Differences between phonological and orthographic vocabulary knowledge among L1-Spanish learners of English as a foreign language

Martín Aoiz Pinillos
Universidad de Navarra

Received: 2023-03-23 / Accepted: 2023-06-30
DOI: *https://doi.org/10.30827/portalin.vi41.27594*
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: The vocabulary size of language learners might predict their success in a second language because of its strong correlation with better performances in that target language. Although previous research has claimed that aural and written vocabulary are two aspects of vocabulary knowledge that need to be estimated separately, very few studies have examined this issue, particularly from an empirical perspective. This paper presents the possible differences in size between the phonological and orthographic vocabulary among learners of English as a second language. A bilingual vocabulary test was delivered, first orally and then in writing, to 209 language learners in Spain. The refined version of the instrument, showed reliabilities ranging between .79 and .92. Statistical analyses confirm that language learners know fewer words in their aural form than in their written form, regardless of the frequency of the word or the learner's language level. This finding supports the claim that aural and written vocabulary are two separate aspects of knowing a word, and impacts on how vocabulary should be taught in L2 classrooms.
Keywords: L1-Spanish EFL learners, phonological vocabulary, orthographic vocabulary, vocabulary testing, vocabulary teaching.

Diferencias entre el conocimiento fonológico y ortográfico del vocabulario entre estudiantes de inglés como lengua extranjera de L1-español

RESUMEN: El tamaño del vocabulario de los aprendices de una lengua puede predecir su éxito en una segunda lengua por su fuerte correlación con mejores desempeños en esa lengua meta. Aunque las investigaciones anteriores han sugerido que el vocabulario oral y escrito son dos aspectos del conocimiento léxico que tienen que ser estimados separadamente, muy pocos estudios han examinado esta cuestión, especialmente desde una perspectiva empírica. Este artículo presenta las posibles diferencias de tamaño entre el vocabulario fonológico y ortográfico en aprendices de inglés como segunda lengua. Una prueba bilingüe de vocabulario fue administrada, primero oralmente y luego por escrito, a 209 aprendices de lengua en España. La versión refinada del instrumento mostró fiabilidades entre .79 y .92. Los análisis estadísticos confirman que los aprendices de lengua conocen menos palabras en su forma oral que en la escrita, independientemente de la frecuencia de la palabra o del nivel lingüístico del aprendiz. Este hallazgo corrobora la afirmación de que el vocabulario oral y escrito son dos aspectos separados del conocimiento de una palabra, e impacta en la forma en que el vocabulario debería enseñarse en las aulas de segunda lengua.
Palabras clave: aprendices de inglés como lengua extranjera con español como lengua materna, vocabulario fonológico, vocabulario ortográfico, test de vocabulario, enseñanza de vocabulario

1. Introduction

Previous research in second language (L2) acquisition has shown the importance of assessing the learners' vocabulary size, as it clearly correlates to better performances, and serves as a predictor of learning success in the target language. This positive correlation has been observed in several studies across all language skills (Mizumoto & Shimamoto, 2008).

Knowing a word involves being familiar with multiple aspects, including the ability to recognize it either orally or in writing (Nation, 2001). Although phonological and orthographic vocabulary knowledge are different and should be assessed separately (Cheng & Matthews, 2018), very few studies have attempted to estimate the possible differences between the learners' ability to recognize the words in their aural and written form.

Unlike previous studies (for example, Masrai, 2019), this investigation intends to quantify how different the learners' aural and vocabulary sizes are by using the same test items for both test modalities. Furthermore, the present study differs from past similar investigations (for example, Aoiz, 2022) in that it encompasses all levels of overall language proficiency from A2 to C2. Another difference with previous research into the vocabulary size of L2-English learners is that it focuses on L1-Spanish speakers, a language that might be more similar to English than Greek and Arabic (Milton & Hopkins, 2006), Chinese (Cheng & Matthews, 2018), or Japanese (Hamada & Yanagawa, 2023).

This article presents the findings from comparing the learners' ability to link the aural and the written form of a word to its meaning. As the study participants show varied language abilities, the impact of their linguistic proficiency on the ability to recognize words is also analysed. Furthermore, analyses of the possible influence of word frequency on the learners' ability to recognize words are also included.

2. Literature review

2.1. Receptive and productive vocabulary knowledge

Based on the multiple aspects of what knowing a word involves (Nation, 2001), a broad distinction might be made between productive and receptive vocabulary assessment. The former focuses on determining the ability of language users to produce the correct word on a given moment, whereas the assessment of receptive vocabulary focuses on the ability to recognize the form-meaning link (Masrai, 2022).

When distinguishing between receptive and productive vocabulary, researchers assume that using a word in an utterance implies knowing it more deeply than simply recognizing it. In this respect, knowledge about a word usually begins receptively and it develops until the language user is able to use that word in later stages of its acquisition (Read, 2000). Consequently, authors have mentioned the existence of a continuum of receptive to productive word knowledge, and they have empirically evidenced the differences between the L2-English learners' ability to recognize a word and to recall and use it when necessary (Masrai, 2022). A possible explanation for those differences might be that during the early stages of vocabulary acquisition, language learners only know if they have seen or heard that word before, and in fewer cases, they are able to relate that aural or written form to a possible meaning (Cheng & Matthews, 2018).

