = -2.810$, $p < .01$) but not different from those on grammar recognition. Scores on the grammar test were still significantly higher than those on the meaning test ($z = -2.823$, $p < .01$). Overall, participants thus showed knowledge of the three dimensions as form > grammar > meaning immediately after listening, and this was form = grammar > meaning two weeks later.

4.3. Retention of the three vocabulary knowledge dimensions

As the two participant groups of the immediate and delayed post-tests were comparable, scores on the delayed post-tests can be interpreted as retention of vocabulary knowledge. Independent samples Mann–Whitney U tests were carried out to compare scores on the immediate and delayed post-tests for each knowledge type. Knowledge of form was significantly lower on the delayed than on the immediate test ($U = 28.5000$, $z = -3.174$, $p < .01$). There was also a significant decrease in knowledge of grammar ($U = 51.5000$, $z = -2.156$, $p < .05$). This suggests there is attrition of both form and grammar over a two-week time span. However, as Table 3 and Fig. 1 show, knowledge retention of these two dimensions seems to differ: whereas on the immediate test participants scored higher on the form than on the grammar test, on the delayed post-test scores of these two tests are quite similar. Although initial learning may be higher for form, it seems that this type of knowledge is also forgotten more easily. Contrary to form and grammar, scores on the meaning test did not differ significantly between the immediate and delayed post-tests. Knowledge of

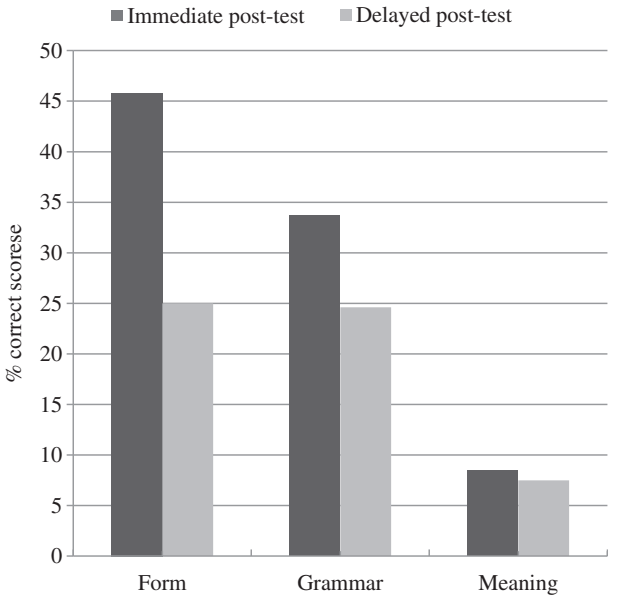

Fig. 1. The percentages of correct scores on the immediate and delayed post-tests of form recognition, grammar recognition, and meaning recall.

meaning is very low on both tests, but any knowledge that learners may have immediately after listening is likely to be retained until two weeks later.

4.4. Frequency of occurrence

Table 4 shows the mean number of items in each of the four frequency bands (3x, 7x, 11x, 15x) answered correctly on the three immediate and delayed post-tests (see also Figs. 2 and 3). On the immediate post-test of form recognition, a Friedman test showed a significant effect of frequency: χ^2 (2, $n = 20$) 21.000, $p < .001$. Post hoc Wilcoxon tests revealed that scores on items in the 3x band were significantly lower than those on items of the other three frequency bands (all $p < .01$). As Table 4 predicted, participants scored significantly better on items of the 11x band than on items of the 7x band ($z = -2.103$, $p < .05$), and there was no significant difference between the knowledge of form for items of the 7x and 15x bands or 11x and 15 bands. This suggests that immediate acquisition of form occurs between 3 and 7 occurrences, yet beyond this point (7–11–15), no real learning gains are found.

A frequency effect was also found on the immediate post-test of grammar (χ^2 (2, $n = 20$) 21.781, $p < .001$). There was an increase in knowledge from the 3x to the three higher bands (all $p < .01$). Additionally, as also found for form, knowledge of grammar was significantly higher for items of the 11x band than the 7x band ($z = -2.288$, $p < .05$), without there being a difference between knowledge of items from the 7x and 15x bands, or the 11x and 15x bands.