Although the distinction between receptive and productive vocabulary has been long accepted by researchers, there have been some discrepancies in the way those two aspects of vocabulary knowledge should be assessed (Masrai, 2022). Since 2000, most studies have estimated the vocabulary knowledge of their participants by determining whether they were able to recognize the target items, particularly in their written form (Smith, 2019). Yes/No tests like the X-Lex (Meara & Milton, 2003) and the A-Lex (Milton & Hopkins, 2006) are among the most widely used forms to assess the receptive vocabulary of L2-English learners (for example, Masrai, 2019).

A second type of tests employed in the assessment of receptive vocabulary among L2-English populations uses either the Vocabulary Levels Test (VLT, Schmitt et al., 2001), or the Vocabulary Size Test (VST, Nation & Beglar, 2007). Both tests consider that a person knows a word if they are able to link its form to the right meaning by selecting the best option among a limited number of short descriptive phrases. Initially designed to present the target lexical items in their English written form, in the past years aural and bilingual versions of those tests have been developed in Vietnamese (Nguyen and Nation, 2011) Iranian (Karami, 2012), and Russian (Elgort, 2013). Furthermore, aural versions of the VST, where the target words are presented orally and the possible options to choose from appear in the target language, have appeared in Japanese (McLean et al., 2015), Vietnamese (Ha, 2021) and Chinese (Du et al., 2022).

With respect to measuring the productive vocabulary size of an L2-English learner, i.e., the number of words they are able to retrieve from memory and use when required, researchers have employed the Productive Vocabulary Levels Test (Laufer & Nation, 1999), where test-takers show that they know a word productively by writing the missing final letters in its orthographic form. The test has been used in different contexts in the past years (for example, Abdullah et al., 2013), and it was subjected to a validation process by its own designers (Laufer & Nation, 1999).

Furthermore, the productive vocabulary knowledge is also assessed with the Lex 30 (Meara & Fitzpatrick, 2000), where test-takers are shown 30 target words and asked to write down up to four related terms for each one of them. In the past 20 years, several investigations have employed this test to assess L2-English learners' vocabulary productively and investigated its validity (Fitzpatrick & Clenton, 2017). A spoken version of the test, the S_Lex 30, has been recently developed to assess the productive vocabulary orally (Uchihara & Clenton, 2023). Furthermore, researchers have also used the G_Lex (Fitzpatrick & Clenton, 2017), a sentence completion test in which participants are asked to provide up to five words to complete each of 24 sentence gaps. As in the Lex 30 test, only the words provided by test-takers that are not among the 1,000 most frequent in English are considered in the assessment.

2.2. Orthographic and phonological vocabulary knowledge

When it comes to receptive vocabulary knowledge, Nation's taxonomy of what knowing a word actually entails (Nation, 2001) clearly distinguishes between being able to recognize what a word sounds like and what a word looks like. Furthermore, Milton et al. (2010) suggested that language learners stored separately the phonological and orthographic forms

of words. Although this distinction represents a solid argument to undertake research where possible individual differences between those abilities are explored, very few studies have attempted to investigate and quantify whether the language learner's orthographic and phonological vocabularies are similar in size.

Table 1 presents a summary of the studies that have investigated both the phonological and the orthographic vocabulary size of L2-English learners. It features the participants' mother tongue, their sample size and their average language level, as well as the correlation between both measures of vocabulary size and which one was bigger.

Table 1. *Summary of studies on receptive phonological and orthographic vocabulary knowledge*

Author(s)	Learners' L1	*N*	Learners' language level	Correlation phonological-orthographic vocabulary size	Bigger vocabulary size
Milton and Hopkins (2006)	Greek Arabic	88 38	A1-C2	$r = .68$, $p < .01$	Orthographic: L1-Greek Phonological: L1-Arabic
Mizumoto and Shimamoto (2008)	Japanese	332	B1	$r = .89$, $p < .01$	Orthographic
Milton et al. (2010)	Chinese Arabic Japanese	10 10 10	B1-C1	$r = .46$, $p < .05$	Phonological: L1-Arabic Orthographic: Other L1s
Alhazmi and Milton (2015)	Arabic	30	B1	$r = .67$, $p < .01$	Orthographic
Oh (2016)	Korean	75	n/a	$r = .58$, $p < .01$	Orthographic
Aizawa et al. (2017)	Japanese	140	A2	$r = .67$, $p < .05$	Orthographic
Uchihara and Harada (2018)	Japanese	35	B1-C1	$r = .73$, $p < .01$	Orthographic
Masrai (2019)	Arabic, Brazilian, Chinese, Iranian, Japanese	130	B2	$r = .58$, $p < .001$	Orthographic
Ha (2021)	Vietnamese	234	B1	$r = .89$, $p < .01$	Phonological
Aoiz (2022)	Spanish	284	B1	$r = .82$, $p < .001$	Orthographic
Hamada and Yanagawa (2023)	Japanese	155	A2-B1	$r = .70$, $p < .001$	Orthographic

In most studies, participants showed a better ability to recognize words in their written than in their aural form, across a variety of participants' L1s which were very different from English such as Korean, Chinese or Greek. Only L1-Arabic (Milton & Hopkins, 2006; Milton et al., 2010) and L1-Vietnamese language learners (Ha, 2021) had better results in the phonological vocabulary tests. The fact that Arabic and English had different scripts might explain why L1-Arabic participants had higher scores in phonological vocabulary tests.