Table 4
The mean number of items from the four frequency bands answered correctly on the six post-tests.

Frequency	Form		Grammar		Meaning	
	Imm	Del	Imm	Del	Imm	Del
3 Max = 6	1.65 (27.5%)	0.50 (8.3%)	1.00 (16.6%)	1.10 (18.3%)	0.25 (4.2%)	0.50 (8.3%)
7 Max = 6	2.85 (47.5%)	1.20 (20.0%)	2.10 (35.0%)	1.20 (20.0%)	0.30 (5.0%)	0.30 (5%)
11 Max = 6	3.55 (59.2%)	2.90 (48.3%)	2.70 (45.0%)	2.00 (33.3%)	1.00 (16.6%)	0.50 (8.3%)
15 Max = 6	2.95 (49.2%)	1.40 (23.3%)	2.30 (38.3%)	1.60 (26.7%)	0.50 (8.3%)	0.50 (8.3%)
Total Max = 24	11.00 (45.8%)	6.00 (25.0%)	8.10 (33.7%)	5.9/24 (24.6%)	2.05 (8.5%)	1.80 (7.5%)

Imm = immediate post-test.
Del = delayed post-test.

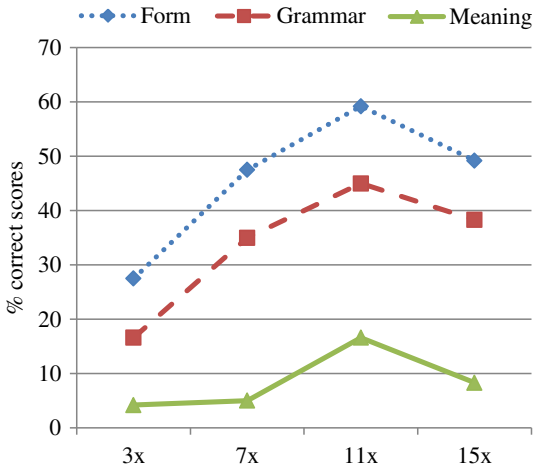

Fig. 2. Percentages of correct scores on the immediate post-tests of form, grammar and meaning by frequency of occurrence.

This shows the same weak frequency effect as found for form: gains between 3 and more occurrences, but no further gains between 7, 11 and 15 occurrences.

Knowledge of meaning on the immediate post-test was also affected by frequency (χ^2 (2, n = 20) 16.875, $p < .01$), yet not as would be expected. Scores on the 11x band were significantly higher than those on the 3x band ($z = 2.982$, $p < .01$), the 7x band ($z = -2.952$, $p < .01$), and the 15x band ($z = -2.500$, $p < .05$), without there being a significant difference between these three bands. In other words, whether listeners heard an item 3, 7, or 15 times did not have an effect on the immediate acquisition of its meaning.

On the delayed post-tests, knowledge of grammar and meaning was completely unaffected by frequency: whether a word had occurred 3, 7, 11 or 15 times in the listening input did not have any lasting effect on listeners' knowledge of these items. Regarding form, the only effect found was the significantly high scores on the 11x band compared to the other three bands ($p < .05$), but the lack of any difference between these three bands indicates that frequency did not play a role in the retained knowledge of form. Retained knowledge of the three dimensions was thus unaffected by the frequency range 3–15.

Thus, beyond the increase from 3 to 7 occurrences (which was not found for meaning), there does not appear to be a frequency effect on the acquisition of any of the three vocabulary knowledge dimensions either immediately or two weeks after listening.

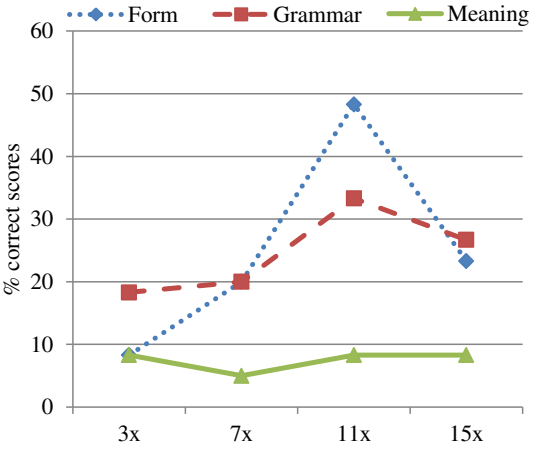

Fig. 3. Percentages of correct scores on the delayed post-tests of form, grammar and meaning by frequency of occurrence.

Table 5
The mean number of concrete and abstract items answered correctly on the six tests (Imm = immediate post-test; Del = delayed post-test).

	Form		Grammar		Meaning	
	Imm	Del	Imm	Del	Imm	Del
Concrete Max = 12	6.30 (52.5%)	3.70 (30.8%)	4.80 (40.0%)	3.80 (31.6%)	1.50 (12.5%)	1.40 (11.7%)
Abstract Max = 12	4.70 (39.1%)	2.30 (19.2%)	3.30 (27.5%)	2.10 (17.5%)	0.55 (4.6%)	0.40 (3.3%)

Imm = immediate post-test.
Del = delayed post-test.

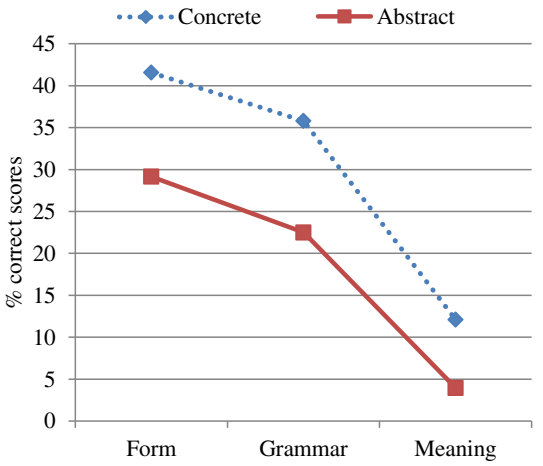

Fig. 4. Percentages of correct scores on concrete and abstract items on the tests of form, grammar and meaning (immediate and delayed post-test scores collapsed).

4.5. Concreteness and part of speech

Although the role of items' concreteness and part of speech is not a main focus of the study, it is of interest to look at their effects.

Table 5 shows the number of correctly answered concrete and abstract items on the immediate and delayed post-tests of form, grammar and meaning. The data of the immediate and delayed post-tests is collapsed for a statistical analysis (Fig. 4). On all three tests, learners demonstrated a better knowledge of concrete than of abstract items. Wilcoxon signed rank tests showed that this advantage of concrete over abstract items was true for all three knowledge dimensions (form, $z = -3.191$, $p < .01$; grammar, $z = -3.419$, $p < .01$; meaning, $z = -3.062$, $p < .01$). This is not surprising for meaning, as the items' concreteness was represented in their meanings, but it is interesting to note that grammar and form recognition were also greater for the concrete items, even though they are not directly related to this characteristic. Concreteness is one factor which facilitates vocabulary acquisition in general (de Groot, 2006), and this advantage seems to apply also to incidental acquisition from listening.

Table 6 gives the number of correctly answered items by part of speech: nouns, verbs and adjectives, and Fig. 5 gives the collapsed data of the immediate and delayed test. Participants generally scored higher on nouns than on the other two parts of speech, as well as higher on verbs than on adjectives. On the form test this difference was significant (χ^2 (2, $n = 30$) 14.080, $p < .01$): participants recognised significantly more nouns than verbs and adjectives (both $p < .001$), but there was no significant difference between their recognition of verbs and adjectives (nouns > verbs = adjectives). Knowledge of grammar also differed between the part of speech types, χ^2 (2, $n = 30$) 18.255, $p < .001$. Participants scored significantly higher on nouns than on verbs ($z = -2.956$, $p < .01$) and adjectives ($z = -4.089$, $p < .001$), and higher on verbs than on adjectives ($z = -2.820$, $p < .01$). Grammatical information therefore seems to be acquired by the profile of nouns > verbs > adjectives. Knowledge of meaning also differed between the three parts of speech (χ^2 (2, $n = 30$) 18.296, $p < .001$), and the

Table 6
The number of nouns, verbs and adjectives answered correctly on the six tests.