However, the case is exactly the opposite with Vietnamese as it is a tonal language and English is a pitch-accent language, but both have the same script (i.e., Latin). A second and more plausible account for the bigger size of the phonological vocabulary of those L1-Arabic and L1-Vietnamese participants refers to the methods employed in the studies, which might have led to an overestimation of those participants' phonological vocabulary size. Milton and Hopkins (2006) had a sample of only 38 L1-Arabic learners, with language proficiencies ranging from A1 to C2. Similarly, only 10 participants had Arabic as their mother tongue in the study carried out by Milton et al. (2010). In the case of Ha's study (2016), although the sample was significantly larger (N = 234) and more homogeneous (average language level B1), it employed two different test formats, multiple choice for the phonological vocabulary test and multiple matching for the orthographic vocabulary test, which might account for the higher scores in the former.

3. Research study

This study sought to find empirical evidence for the claim that L2 aural and written vocabulary knowledge are two separate aspects of knowing a word (McLean et al., 2020), and should be assessed separately (Cheng & Matthews, 2018). Firstly, the aural vocabulary size of L2-English learners with varied proficiencies in the target language and whose L1 is Spanish was assessed with a 4-option multiple-choice bilingual vocabulary test. Then, the same items were presented to those learners, but in writing. Answers on both test modalities were compared to study both possible differences and whether learners' language ability and word frequency impact on those disparities.

3.1. Research Instruments

The items for both modalities of the vocabulary test were selected from wordlists compiled by Nation (2012, 2022) from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). The use of such wordlists implies an update with respect to research that has employed the VST (Nation & Beglar, 2007) or the VLT (Schmitt et al., 2001) because they are a more recent and complete compilation than the wordlists upon which those tests are based. Furthermore, they are more balanced because they include other varieties of English apart from the British one, their items are sampled from spoken (60%) rather from written texts (40%), and they are more teaching-oriented (Dang et al., 2022).

The creation of the present vocabulary test entailed several steps. Firstly, each item in the first eight bands of frequency (1-8k) was selected and copied onto a spreadsheet. Then, the main researcher drew on his 20 years of experience teaching English to L1-Spanish learners to write down the most suitable and frequent translation into Spanish for each of those 8,000 terms. Once all items had a translation into Spanish, 30 words were randomly selected from each band until a total of 240 items were sampled. Then, the other three translations for each question in the test were taken from the same frequency band as the target item, as well as from the same part of speech. Unlike the case with the target items in the test, the actual selection of the three other answers to each question was based on

the main researcher's intuitions and experience in language teaching and vocabulary testing. The 960 possible answers featured in this version of the vocabulary test (240 target items * 4 options) were used only once.

Each question presented the target item first on its own, and then embedded within a sentence that only helped to determine which part of speech was tested in each case. By presenting the items in this way the participants were shown where to focus their attention, so that problems derived from a possible lack of noticing (van Zeeland, 2014) were minimised. The three incorrect options in each question were carefully selected to avoid ambiguity and confusion, and the contextualising sentences for the target word were written in a manner that no additional information about its meaning was revealed. Furthermore, in the case of polysemic headwords with two or more associated lemmas, the translation into Spanish of the target item and the featured distractors helped the test-takers to realise what part of speech and what meaning each question was referring to.

The listening vocabulary test was created from the written version. A native English speaker was recorded in a studio while he read out each of the items in the test without concealing his usual accent, prosody, or intonation. The resulting audio file was edited with the software Audacity®, and the questions in the test were separated from each other with 5 seconds, sufficient time to read the four options and select the correct one in each case (van Zeeland, 2014). Figure 1 shows a screenshot with some of the items featured in the test.

Figure 1. *Screenshot with some items featured in the Written Vocabulary Test (WVT)*

3.3. Data collection

A total of 209 learners attending A2- (*N* = 25), B1- (*N* = 60), B2- (*N* = 50), C1- (*N* = 51), and C2- (*N* = 23) L2-English courses participated in the study. All participants were recruited at either State Language Schools or at tertiary education centres in Spain and had a minimum age of 16. They received an email with an invitation to participate in the study and a link to the online version of the test on Google Forms®. Study participants were told to answer all the questions, including those that were totally unknown to them, because in real-life situations they have to make tentative guesses at the meaning of unknown words. Participants were also unaware of the fact that both the listening vocabulary test (LVT) and the written vocabulary test (WVT) presented the same target items first orally and then in writing, but once in the WVT, they were asked not to change any of their answers in the previous section.

4. Results

4.1. Descriptive Statistics: Impact of word frequency, language level and test modality on item difficulty

The original dataset was refined by means of a Rasch analysis that implied the exclusion from further analyses of items with perfect or nearly perfect scores because they conveyed too little information about the participants' performance. However, as the main purpose of the present study was to investigate differences between aural and written vocabulary size, items were excluded only when there were minimal differences in scores from one test modality to the other (score difference <3%). The number of excluded items was similar across all bands of frequency, between 26.7% and 43.3% of all the words originally featured in the test.

Table 2 shows the number of items with perfect scores and their corresponding separation and reliability indices (expressed in logits), depending on how many items are analysed. The item reliability and separation indices increased in both test modalities (aural and written) and across language levels. Furthermore, the refined datasets, i.e., after most items with perfect scores were excluded from further analyses, showed better separation and reliability indices for the aural than for the written vocabulary test.

Table 2. *Number of items with perfect scores depending on number of items analysed and corresponding values for separation and reliability*

Language level	Analysed items	Count perfect scores		Item separation		Item reliability	
		LVT	*WVT*	*LVT*	*WVT*	*LVT*	*WVT*
A2-B1 original	90	16	38	2.30	1.54	.84	.70
A2-B1 refined	61	2	10	3.43	2.26	.92	.84
B2 original	150	51	67	1.80	1.33	.76	.64
B2 refined	101	12	18	2.53	1.95	.87	.79
C1-C2 original	240	81	122	1.98	1.52	.80	.70
C1-C2 refined	160	23	46	2.82	2.10	.89	.81

Once the refinement process concluded, the best performing items were kept for further analysis and the participants' answers studied. Table 3 shows the percentages of correct answers participants had in both modalities of the test. The higher the participant's linguistic level, the more correct answers they had in each of the bands of frequency, and in both the aural and the written vocabulary test. On the other hand, for participants with similar language proficiency, the less frequent the word, the lower the percentage of correct answers. Finally, for participants with similar language levels and for the same band of frequency, the results were better in the written vocabulary test (WVT) than in the listening vocabulary test (LVT).

Table 3. *Percentages of correct answers across language levels and frequency bands*

	Language level					
	A2-B1 (*N* = 85)		B2 (*N* = 50)		C1-C2 (*N* = 74)	
Frequency band	LVT	WVT	LVT	WVT	LVT	WVT
1k (N =20)	81.47%	89.65%	86.60%	92.10%	93.92%	97.09%
2k (N=22)	70.21%	82.46%	77.64%	87.09%	85.57%	92.38%
3k (N=19)	67.86%	78.20%	75.16%	83.68%	84.50%	89.12%
4k (N=19)			60.63%	70.74%	75.25%	82.93%
5k (N=21)			57.14%	68.00%	65.44%	74.71%
6k (N=21)					58.37%	71.04%
7k (N=17)					60.65%	68.52%
8k (N=21)					55.86%	66.41%

Many items showed ceiling effects in the original dataset because all participants answered them correctly, especially among the first bands of frequency in the WVT. On the other hand, in the refined dataset there were still many items from the most frequent words in English that participants were unable to recognize, either aurally or in writing. The percentages of unknown words from high-frequency vocabulary (1-3k) ranged from 10.35 to 32.14% among the B1-level learners; from 7.90 to 24.84% for the B2-level group; and from 2.91 to 15.50% with the most proficient participants.

Pearson's product-moment correlations were computed for the participants' measures expressed in logits in both the LVT and the WVT. The overall correlation for the whole dataset was .59 (p <.001). Then, further calculations were made to establish that correlation depending on the band of frequency and the participant's language ability. Table 4 shows how the LVT and the WVT correlated positively in all the bands and for all participants, regardless of their language level. Although all these correlations were statistically significant (p <.01), they were moderate as their values ranged from .47 to .62.

Table 4. *Pearson product-moment correlations between measures in LVT and WVT in each band of frequency*

		Language level		
		A2-B1 (*df* =83)	B2 (*df* =48)	C1-C2 (*df* =72)
Frequency band	*1k*	.47	.54	.51
	2k	.54	.59	.59
	3k	.46	.56	.62
	4k		.48	.47
	5k		.56	.59
	6k			.53
	7k			.60
	8k			.56

Paired *t*-tests were subsequently employed to determine the significance of the observed differences between the results in both vocabulary tests. In the LVT, when the differences in results from one frequency band to the next were examined, most differences were significant across all language levels ($p < .01$). Likewise, all differences reached the significance level in the WVT ($p < .05$), except for the 4k-5k comparison among B2-level. Furthermore, most of the observed differences yielded medium to large effect sizes according to Cohen's typology (Table 5).