	Form		Grammar		Meaning	
	Imm	Del	Imm	Del	Imm	Del
Nouns Max = 8	4.80 (60%)	2.30 (28.8%)	3.95 (49.4%)	2.80 (35.0%)	1.25 (15.6%)	1.40 (17.5%)
Verbs Max = 8	3.05 (38.1%)	2.00 (25.0%)	2.75 (34.4%)	1.90 (23.8%)	0.50 (6.25%)	0.20 (2.5%)
Adjectives Max = 8	3.15 (39.4%)	1.70 (21.3%)	1.40 (17.5%)	1.20 (15.0%)	0.30 (3.8%)	0.20 (2.5%)

Imm = immediate post-test.
Del = delayed post-test.

significant difference was found to exist between nouns and the two other parts of speech, with nouns > verbs ($z = -2.994$, $p < .01$) and nouns > adjectives ($z = -3.753$, $p < .001$). Acquiring the meaning of the three parts of speech thus reflects nouns > verbs = adjectives, as also found for form. This is congruent with most previous lexical acquisition research (e.g. Ellis and Beaton, 1993) showing nouns as the easiest word class to acquire.

Overall, it can be taken from this analysis that item-related variables such as concreteness and part of speech have an effect on their incidental acquisition. Although the effect of these two aspects is not the same for the three knowledge dimensions, in general concrete items were acquired better than abstract ones, and nouns better than verbs and adjectives.

5. Discussion

5.1. Main findings

This study has shown that L2 listening is a source of incidental vocabulary learning. Learners showed an improved knowledge of about 7 (29%) of the nonwords immediately after listening, and retained knowledge of 4 or 5 (19%) of the words until two weeks later. The first research question involved the relative acquisition of the three knowledge dimensions (form, grammar and meaning). Most of this learning took place in the dimension of word form, i.e. learners primarily developed the ability to recognise the word when it was heard again. Least knowledge was acquired of word meaning: after listening for 20 min to a repetition-rich and simplified input, learners could recall the meaning of only two words. Moreover, not much knowledge was retained over the course of two weeks (Research question 2): although learners developed initial knowledge of word form and grammar, much of this knowledge was lost. In contrast, knowledge of meaning seemed harder to

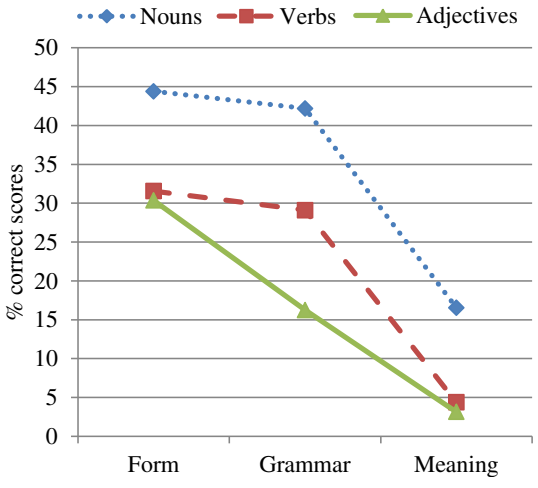

Fig. 5. Percentages of correct scores on nouns, verbs and adjectives on the tests of form, grammar and meaning (immediate and delayed post-test scores collapsed).

acquire, but once learners had linked a meaning to a form, this knowledge was likely to be retained for two weeks.