Table 5. *Paired* t*-tests and effect sizes for the differences in measures between frequency bands*

			Language level					
			A2-B1 (*N* = 85)		B2 (*N* = 50)		C1-C2 (*N* = 74)	
			LVT	WVT	LVT	WVT	LVT	WVT
Frequency band	*1k vs 2k*	*p*-value	<.001	<.001	<.001	<.001	<.001	<.001
		Effect size	.97	.77	.96	.68	1.20	.92
	2k vs 3k	*p*-value	.097	<.001	.063	.015	.312	<.001
		Effect size	.21	.37	.26	.35	.11	.42
	3k vs 4k	*p*-value			<.001	<.001	<.001	<.001
		Effect size			1.04	.98	.69	.55
	4k vs 5k	*p*-value			.114	.104	<.001	<.001
		Effect size			.22	.19	.74	.60
	5k vs 6k	*p*-value					<.001	.002
		Effect size					.58	.26
	6k vs 7k	*p*-value					.089	<.001
		Effect size					-.17	.39
	7k vs 8k	*p*-value					.001	<.001
		Effect size					.39	-.09

In a second batch of paired *t*-tests, results in the LVT from each band of frequency for the same group of language ability were compared to their counterparts in the WVT. Table 6 shows that all comparisons presented significant differences from one test modality to the other ($p < .001$ and $p < .05$), and that those differences yielded medium to large size effects.

Table 6. *Paired t-tests and effect sizes for the differences in measures between LVT and WVT*

			Languaje level		
			A2-B1 (*N* = 85)	B2 (*N* = 50)	C1-C2 (*N* = 74)
Frequency band	*1k*	*p*-value	<.001	<.001	<.001
		Effect size	-.830	-.638	-.712
	2k	*p*-value	<.001	<.001	<.001
		Effect size	-1.099	-1.174	-.899
	3k	*p*-value	<.001	<.001	<.001
		Effect size	-.872	-.767	-.465
	4k	*p*-value		<.001	<.001
		Effect size		-.628	-.521
	5k	*p*-value		<.001	<.001
		Effect size		-.762	-.756
	6k	*p*-value			<.001
		Effect size			-.912
	7k	*p*-value			<.024
		Effect size			-.292
	8k	*p*-value			<.001
		Effect size			-.751

Results were also grouped into three categories according to how frequent the target words in the test were: high-frequency (1-3k), mid-frequency (4-5k) and low-frequency (6-8k). For those test-takers with a B2-level, their results in items from the first three bands (1-3k) were clearly better than in the mid-frequency vocabulary (4-5k). The differences were statistically significant ($p < .001$) and with a large effect size in both the LVT and the WVT (1.95 and 1.43, respectively). For the participants with a C1-C2 proficiency, their results in the most frequent vocabulary were better than in the mid-frequency, and those in bands 4-5k were comparatively better than in the low-frequency vocabulary (6-8k). The differences in the comparisons for that group of participants were statistically significant ($p < .001$) and had medium to large effect sizes. Table 7 shows the significance of those differences and their corresponding effect sizes.

Table 7. *Paired t-tests and their effect sizes for the differences in measures between high-frequency, mid-frequency and low-frequency vocabulary*

			Language level			
			B2 (N = 50)		C1-C2 (N = 74)	
			LVT	WVT	LVT	WVT
Frequency band	*1-3k → 4-5k*	*p*-value	<.001	<.001	<.001	<.001
		Effect size	1.95	1.78	1.90	1.60
	4-5k → 6-8k	*p*-value			<.001	<.001
		Effect size			1.04	.75

Finally, results from A2-B1 participants were compared to the ones from the B2-level participants, and theirs with the results from the C1-C2 group for the same band of frequency. All the comparisons across levels and bands of frequency reached the significance level in both tests ($p < .05$) and yielded medium to large effect sizes. The results from the paired *t*-tests employed to analyse the significance of the differences and the effect sizes of those differences are shown in Table 8.

Table 8. *Paired t-tests and their effect sizes for the differences in measures between language levels*

		Language level							
		A2-B1 → B2				B2 → C1-C2			
		LVT		WVT		LVT		WVT	
		p-value	Effect size	*p*-value	Effect size	*p*-value	Effect size	*p*-value	Effect size
Frequency band	*1k*	.001	-.26	.027	-.15	.002	-.47	.009	-.42
	2k	<.001	-.25	.001	-.20	<.001	-.29	.001	-.30
	3k	<.001	-.29	.004	-.25	.002	-.43	.023	-.30
	4k					<.001	-.73	<.001	-.60
	5k					.001	-.27	<.001	-.25

5. Discussion

5.1. How different are the aural and written vocabulary size of L2-English learners?

Participants' answers for the items in all the bands in the listening vocabulary test correlated moderately with their corresponding answers for the same items in the written vocabulary test, with positive Pearson product-moment correlations ranging from .47 to .62 (Table 4) and an overall correlation for all the measures in both tests established at .59 ($p <$ .001). It is worth mentioning that the calculations were made from the measures expressed in logits and not directly from the scores obtained by participants in each of the tests. If that were the case, the overall correlation would be higher (.74, $p < .001$).

Nevertheless, these results are in line with what recent research has shown about the differences between the orthographic and the phonological vocabulary size of L2-English learners in receptive vocabulary tests. For example, Milton and Hopkins (2006) found a strong correlation (r = .68, p < .01) between the orthographic and phonological vocabulary among L2-English learners whose first language was either Greek or Arabic. Similarly, Milton et al. (2010) showed that the Spearman correlation between both measures was .46. Uchihara and Harada (2018) also studied the vocabulary size of L1-Japanese learners of English, first orally and then in writing, and found a significant correlation between the two measures (r = .73, p < .01). Furthermore, Masrai (2019) found that measures of receptive written and aural vocabulary correlated moderately (r = .58, p < .001).

Moreover, unlike the above-mentioned research, some studies have investigated the differences between the phonological and orthographic vocabulary size of L2-English learners with receptive vocabulary tests featuring the same target words in both test modalities. For example, Mizumoto and Shimamoto (2008) set the correlation of both test scores at.89, Ha (2021) at .88, Aoiz (2022) at .82, and Hamada and Yanagawa (2023) at .70. The higher correlations presented by these sets of test scores support the use of identical versions of the same test to investigate the possible differences between aural and written vocabulary size (Hamada & Yanagawa, 2023).

Although correlation values show that there is some range of overlap, when results are analysed in terms of item difficulty, aural and written vocabulary might be considered different aspects of the construct of vocabulary knowledge (González-Fernández & Schmitt, 2020). Fewer items in the LVT were answered correctly by all participants (perfect scores) than in the WVT (Table 2). Likewise, the percentages of correct answers across all bands of frequency show that study participants, regardless of their overall language level, found the items in the LVT more difficult than their counterparts in the WVT (Table 3). The observed differences between the LVT and the WVT reached the significance level in all bands (p < .05). Furthermore, most of the observed differences between results in the LVT and the WVT for the same band of words and the same group of language proficiency yielded medium to large effect sizes (Table 6).

These differences are in line with what most previous research has claimed. For example, Milton and Hopkins (2006) found that Greek learners showed bigger orthographic vocabulary sizes than their phonological ones, whereas the opposite was true among the L1-Arabic participants in their study. In a similar study with a group of L2-English learners from a variety of L1s, Milton et al. (2010) showed that only L1-Arabic learners of English had smaller orthographic than phonological vocabulary sizes.

Furthermore, among L1-Japanese students, Mizumoto and Shimamoto (2008) found that the scores in the orthographic version of the test were statistically higher than the ones in the aural test, regardless of word frequency. Similarly, Uchihara and Harada (2018) and Hamada and Yanagawa (2023) presented results where Japanese native speakers had significantly better scores in the written version of the test. In a study on the impact of vocabulary knowledge on listening comprehension among L2-English learners with different L1s, Masrai (2019) also discovered that language learners showed higher abilities to recognize words when they were written and not spoken. Finally, Aoiz (2022), in his study of the impact of aural and written vocabulary on listening comprehension, showed that L2-English learners found it easier to recognize the orthographic form of words.

On the other hand, L1-Arabic learners showed better results in phonological vocabulary tests (Milton & Hopkins, 2006; Milton et al. 2010), and among L1-Vietnamese L2-English learners, Ha (2021) found slightly bigger aural vocabulary sizes than written ones. Two possible reasons might account for this inconsistency of results on whether the L2-English learners' orthographic or phonological vocabulary size is bigger (Table 1). The first one refers to the possible influence of the learners' L1 on the way they learn vocabulary, as it was claimed by Milton and Hopkins (2006) with respect to their L1-Arabic participants when they had difficulties to recognize words written in a completely different alphabet. However, learners with such different scripts from English like Japanese or Korean were more proficient at recognizing words in their written than in their oral form (Table 1). Furthermore, in Ha's study (2021) L1-Vietnamese participants showed better results in the recognition of the aural form of words although their mother tongue is a tonal language. Secondly, and more importantly, the methods employed for the estimation of vocabulary sizes might have impacted on this difference in findings as the sample was relatively small (Milton & Hopkins, 2006; Milton et al., 2010), or because a different test format was used in each test modality (Ha, 2021).

5.2. How do L2-English learners' aural and written vocabulary size correlate with their overall linguistic proficiency?

The Pearson product-moment correlations for the results in the LVT for each of the bands with their corresponding results in the WVT increased with the participants' language ability (Table 4). In the group of test-takers attending A2-B1 classes, aural and written vocabulary correlated at .50 for the most frequent vocabulary items (1-3k), at .57 for the B2-level, and .60 for the participants in the C1-C2 level of proficiency. For mid-frequency vocabulary (4-5k), results in the LVT and WVT for the B2 group correlated at .52, whereas the C1-C2 participants showed correlations at .54. Finally, this group of language learners had results in both the LVT and WVT for the low-frequency vocabulary (6-8k) correlating at .56. In other words, the higher the language level a learner has, the more similar are their aural and written vocabulary sizes.

When the scores were studied according to the participants' language ability, the percentage of correct answers increased: the higher the language ability the higher the score, in all the bands and for both tests (Table 3). Furthermore, when the scores among the A2-B1 learners were compared with the ones obtained in the same band and test modality by the B2 learners, the differences were significant and yielded moderate effect sizes. Likewise, the scores among the C1-C2 group for each band of frequency were significantly better in both tests than their counterparts among the B2 group, with differences showing moderate to large effect sizes (Table 6).