The third research question considered the role of frequency of occurrence. Results showed there was no clear relationship between learners' acquisition of the target items and their frequency of occurrence in the input. Immediately after listening there was a significant advantage of 7 over 3 occurrences in the acquisition of form and grammar but the results show this to be initial learning: even if a word has just been heard seven times, there is only 47.5% chance a learner will recognise it when heard again. Moreover, this frequency effect was not sustained over a period of two weeks. It appears that immediate, short-term knowledge of form and grammar starts developing with relatively few exposures, yet it needs to be heard considerably more than fifteen times for this knowledge to fully develop and be retained. Knowledge of meaning was not affected by frequency either immediately or two weeks after listening. Many more than 15 occurrences appear to be needed for durable learning of form and grammar, and for any meaningful learning of meaning to occur from listening. As also found in reading research (Horst et al., 1998), frequency is only one of many factors affecting vocabulary acquisition.

So what gains can we expect from L2 listening? Of course this depends on many factors, one being what type of knowledge is aimed for. Mastery of the form-meaning link is a good initial goal, as it makes listening comprehension possible. However, it appears that listening will struggle to provide this learning by itself, even if the input contains 15 repetitions of the target items. This is in line with Brown et al.'s (2008, p. 153) suggestion that considerably more than 20, perhaps 50 or 100, repetitions are needed for listeners to develop the ability to recall a word's meaning. However, building memory of the word form can be considered one step towards developing a form-meaning link, so listening can certainly be considered beneficial in this respect.

5.2. Incidental learning from reading and listening

Although it is a well-established principle that reading leads to more learning than listening does (Section 2.2), this has not been analysed from a dimensions approach; research has merely looked at the acquisition of the form-meaning link through the two modalities. But would the acquisition of form, grammar and meaning develop similarly through spoken and written input? Pellicer-Sánchez and Schmitt's (2010) reading study (Section 2.1) explored three knowledge dimensions also covered in this study: spelling recognition (here *form recognition*); word class recall[4] (here *grammar recognition*); and meaning recall. The frequency bands used in their study also overlap partly with those covered here: 2–4 (here *3x*); 5–8 (here *7x*); and 10–17 (here the *11x and 15x bands collapsed*). Table 7 provides a comparison of the results of these two studies by knowledge dimension and frequency band.[5] Of course this comparison should be interpreted with care, as it is based on the results of two separately designed and executed studies.

One difference between the two modalities is the role of frequency. Table 7 shows that initial learning of form and grammar is greater from listening, but gains from reading accelerate after more exposures (10+) whereas gains from listening do not. This ties in with earlier research: reading studies (e.g. Pigada and Schmitt, 2006; Waring and Takaki, 2003) have found that learning accelerates at 8–10 exposures, while comparative studies (Brown et al., 2008; Vidal, 2011) have found the repetition effect to be smaller in listening than in reading. However, although more learning may occur through reading than listening, these results indicate that a great number of exposures (10+) are needed for reading to have this advantage over listening.

Another notable difference is the acquisition of the form-meaning link. Although the development of this knowledge type lags behind in both modalities, this gap appears to be greater in listening. Here scores on the meaning test are lower in comparison to the other two tests than is found in reading; furthermore, listeners scored lower than readers on the meaning test regardless of frequency. Several reading studies have found that meaning is hardest to acquire incidentally, and that the acquisition of meaning has a particularly weak relationship with frequency of occurrence. It seems that these two findings apply also, perhaps even more strongly, to learning from listening.

[4] This dimension was called 'recall' because participants were asked to write down the correct part of speech rather than circle or tick it (Pellicer-Sánchez and Schmitt, 2010). However, similarly to this study, they were given the target item and asked what the part of speech was, so the actual knowledge type measured was the same.

[5] Pellicer-Sánchez and Schmitt did not use a delayed post-test, so Table 7 gives only the immediate post-test results from this study.

Table 7
A comparison of the percentage of correctly scored items in the tests of form, grammar and meaning in the current study (Listening) and Pellicer-Sánchez and Schmitt's (2010) study (Reading).

Frequency		Form (%)	Grammar (%)	Meaning (%)	All three (%)
3	Listening	27.5	16.6	4.2	16.1
2–4	Reading	16.0	7.0	5.0	9.3
7	Listening	47.5	35.0	5.0	29.2
5–8	Reading	37.0	20.0	11.0	22.7
11–15	Listening	54.2	41.7	12.5	36.1
10–17	Reading	85.0	54.0	48.0	62.3

Interestingly, the general acquisition profile revealed by the two studies is similar: form > grammar > meaning. This was found by earlier reading studies too: Waring and Takaki (2003), Webb (2007), and Chen and Truscott (2010) all found that form recognition was acquired with fewer occurrences than meaning recall, and Webb (2007) also reported that receptive knowledge of grammar fell somewhere between knowledge of form and meaning. This suggests a predictable development of these three knowledge types, regardless of whether written or spoken language is used as a learning source.

5.3. Limitations of the study

There are a number of limitations to this study. The first is that the number of participants in the delayed post-test group was rather small ($N = 10$). Although their results give us some idea of the knowledge retention that can be expected after two weeks, these findings should be taken as indicative rather than conclusive.

Another limitation is that the context in which the target items occur can never be fully controlled. It seems that the meaning of some target items was easier to infer than others, leading to variation in the acquisition rate of individual items. One example was the item *sulp* (*blood*), which learners showed considerably better knowledge of, particularly on the meaning recall test (Table 3). This lead to a high average score on the 11x frequency band, as also found in the analysis (section 4.4). To see if this item interfered with the frequency effect, the analysis was repeated after excluding this success-item, but still no frequency effect was found between 7–11–15 occurrences. In future research, target items could usefully be analysed for degree of concreteness (e.g. with Coh-Matrix: http://cohmetrix.memphis.edu/cohmetrixpr/index.html), so that the items in different variable categories would be more equal in their inferability.

Lastly, all exposures occurred within one listening session. It is not unlikely that listeners would have acquired more knowledge if the exposures had been spaced. This expectation follows the so-called 'distributed practice' principle (Cepeda et al., 2006; Nakata, 2008), which posits that the repetitions of an item have a more positive impact on learning if they are spread out over time rather than occurring within a close time period (the latter case is referred to as 'massed practice'). In this study, all items were presented to the participants within one input session of approximately 20 min, which is a clear case of massed practice. This matter should be taken into account when considering the results.

6. Conclusion

This study appears to have been the first to explore incidental vocabulary acquisition through L2 listening using a dimensions framework, an approach adopted from reading research. Results revealed knowledge gains that would not have been found if only a traditional form-meaning test format would have been used. It appears that some types of knowledge (i.e. word form) are acquired relatively easily through L2 listening, while others (i.e. meaning) are not. The low acquisition rate of word meaning found here, as well as in other incidental learning studies, emphasises once more the importance of combining incidental learning with some sort of explicit focus. It is still unclear where the real potential of listening for vocabulary acquisition lies. Rather than in learning the form-meaning link, it may be particularly useful in learning aspects that concern language usage, such as register characteristics and formulaic sequences. Future research should give learners more input and test a wider variety of knowledge types. This should reveal the strengths and weaknesses of listening, and reveal how listening can be used in and outside the language classroom based on the knowledge types aimed for.

Appendix A

Examples of the three tests used in this study.

1. Form recognition test:

The participant sees on paper:

1 □ A □ B □ C □ D □ I don't remember any of these

At the same time the participant hears on the recording:

'Number one [1sec.] A, rulck [2sec.] B, vince [2sec.], C, preave [2 sec.], D, storce [4sec.]'

2. Grammar recognition test:

The participant sees on paper:

1 sulp Noun □ There was a lot of **sulp**.
Verb □ I **sulped** all evening.
Adjective □ It was a **sulp** film. □ I don't know

At the same time the participant hears on the recording:

'Number one [1sec.] sulp [6 sec.]'