These results align with what research has claimed about the positive correlation between L2-English learners' language proficiency and their vocabulary size (Zhang & Zhang, 2022). Furthermore, several findings from this investigation provide an additional set of empirical evidence to consider vocabulary acquisition as a developmental process, where "stronger forms of vocabulary knowledge suggest greater mastery of the words in question" (McLean et al., 2020, p.407). In this respect, Milton and Hopkins (2006) found that although at the outset of learning, among less proficient learners, the phonological vocabulary size might

be bigger than the orthographic one, once they have gained more proficiency in the target language their knowledge of the written form of words becomes comparatively larger.

On the other hand, the present study shows that even among the least proficient learners (A2-B1), scores in the written vocabulary tests were comparatively higher than in the aural vocabulary tests. Furthermore, similar to the results in this study (Table 3), Mizumoto and Shimamoto (2008) found that the differences between aural and written vocabulary sizes were more acute among low-level learners. A possible explanation for these differences is that Milton and Hopkins (2006) determined the participants' overall language proficiency depending on their results in the written vocabulary tests they had taken and not on the level of the courses they were attending.

The present study has shown that L2-English learners develop knowledge of the aural forms of words later than of written forms, regardless of their language level or the word frequency (Table 3). Additionally, it has shown that their knowledge of words from the same bands of frequency and in the same test modality significantly increases with their overall language proficiency (Table 3), but also that their knowledge of each word in either its aural or written form shows stronger correlations among more proficient learners (Table 4).

5.3. How does word frequency impact on L2-English learners' oral and written vocabulary size?

Results in both the LVT and the WVT across all levels of language proficiency are clearly influenced by word frequency: the higher the frequency of a word, the more easily it is recognised (Table 3). This finding is in line with what previous empirical research has claimed (for example, Alhazmi & Milton, 2015). Furthermore, the differences between results in one band of frequency with respect to the next were significant in most of the cases, particularly in the WVT (Table 5). These findings align with what Hamada and Yanagawa (2023) found, i.e., the differences between aural and written vocabulary size among L2-English learners depend on how frequent the words are: the more frequent the word, the smaller the difference. Finally, what this research and previous studies (Zhang & Zhang, 2022) have shown with respect to the positive correlation between word frequency and vocabulary size supports the perspective that vocabulary knowledge is acquired depending on how often the learner is in contact with a word (Ellis, 2002).

6. Implications for researchers and teachers

This study has presented solid empirical evidence to support the claim that aural and written vocabulary knowledge are two separate aspects and should be assessed with different instruments. At the same time, vocabulary teaching should begin to assume that knowing a word is more than being able to recognize it in writing (Nation, 2001), and make sure that L2-English learners develop knowledge of both the aural and written form of words.

Furthermore, the perspective on the developmental acquisition of vocabulary (McLean et al., 2020) also calls for the inclusion of pedagogies aiming at the acquisition of other aspects of knowing a word, without assuming that being familiar with one aspect implies doing the same with the rest (González-Fernández & Schmitt, 2020).

Finally, as the present study assessed only the form-meaning link, which is the first aspect of knowing a word that leaners acquire (Schmitt, 2008), language teachers should insist more on learning the most frequent vocabulary in English, even among very proficient learners, as there are still many items unknown to them (Table 3).

7. Limitations and future research

One possible limitation of this study is the absence of measures to prevent practice-of-order effects because of the decision of delivering the same target items in both tests and in the same order, first orally and then in writing. However, the results featured in Table 3 confirm that percentages of participants correct answers were not influenced by practice-of-order effects because they depended on how frequent the word was, how proficient those participants were, and how the items were delivered. The more frequent the word and the higher the participant's linguistic proficiency, the better the results. Likewise, for all frequency bands and participants, results were comparatively better in the WVT. Although it was delivered after the LVT, the logical tiredness when answering the last questions in a test might not have appeared among the study participants. Moreover, the possible washback effect of answering the same questions for a second time, with the logical improvement in the results, was prevented as the target items were first presented orally, and the scores and the correct answers in the first test or any other relevant clues were not given to the test-takers.

Among the aspects that might be improved in future investigations is the use of a larger sample size, particularly for the lowest and highest level of language proficiency (A2 and C2, respectively), so that enhanced item reliability values are shown (Table 2). Furthermore, future research should also assess other forms of knowing a word (Nation, 2001) that might be acquired later because they imply more effort on the learner's part (McLean et al., 2020), instead of just identifying who is able to recognize the form-meaning link by choosing the correct option in a bilingual multiple-choice test.

Another line for future research refers to continuing the investigations on the influence of the vocabulary size of language learners on their ability to understand written or aural texts, but also on the influence of language activities like listening or reading on how they learn vocabulary.