3. Meaning recall test (first six items):

1 I had a **mulse** dinner. _____ □ I don't know
2 She always **nershes** strange things. _____ □ I don't know
3 He was too **drepe** for me. _____ □ I don't know
4 His **nunce** was a good one. _____ □ I don't know
5 My mother is always too **krung**. _____ □ I don't know
6 I left my **vinse** in the living room. _____ □ I don't know

References

Brown, R., Waring, R., Donkaewbua, S., 2008. Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. Reading a Foreign Lang. 20, 136–163.
Bruton, A., 2007. Partial lexical knowledge in tests of incidental vocabulary learning from L2 reading. Can. Mod. Lang. Rev. 64, 163–180.
Cepeda, N.J., Pashler, H., Vul, E., Wixted, J.T., Rohrer, D., 2006. Distributed practice in verbal recall tasks: a review and quantitative analysis. Psychol. Bull. 132, 354–380.
Chen, C., Truscott, J., 2010. The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. Appl. Linguistics 31, 693–713.
Cobb, T. (n.d.). The Compleat Lexical Tutor [Computer Software]. Retrieved from http://www.lextutor.ca/.
de Groot, A.M.B., 2006. Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. Lang. Learn. 56, 463–506.
Donkaewbua, S., 2007. The Effects of Previous Partial Word Knowledge on Vocabulary Learning through Listening. Voctoria University of Wellington (Unpublished PhD thesis).
Dupuy, B., Krashen, S.D., 1993. Incidental vocabulary acquisition in French as a foreign language. Appl. Lang. Learn. 4, 55–63.
Ellis, N.C., Beaton, A., 1993. Psycholinguistic determinants of foreign language vocabulary learning. Lang. Learn. 43, 559–617.
Glover, J.A., 1989. The 'testing' phenomenon: not gone but nearly forgotten. J. Educ. Psychol. 81, 392–399.
Goh, C.C.M., 2000. A cognitive perspective on language learners' listening comprehension problems. System 28, 55–75.

Horst, M., Cobb, T., Meara, P., 1998. Beyond a clockwork orange: acquiring second language vocabulary through reading. Reading Foreign Lang. 11, 207−223.

Hulstijn, J.H., 1992. Retention of inferred and given word meanings: experiments in incidental vocabulary learning. In: Arnaud, P., Bejoint, H. (Eds.), Vocabulary and Applied Linguistics, pp. 113−125.

Kelly, P., 1992. Does the ear assist the eye in the long-term retention of lexis? Int. Rev. Appl. Linguistics Lang. Teach. 30, 137−160.

Nakata, T., 2008. English vocabulary learning with word lists, word cards and computers: implications from cognitive psychology research for optimal spaced learning. ReCALL 20, 3−20.

Nation, I.S.P., 2001. Learning Vocabulary on Another Language. Cambridge University Press, Cambridge.

Pellicer-Sánchez, A., Schmitt, N., 2010. Incidental vocabulary acquisition from an authentic novel: do things fall apart? Reading a Foreign Lang. 22, 31−55.

Pigada, M., Schmitt, N., 2006. Vocabulary acquisition from extensive reading: a case study. Reading a Foreign Lang. 18, 1−28.

Pitts, M., White, H., Krashen, S., 1989. Acquiring second language vocabulary through reading: a replication of the clockwork orange study using second language acquirers. Reading a Foreign Lang. 5, 271−275.

Rastle, K., Harrington, J., Coltheart, M., 2002. 358,534 nonwords: the ARC nonword database. Q. J. Exp. Psychol. 55, 1339−1362.

van Zeeland, H., Schmitt, N., 2012. Lexical coverage and L1 and L2 listening comprehension: the same or different from reading comprehension? Appl. Linguistics http://dx.doi.org/10.1093/applin/ams074.

Vandergrift, L., 2007. Recent developments in second and foreign language listening comprehension research. Lang. Teach. 40, 191−210.

Vidal, K., 2003. Academic listening: a source of vocabulary acquisition? Appl. Linguistics 24, 56−89.

Vidal, K., 2011. A comparison of the effects of reading and listening on incidental vocabulary acquisition. Lang. Learn. 61, 219−258.

Waring, R., Takaki, M., 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? Reading a Foreign Lang. 15, 130−163.

Webb, S., 2005. Receptive and productive vocabulary learning: the effects of reading and writing on word knowledge. Stud. Second Lang. Acquisition 27, 33−52.

Webb, S., 2007. The effects of repetition on vocabulary knowledge. Appl. Linguistics 28, 46−65.