A final limitation of this study refers to the sample of participants because they all had Spanish as their first language, so findings should be generalized with caution to learners with other L1s. On the other hand, findings might be more generalizable to learners with languages more similar to Spanish than for example Vietnamese (Ha, 2021) or Japanese (Hamada & Yanagawa, 2023), particularly those of the Romance family like French, Italian or Portuguese (Schmitt et al., 2001).

8. References

Abdullah, K. I., Puteh, F., Azizan, A. R., Hamdan, N. N. I., & Saude, S. (2013). Validation of a controlled productive Vocabulary Levels Test below the 2000-word level. *System*, *41*(2), 352-364. https://doi.org/10.1016/j.system.2013.03.005

Aizawa, K., Iso, T., & Nadasdy, P. (2017). Developing a vocabulary size test measuring two aspects of receptive vocabulary knowledge: visual versus aural. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 1-6). https://doi.org/10.14705/rpnet.2017.eurocall2017.679

Alhazmi, K., & Milton, J. (2015). Phonological vocabulary size, orthographic vocabulary size, and EFL reading ability among native Arabic speakers. *Journal of Applied Linguistics*, *30*, 26-43. https://doi.org/10.26262/jal.v0i30.8297

Aoiz, M. (2022). Relationship Between L2 Vocabulary Size and Listening Ability. *Huarte De San Juan. Filología Y Didáctica De La Lengua*, *21*, 133–166. https://doi.org/10.48035/rhsj-fd.21.6

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing, 35*(1), 3-25. https://doi.org/10.1177/0265532216676851

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*(4), 617-641. https://doi.org/10.1177/1362168820911189

Du, G., Hasim, Z., & Chew, F. P. (2022). Contribution of English aural vocabulary size levels to L2 listening comprehension. *International Review of Applied Linguistics in Language Teaching*, *60*(4), 937-956. https://doi.org/10.1515/iral-2020-0004

Ha, H. T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia*, *11*(1), 1-20. https://doi.org/10.1186/s40468-021-00131-8

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253-272. https://doi.org/10.1177/0265532212459028

Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? A Review of the Research. *Studies in Second Language Acquisition, 24*(2), 223-236. https://doi.org/10.1017/S0272263102002073

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL QUARTERLY*, *51*(4), 844-867. https://doi.org/10.1002/tesq.356

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481-505. https://doi.org/10.1093/applin/amy057

Hamada, Y., & Yanagawa, K. (2023). Aural vocabulary, orthographic vocabulary, and listening comprehension. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2022-0100

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC journal*, *43*(1), 53-67. https://doi.org/10.1177/0033688212439359

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing*, *16*(1), 33-51.

Masrai, A. (2019). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review, 11*(3), 423-447. https://doi.org/10.1515/applirev-2018-0106

Masrai, A. (2022). The relationship between two measures of L2 phonological vocabulary knowledge and L2 listening comprehension. *TESOL Journal,13*(1),e612. https://doi.org/10.1002/tesj.612

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

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389-411. https://doi.org/10.1177/0265532219898380

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

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19-30. https://doi.org/10.1016/S0346-251X(99)00058-5

Milton, J., & Hopkins, N. (2006). Comparing Phonological and Orthographic Vocabulary Size: Do Vocabulary Tests Underestimate the Knowledge of Some Learners. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, *63*(1), 127–147. https://doi.org/10.1353/cml.2006.0048

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. *Insights into non-native vocabulary teaching and learning*, *52*, 83-98.

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners. *Language Education & Technology*, *45*, 35-51. https://doi.org/10.24539/let.45.0_35

Nation, I. S. P. (2001) Learning vocabulary in another language. Cambridge University Press.

Nation, I. S. P. (2012, 2022). The BNC/COCA word family lists (17 September 2012). Unpublished paper. [online] Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher, 31*(7)*,* 9–13.

Nguyen, L. T. C. & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC journal*, *42*(1), 86-99. https://doi.org/10.1177/0033688210390264

Oh, E. (2016). Comparative studies on the roles of linguistic knowledge and sentence processing speed in L2 listening and reading comprehension in an EFL tertiary setting. *Reading Psychology*, *37*(2), 257-285. https://doi.org/10.1080/02702711.2015.1049389

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language teaching research*, *12*(3), 329-363. https://doi.org/10.1177/1362168808089921

Smith, G. (2019). *The relationship between L2 vocabulary knowledge and listening comprehension ability: A metanalysis*. Paper presented at the 2019 Conference of the American Association for Applied Linguistics. Atlanta, Georgia. Retrieved from https://www.researchgate.net/publication/331876888_The_Relationship_between_L2_Vocabulary_Knowledge_and_Listening_Comprehension_Ability_A_Meta- Analysis.

Uchihara, T., & Clenton, J. (2023). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal*, *51*(3), 376-393. https://doi.org/10.1080/09571736.2022.2080856

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly, 52*(3), 564-587. https://doi.org/10.1002/tesq.453

van Zeeland, H. (2014). *Second language vocabulary knowledge in and from listening* (Doctoral dissertation, University of Nottingham).

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696-725. https://doi.org/10.1177/1362168820913998